

**Proceedings of
The 5th SEAMS – GMU INTERNATIONAL CONFERENCE
ON MATHEMATICS AND ITS APPLICATIONS
Yogyakarta – Indonesia, 24th – 27th July 2007**

**THE ROLE OF MATHEMATICS,
COMPUTATIONAL MATHEMATICS,
AND ITS APPLICATIONS
IN THE DEVELOPMENT OF SCIENCE AND
TECHNOLOGY**

**Editors:
Supama
Irwan Endrayanto
Khabib Mustofa
Indah Emilia Wijayanti
Herni Utami**

**Published by
DEPARTMENT OF MATHEMATICS
GADJAH MADA UNIVERSITY, YOGYAKARTA – INDONESIA
Phone: +62 274 552243, +62 274 7104933
Fax.: +62 274 555131**

REFEREES

Zhao Dongsheng
Nanyang Institute of Education
SINGAPORE

J.A.M van der Weide
Technical University Delft
The NETHERLANDS

Peter F. Bullen
University of British Columbia
CANADA

Hidetoshi Marubayashi
Naruto University
JAPAN

Hendra Gunawan
Institute Technology Bandung
INDONESIA

Edy Soewono
Institute Technology Bandung
INDONESIA

Edy Tri Baskoro
Institute Technology Bandung
INDONESIA

Roberd Saragih
Institute Technology Bandung
INDONESIA

Irawati
Institute Technology Bandung
INDONESIA

Yudi Soeharyadi
Institute Technology Bandung
INDONESIA

Marjono
Brawijaya University
Malang – INDONESIA

Sri Nurdianti
Institute Pertanian Bogor
Bogor – INDONESIA

Ari Suparwanto
Gadjah Mada University
Yogyakarta – INDONESIA

Budi Surodjo
Gadjah Mada University
Yogyakarta – INDONESIA

Ch. Rini Indrati
Gadjah Mada University
Yogyakarta – INDONESIA

Salmah
Gadjah Mada University
Yogyakarta – INDONESIA

Indah Emilia Wijayanti
Gadjah Mada University
Yogyakarta – INDONESIA

Sri Haryatmi
Gadjah Mada University
Yogyakarta – INDONESIA

Dedi Rosadi
Gadjah Mada University
Yogyakarta – INDONESIA

Danardono
Gadjah Mada University
Yogyakarta – INDONESIA

Widodo
Gadjah Mada University
Yogyakarta – INDONESIA

Khabib Mustofa
Gadjah Mada University
Yogyakarta – INDONESIA

Lina Aryati
Gadjah Mada University
Yogyakarta – INDONESIA

Edi Winarko
Gadjah Mada University
Yogyakarta – INDONESIA

PREFACE

It is an honor and great pleasure for the Department of Mathematics – Gadjah Mada University, Yogyakarta – INDONESIA, to be entrusted by the South East Asian Mathematical Society (SEAMS) with the organization of its international conference every four years. Appreciation goes to those who have developed and established this tradition of the succesful series of conferences.

The SEAMS – Gadjah Mada University International Conference 2007 on Mathematics and Its Applications was held on 24 – 27 July 2007 at Gadjah Mada University. The conference was the follow up of the succesful series of events which have been held in 1989, 1995, 1999, and 2003.

The conference has achieved its main purposes of promoting the exchange of ideas and presentation of recent development, particularly in the areas of pure, applied, and computational mathematics which are represented in South East Asian Countries. The conference has also provided a forum of researchers, developers, and practitioners to exchange ideas and to discuss future direction of research. Moreover, it has enhanced collaboration between researchers from countries in the region and those from outside.

The conference was attended by more than 200 participants. They come from Japan, Canada, India, Nepal, Taiwan, The Netherlands, Germany, Oman, Malaysia, Singapore, The Philippines, and Indonesia. During the four days conference, there were 15 plenary lectures and than 148 contributed papers communications. The plenary lectures were delivered by Prof. Chew Tuan Seng (Singapore), Prof. Zhao Dongsheng (Singapore), Prof. J.A.M. van der Weide (the Netherlands), Prof. R. Helmers (the Netherlands), Prof. J.C. Engwerda (the Netherlands), Prof. Holger Hermans (Germany), Prof. Saburou Saitoh (Japan), Prof. Y. Sawano (Japan), Prof. H. Marubayashi (Japan), Prof. Y. Hirano (Japan), Prof. Hendra Gunawan (Indonesia), Prof. Sri Wahyuni (Indonesia), Andonowati, PhD (Indonesia), Edi Winarko, PhD (Indonesia), Dr. Dedi Rosadi (Indonesia). Most of the contributed papers were delivered by mathematicians from Asia.

The proceedings consists of 9 invited lectures and 54 refereed contributed papers.

We have the good fortune by getting some support from friends, many organizations, parties, and institutions. Therefore, on behalf of the organizing committee, may I extend our thanks to the Abdus Salam International Centre for Theoretical Physics (ICTP) Trieste Italy and UNESCO Jakarta Office, without their generous support this conference could not have an international character. We would like also to acknowledge and express our deepest gratitude to Gadjah Mada University (UGM), Department of Math. UGM, Faculty of Math. And Natural Sciences UGM, Graduate School of Mathematics UGM, Graduate School of Computer Science UGM, and Asram Furniture and Interior for their assistance and support. We would like also to extend our appreciation and deepest gratitude

to all the invited speakers, all participants, and referees for the wonderful cooperation, the great coordination, and the fascinating efforts. Appreciation and special thanks are addressed to our colleagues and staffs who help in editing papers. Finally, we acknowledge and express our thanks to all friends, colleagues, and staffs in the Department of Mathematics UGM for their help and support in the preparation for and during the conference.

Editorial Board
Supama
Irwan Endrayanto
Khabib Mustofa
Indah Emilia Wijayanti
Herni Utami.

Table of Contents

Invited Talks

Scott Closed Set Lattices and Applications <i>Zhao Dongsheng</i>	1
Multicriteria Dynamic Optimization Problems and Cooperative Dynamic Games <i>Jacob Engwerda</i>	17
On the Bootstrap Accuracy for the Trimmed Mean: A Survey of Some Recent Developments <i>N.V. Gribkova and R. Helmers</i>	55
A MoDeST Evaluation of the Interplay between Energy Consumption and Clock Drift in ZigBee <i>Christian Groß, Holger Hermanns and Reza Pulungan</i>	69
Testing For Independence In Gaussian Time Series Using the Codifference Function: Some Finite Sample Properties <i>Dedi Rosadi</i>	95
Applications of Reproducing Kernels to Numerical Analysis through the Tikhonov Regularization <i>Saburou Saitoh</i>	105
Atomic Decomposition of Morrey Spaces <i>Yoshihiro Sawano</i>	125
Mathematical Finance And Probability Theory <i>J.A.M. van der Weide</i>	133
The Discovery of Temporal Patterns and Rules from Sequential Data <i>Edi Winarko</i>	145

Contributed Papers

Part I: Algebra and Combinatorics

Sufficient and Necessary Conditions of a Bounded Krull Order <i>Santi Irawati and Hidetoshi Marubayashi</i>	161
Fuzzy Version of the Fundamental Theorems of Homomorphism Groups <i>Karyati</i>	167
A Tutorial and Survey on Codes Over Rings <i>Maheshanand and Siri Krishan Wasan</i>	177

The Existence of a Solution of the Non-Homogeneous Systems over an Elementary Divisor Domain on Solvable Module <i>Ikrar Pramudya, Sri Wahyuni, Ari Suparwanto and Widodo</i>	195
On Edge Magic Total Labeling of Caterpillar-Like Trees <i>Kiki A. Sugeng and Denny R. Silaban</i>	203
Semiprime Submodules Characterizations <i>Suprpto and Indah Emilia Wijayanti</i>	209
On Prime L -Subcomodules <i>Indah Emilia Wijayanti</i>	213

Part II: Analysis

On Certain Classes of Close-To-Convex Functions <i>Khalifa Al-Shaqsi and Maslina Darus</i>	227
Representation of SM -Operators on Classical Banach Spaces <i>Muslim Ansori, Soeparna Darmawijaya and Supama</i>	237
A Real Measure versus A Vector Measure <i>Soeparna Darmawijaya</i>	249
Some Convergence Theorems for the Henstock-Stieltjes Integral <i>Umi Mahnuna Hanung</i>	257
On the Young Inequality and Locally Boundedness of Superposition Operators on Modularized Spaces <i>Supama</i>	267

Part III: Applied Mathematics

A Mathematical Scenario of an Optimum and SustainableHarvesting of Teak Wood <i>Arman and Edi Cahyono</i>	279
Inverse Problem of Laplace Equation in Plane Domain Containing a Crack <i>Wono Setya Budhi and Marwan Wirianto</i>	289
Optimal Downlink Rate Allocation in Multicell CDMA Networks <i>A.I. Endrayanto, A.F. Gabor and R.J. Boucherie</i>	297
Beggs-Brill Corelaton Discritization For Predicting Pressure Distribution In Two Phase Flow Pipeline Network <i>A.I. Hasan, R.P. Pamungkas, L.S. Riza, W.U. Baroto, K.A. Sidarto and L. Mucharam</i>	309

Hypertranscendental Equation <i>Chin Chou Hsin</i>	321
Bifurcation Analysis of a New-Product Growth Model <i>Joice Ruth Juliana and M. Setijo Winarko</i>	327
A Contextual Approach, using Problem-Based, to Teaching and Learning Differential Equations in the Universiti Tun Hussein Malaysia <i>Nafisah Kamariah MD Kamaruddin and Wan Mohd Rashid Wan Ahmad</i>	339
Nonlinear MHD Mixed Convection Flow and Heat and Mass Transfer of an Electrically Conducting Variable Viscosity over a Porous Wedge <i>R. Kandasamy, I. Hashim , Muhaimin and Ruhaila</i>	347
Derivation of the NLS Breather Solutions using Displaced Phase-Amplitude Variables <i>Natanael Karjanto and E. van Groesen</i>	357
Symmetric Rank One Method for Unconstrained Optimization Problems <i>M.B Mamat, K.H. Sow , Y.B Dasril, and I.B Mohd</i>	369
Functional Analysis of the Linearized Stokes Problem in a Heterogeneous Media <i>H.A. Parhusip</i>	377
The Application Mathematics for Inverse Acoustics Problem <i>Ratnadewi</i>	389
On the GQ-Consistent Quantum Borel Kinematics: Unitary Representations of the Groups of Canonical Transformations <i>M.F. Rosyid</i>	397
On the Relation Between Configuration and Phase Space Quantization II <i>M.F. Rosyid</i>	411
N-Player Linear Quadratic Dynamic Game for Descriptor System <i>Salmah</i>	423
Tracking Optimal Control for Flexible System <i>Roberd Saragih</i>	433
Performance of Genetic Algorithm for Permutation Flowshop Scheduling Problem <i>D.R. Silaban, D. Kusharyadi and Y. Satria</i>	441
The Immersed Interface Method for Two Channels Dissipation Model with Discontinue Constants Velocity <i>Sumardi, Soeparna Darmawijaya and Lina Aryati</i>	453
Estimating the Value of the Basic Reproduction Number in an Age-Structured SIR Host-Vector Model <i>Asep K. Supriatna and Edy Soewono</i>	465

Stability Analysis of FTBS on Shallow Water Equations <i>L.H. Wiryanto and Warsoma Djohan</i>	469

The Initial Functions in Wavelet Frame Construction <i>Mahmud Yunus, Hendra Gunawan and Janny Lindiarni</i>	475

Part IV: Computer Science

Texture Fabric Defect Detection System Utilizing Image Processing and Artificial Neural Network <i>Sri Hartati</i>	485

Knowledge Management System based on Ontology Model a Case Study on Plant Data Searching <i>Herlina Jayadianti, F.Soesianto, Abdul Kadir and Azhari</i>	495

A Case Study of XML to Relational Database Translation using XML-Enabled Database <i>Anny Kartika Sari</i>	507

TAR-QL: An SQL-like Query Language for Querying Discovered Temporal Association Rules <i>Edi Winarko</i>	519

Part V: Statistics

An Introduction of Spatial Structural Equations Model <i>Isah Aisah, Budi Nurani and Sutawanir Darwis</i>	531

Estimating the Ability of <i>Scylla Serata</i> , <i>Pinctada Margaritifera</i> , <i>Perna Viridis</i> as Adsorbent for Lead from Aqueous Solution: Using Linear Regression Model <i>Kamsia Budin, Amran Ahmed, Siti Rahayu Mohd. Hashim, Noraini Abdullah and Alice Ngo Chin-Chin</i>	539

Dual Time Scales in Event History Analysis <i>Danardono</i>	547

Development Of Bootstrap Simulation For Tracer Modeling <i>Sutawanir Darwis, Agus Yodi Gunawan, M. Ali Ashat, Sri Wahyuningsih and Nurtiti Sanusi</i>	557

Numerical Simulation of Statistical Tests to Determine the Best Subinterval in Global Optimization <i>Yosza Dasril, Ismail Bin Mohd and Siska Candra Ningsih</i>	563

Generating Square-Root Diffusions Process by Monte Carlo Simulation with Application in Syareea Insurance <i>Adhitya Ronnie Effendie, Subanar and Danardono</i>	573

Binomial Method for Indonesian Option Pricing <i>Gunardi, Subanar, J.A.M. van der Weide and Sri Haryatmi</i>	579
Regime Switching Garch (an Application to Dowjones Index Return) <i>M. Arbi Hadiyat, Nur Iriawan and Slamet Mulyono</i>	589
An Application of the GSTAR-Kriging Model <i>Budi Nurani Ruchjana and Sutawanir Darwis</i>	597
Statistical Significance Test For Neural Network Classification <i>Sri Rezeki, Subanar and Suryo Guritno</i>	605
The Technique of Data Robustness in Logit Modeling <i>Mohd Saifullah Bin Rusiman</i>	615
Stability of a Neuronal Model with Distributed Delay and Time-Dependent Parameters <i>Sariyasa</i>	623
Asymptotic Normality and Model Selection in Neural Networks by using Inference of R^2 Incremental and Wald Test for Time Series Forecasting <i>Suhartono, Subanar and Suryo Guritno</i>	629
Optimization Radila Basis Function Neural Network <i>Brodjol Sutijo, Subanar and Suryo Guritno</i>	641
Asymptotic Properties of Integrated Renewal Processes <i>Suyono and J.A.M. van der Weide</i>	649
Model, Calculations and Applications of Spatial Scan Statistics <i>Yekti Widyaningsih</i>	657

Invited Talks

SCOTT CLOSED SET LATTICES AND APPLICATIONS

ZHAO DONGSHENG

Abstract. Given a poset P , all the Scott closed sets of P form a complete lattice. Different types of P yield different lattices of Scott closed sets. In this survey paper, we shall examine some of the main results on such lattices and their applications in constructing other objects in domain theory.

Key words and Phrases: D-completion, continuous poset, continuous dcpo, local dcpo, completely distributive lattice, co-prime element, Hoare powerdomain, C-algebraic lattice.

1. INTRODUCTION

Domain theory is a relatively new branch of mathematics that studies special kinds of partially ordered sets- continuous directed complete posets, or domains, and its applications. The main applications are in theoretical computer sciences(see[1]). Domains can be used to model the idea of approximation and convergence in a very general way. This theory also has close relations to topology, metric spaces, lattice theory and category theory. One of the fundamental structures in domain theory is the Scott topology, introduced first by Dana Scott in his searching for denotational semantics. The set of all Scott open sets of domains form a special kinds of complete lattices—completely distributive lattices, which have been studied longtime before domains. In the early time, the lattices of Scott open sets of continuous posets have been extensively studied and used in many aspects. However, the recent development of domain theory shows that the lattices of Scott

closed sets are also useful and important in many aspects. In this survey paper we shall examine the recent development of researches in lattices of Scott closed sets—called Scott closed set lattices. In Section 2, we present the basic definitions and results on posets, Scott topology and continuous posets. In Section 3, a dcpo completion of posets, called D-completion is constructed. For any poset P , P is continuous if and only if its D-completion is continuous. Every poset and its D-completion have the isomorphic Scott closed set lattice. In Section 4, a functor from the category of continuous dcpos to the category of completely distributive lattices is constructed, using the Scott closed set lattices. This functor establishes an equivalence between the above two categories. Section 5 displays the relationship between the Scott closed set lattices and power domains. The Hoar power domain of a dcpo consists of all nonempty Scott closed sets of the dcpo. In Section 6, some general properties of the Scott closed set lattices of posets are presented. For complete semilattices and complete lattices, a complete characterization of the Scott closed set lattices are given. In the last section, we list some problems on Scott closed set lattices.

2. PRELIMINARIES

In this section we recall the basic definitions and results on posets and Scott topology.

A partially ordered set, or poset for short, is a nonempty set P together with a binary relation $\leq$ which is

- (a) reflexive (for all $x \in P, x \leq x$),
- (b) transitive (for all $x, y, z \in P, x \leq y, y \leq z$ imply $x \leq z$), and
- (c) antisymmetric (for any $x, y \in P, x \leq y$ and $y \leq x$ imply $x = y$).

The supremum of a subset A of a poset P , denoted by $\sup A$ or $\bigvee A$, is the least upper bound of A . The infimum of a subset A of P , denoted by $\inf A$ or $\bigwedge A$, is the greatest lower bound of A . For convenience, we write $x \vee y$ for $\sup\{x, y\}$ and $x \wedge y$ for $\inf\{x, y\}$, respectively.

A poset is called a complete lattice if every subset has a supremum and an infimum.

A nonempty subset D of a poset is directed if for any two elements x and y in D , there is $z \in D$ such that $x \leq z$ and $y \leq z$.

A poset P is a directed complete poset, or dcpo for short, if every directed subset of P has a supremum in P .

If X is a subset of poset P , we write $\downarrow X = \{y \in P : y \leq x \text{ for some } x \in X\}$ and dually, $\uparrow X = \{y \in P : y \geq x \text{ for some } x \in X\}$. A subset X is a lower (upper) set if $X = \downarrow X$ ($X = \uparrow X$).

A subset X of a poset P is called Scott closed if

- i) $X = \downarrow X$ and
- ii) for any directed subset D , $D \subseteq X$ implies $\bigvee D \in X$ whenever $\bigvee D$ exists.

The complements of Scott closed sets are called Scott open sets. Thus a subset U of P is Scott open if U is an upper set, and for any directed set $D \subseteq P$, $\bigvee D \in U$ implies $D \cap U \neq \emptyset$ whenever $\bigvee D$ exists.

All the Scott open sets form a topology on P , which is called the Scott topology on P and is denoted by $\sigma(P)$. Also we denote $\Sigma P = (P, \sigma)$ and call it the Scott space of P .

The set of all Scott closed subsets of P will be denoted by $C(P)$. Then $C(P)$ is a complete lattice with respect to the inclusion relation. For more about posets, dcpos and Scott topology we refer to [6].

Example 1 (1) Let $\mathbf{R}$ be the set of all real numbers equipped with the conventional order of numbers. Then a subset $U \subseteq \mathbf{R}$ is Scott open iff $U = \mathbf{R}$ or $U = (a, +\infty)$ for some $a \in \mathbf{R}$.

(2) Let X be a set and $\mathcal{P}(X)$ be the power set (the set of all subsets of X). With respect to the inclusion relation $\subseteq$, $\mathcal{P}(X)$ is a complete lattice. A subset $\mathcal{U} \subseteq \mathcal{P}(X)$ is Scott open if and only if there exists a collection $\{A_i : i \in I\}$ of finite subsets of X such that

$$\mathcal{U} = \{B \subseteq X : \text{there exists } A_i, B \supseteq A_i\}.$$

(3) Let $P = \{[a, b] : a \leq b, a, b \in \mathbf{R}\}$ be the set of all closed intervals of real numbers. Define $[a, b] \leq [c, d]$ if $a \leq c \leq d \leq b$. Then $(P, \leq)$ is a dcpo by the compactness of closed intervals. An upper set U of P is Scott open iff for any $[a, b] \in U$ there is $[c, d] \in U$ such that $[a, b] \subseteq (c, d)$.

(4) If A is a discrete poset, i.e., $x \leq y$ in A only when $x = y$. Then every subset of A is Scott closed, so $C(P) = \mathcal{P}(A)$.

Definition 1 A complete lattice L is called a Scott closed set lattice, if there is a poset P such that L is order isomorphic to the lattice $C(P)$.

A basic problem on Scott closed lattices is: Which complete lattices are Scott closed set lattice?

An element r of a poset P is called a co-prime element if $x \vee y \geq a$ implies either $x \geq a$ or $y \geq a$ whenever $x \vee y$ exists. The set of all co-prime elements of P will be denoted by $\text{Spec}(P)$. By [14], a distributive complete lattice L is isomorphic to the lattice of all closed sets of a topological space if and only if $\text{Spec}(L)$ is join dense in L , i.e. every element of L is the join of all co-primes below it.

For any $x \in P$, the set $\downarrow x$ is a co-prime element of $C(P)$.

Example 2 Let L be the set of all closed sets of the real line $\mathbf{R}$. Is L a Scott closed set lattice? Note that the co-prime elements of L are exactly the singleton sets. Thus if P is a poset such that $C(P)$ is isomorphic to L , then for any two distinct $x, y \in P$, neither $\downarrow x \subseteq \downarrow y$ nor $\downarrow y \subseteq \downarrow x$, it follows that x and y are not comparable. Hence P must be a discrete poset, thus $C(P) = \mathcal{P}(P)$ which is

an atomic complete Boolean algebra. However, it is well known that L is not an atomic complete Boolean algebra. Hence L is not a Scott closed set lattice.

A mapping $f : P \rightarrow Q$ between posets is called monotone (or order-preserving) if $x \leq y$ in P implies $f(x) \leq f(y)$ in Q . Let POS be the category of posets and monotone mappings.

A mapping $f : P \rightarrow Q$ between posets is Scott continuous if for any directed subset $D \subseteq P$ with $\bigvee D$ exists in P , $f(\bigvee D) = \bigvee f(D)$ holds.

Let POS_d denote the category of all posets and Scott continuous mappings. Since every Scott continuous mapping is monotone, POS_d is a subcategory of POS .

The full subcategory of POS_d consisting of dcpos is denoted by $DCPO$.

Let P be a poset. For two elements a and b in P , we say that a is way-below b , denoted by $a \ll b$, if for any directed subset D of P , if $\bigvee D$ exists and $b \leq \bigvee D$ then there exists $d \in D$ such that $a \leq d$.

A poset P is called a continuous poset if for every element $a \in P$, the set $\{x : x \in P, x \ll a\}$ is a directed set and

$$a = \bigvee \{x : x \in P, x \ll a\}.$$

A continuous dcpo is also called a domain by many authors. If P is a continuous poset, then the way-below relation has the interpolation property: if $a \ll b$ then there exists x such that $a \ll x \ll b$. We refer to [6] for more about continuous posets.

Example 3 (1) For any any $A, B \subseteq X$, $A \ll B$ holds in $\mathcal{P}(X)$ iff A is a finite subset of B . Thus $(\mathcal{P}(X), \subseteq)$ is a continuous poset.

(2) From any poset P , one can construct a continuous poset. A subset I of poset P is called an ideal if I is a directed lower set. Let $\text{Idl}(P)$ be the set of all ideals of P . For any two ideals I and J , $I \ll J$ holds in $(\text{Idl}(P), \subseteq)$ iff $I \subseteq \downarrow x$ for some $x \in J$. It follows then that $(\text{Idl}(P), \subseteq)$ is a continuous poset.

(3) Let X and Y be two sets. A partial mapping f from X to Y is a mapping from a subset of X to Y . Let $PM(X, Y)$ denote the set of all partial mappings from X to Y . Let f, g be two partial mappings from X to Y , define $f \leq g$ if $\text{dom}(f) \subseteq \text{dom}(g)$ and $g|_{\text{dom}(f)} = f$, where $\text{dom}(f)$ and $\text{dom}(G)$ denote the domain of f and g respectively. Then $(PM(X, Y), \leq)$ is a poset. For any $f, g \in PM(X, Y)$, $f \ll g$ if and only if $\text{dom}(f)$ is a finite set and $f \leq g$. Thus $(PM(X, Y), \leq)$ is a continuous poset.

For a complete lattice L and any elements x, y in L , x is long way-below y , written $x \triangleleft y$, if $y \leq \bigvee D$ implies $x \leq d$ for some $d \in D$ for every subset $D \subseteq L$. By Raney [13], L is completely distributive if and only if $y = \bigvee \{x \in L : x \triangleleft y\}$ holds for every $y \in L$.

Lemma 1 *Let P be a poset.*

- (1) *If $\downarrow x \triangleleft \downarrow y$ holds in $C(P)$, then $x << y$ holds in P .*
- (2) *If P is continuous, then for any $A \subseteq P$, $cl(A) = \downarrow \{\bigvee E : E \subseteq A \text{ is directed and } \bigvee E \text{ exists}\}$, where $cl(A)$ is the closure of A with respect to the Scott topology;*
- (3) *If P is continuous, then $x << y$ in P implies $\downarrow x \triangleleft \downarrow y$ in $C(P)$.*

3. D-COMPLETION OF POSETS

For any poset P , $Idl(P)$ is a dcpo which is called the ideal completion. The mapping $\eta_P : P \rightarrow Idl(P)$, sending $x \in P$ to $\downarrow x$ in $Idl(P)$, is a universal monotone mapping from P into a dcpo. That is η_P is monotone and for every monotone mapping $f : P \rightarrow Q$ into a dcpo Q , there is a (Scott continuous) monotone mapping $\hat{f} : Idl(P) \rightarrow Q$ such that $f = \hat{f} \circ \eta_P$. The ideal completion is not a real dcpo completion because the ideal completion of a dcpo is not necessarily isomorphic to P .

In this section we present a dcpo completion of posets which are constructed in terms of the Scott closed set lattice. This completion is idempotent and it has closer relation with P than $Idl(P)$.

The proofs for the results in this section can be found in [15].

Definition 2 *A subset X of a dcpo P is called a subdcpo of P if for any directed subset $D \subseteq X$, $\bigvee D \in X$.*

Remark 1 (1) *The empty set and P itself are subdcpos. The intersection of any collection of subdcpos is a subdcpo.*

(2) *Every Scott closed set of P is a subdcpo of P .*

(3) *The union of two subdcpos is again a subdcpo. As a matter of fact, suppose that A and B are subdcpos of P and $D \subseteq P$ is a directed set such that $D \subseteq A \cup B$. Then either one of the following cases is true: (a) for any $x \in D$ there is $d \in D, d \in A$ and $x \leq d$; (b) for any $x \in D$ there is $d \in D, d \in B$ and $x \leq d$. If (a) is true, then $D \cap A$ is a directed set and $\bigvee D = \bigvee(D \cap A) \in A$. If (b) is true, then $\bigvee D \in B$.*

If X is a subset of P , we denote by $cl_d(X)$ the intersection of all subdcpos containing X .

For any $X \subseteq P$, the closure $cl(X)$ of X with respect to the Scott topology contains $cl_d(X)$ and thus $\downarrow cl_d(X) \subseteq cl(X)$. However, generally $cl(X) \neq \downarrow cl_d(X)$.

Lemma 2 *If $f : P \rightarrow Q$ is Scott continuous, then for each $X \subseteq P$, $f(cl_d(X)) \subseteq cl_d(f(X))$.*

Lemma 3 *For any dcpo P , $Spec(P)$ is a subdcpo.*

Corollary 1 *For any directed subset D of poset P ,*

$$cl(D) \in Spec(C(P)).$$

Let X be a subset of a dcpo P . For each ordinal α define a subset X^α as follows:

$$\begin{aligned} X^0 &= X, X^{\alpha+1} = \{\vee D : D \text{ is a directed subset of } X^\alpha\}, \\ X^\lambda &= \bigcup_{\alpha < \lambda} X^\alpha, \text{ if } \lambda \text{ is a limit ordinal.} \end{aligned}$$

Then due to the cardinal reason, there is an α such that $X^\alpha = X^{\alpha+1}$, and in this case it is clear that $cl_d(X) = X^\alpha$.

Remark 2 *From the above observation, it is easy to see that if X is a subset of dcpo P and $p \in cl_d(X)$, then $p \in cl_d(\downarrow p \cap X)$.*

Lemma 4 *If X is a subset of dcpo P and $f, g : cl_d(X) \rightarrow A$ are Scott continuous mappings into a dcpo A such that $f|_X = g|_X$, then $f = g$.*

A monotone mapping $f : P \rightarrow Q$ between posets is a lower adjoint of a monotone mapping $g : Q \rightarrow P$ if for any $x \in P$ and $y \in Q$, $f(x) \leq y$ if and only if $x \leq g(y)$ (here g is called the upper adjoint of f). Every lower adjoint preserves existing joins and every upper adjoint preserves existing meets (see [6]). If $f : L \rightarrow M$ is a mapping between complete lattices that preserves arbitrary infimum, then f has a lower adjoint [6].

Lemma 5 *If $f : P \rightarrow Q$ has an upper adjoint that preserves existing finite joins, then $f(Spec(P)) \subseteq Spec(Q)$.*

Definition 3 *A D -completion of a poset P is a dcpo A together with a Scott continuous mapping $\eta : P \rightarrow A$, such that for any Scott continuous mapping $f : P \rightarrow B$ into a dcpo B there exists a unique Scott continuous mapping $\hat{f} : A \rightarrow B$ satisfying $f = \hat{f} \circ \eta$.*

By a standard argument, it can be shown that if the D -completion of P exists, it is unique up to isomorphic. We shall use $E(P)$ to denote the D -completion of P if it exists.

Theorem 1 *Let P be any poset. Then the smallest subdcpo of $C(P)$ containing the subset $\Psi(P) = \{\downarrow x : x \in P\}$ is a D -completion of P .*

Corollary 2 *The subcategory $DCPO$ of POS_d consisting of all dcpos is reflexive in POS_d .*

Example 4 (1) *Let $P = \mathbb{Q}$ be the set of all rational numbers with the ordinary order of numbers. Then $\Gamma(P)$ is isomorphic to the poset $\mathbb{R}^* = \mathbb{R} \cup \{\infty\}$ of all real*

numbers associated with the top element ∞ . Furthermore, $E(P) = \Gamma(P) \cong \mathbb{R}^*$, which is different from the set $\text{Id}(P)$ of all ideals of P .

In general, let P be a chain, then $E(P) \cong \mathcal{D}_0(P)$, where $A \in \mathcal{D}_0(P)$ if and only if it is a lower set and either $\bigvee A$ does not exist in P , or $A = \downarrow a$ when $\bigvee A = a$ exists in P .

(2) Let X be a nonempty set and $\mathcal{P}_0(X)$ be the set of all finite subsets of X . Then $\mathcal{P}_0(X)$ is a poset under the order of inclusion, which is not a dcpo except when X is a finite set. Note that a directed subset of $\mathcal{P}_0(X)$ has a supremum only when it is a finite set. Thus $\Gamma(\mathcal{P}_0(X)) = \mathcal{D}(\mathcal{P}_0(X))$, the set of all lower sets of $\mathcal{P}_0(X)$. It is easy to see that $E(\mathcal{P}_0(X)) = \text{Id}(\mathcal{P}_0(X))$, the set of all ideals of $\mathcal{P}_0(X)$.

(3) Let $P = \mathcal{P}_N(X)$ be the set of all countable subsets of X . Then P is not a dcpo under the order of inclusion unless X is a countable set. If X is not countable, then $\Gamma(P) \neq \text{Id}(P)$.

Recall that if $f : P \rightarrow Q$ is lower adjoint to the mapping $g : Q \rightarrow P$, then f is surjective if and only if g is injective(see Proposition O-3.7[6]).

Lemma 6 Suppose that A is a subset of a dcpo P such that $P = \text{cl}_d(A)$. Then for any $F \in \Gamma(P)$, $F = \text{cl}(F \cap A)$.

The following result reveals that a poset P and its D-completion have isomorphic Scott closed set lattices.

Theorem 2 If $E(P)$ is a D-completion of poset P , then $C(P)$ is isomorphic to $\Gamma(E(P))$.

A meet-semilattice P is a poset that has a top element, such that $x \wedge y$ exists for any two elements x, y in P . A semilattice S is meet-continuous if

$$x \wedge \bigvee D = \bigvee \{x \wedge y : y \in D\}$$

holds for every directed subset D whose join exists. A complete lattice L is a frame if

$$a \wedge \bigvee B = \bigvee \{a \wedge x : x \in B\}$$

holds for every $a \in L$ and $B \subseteq L$ (see [10] for more about frames, or locales).

Theorem 3 Let P be a semilattice. Then the following conditions are equivalent:

- (1) P is meet-continuous;
- (2) $E(P)$ is meet-continuous;
- (3) $C(P)$ is a frame.
- (4) $C(E(P))$ is a frame.

Theorem 4 For every poset P , the following statements are equivalent:

- (1) P is continuous;
- (2) $C(P)$ is completely distributive under the inclusion order.
- (3) $E(P)$ is a continuous dcpo.

In [12] Mislove defined a functor $BSpec : POS_d \rightarrow LD$ from the category POS_d to the category LD of local dcpo and claims that this functor is left adjoint to the forgetful functor from LD to POS_d .

A poset P is called a local dcpo (or bounded complete dcpo) if every upper bounded directed subset has a supremum [12]. Let LD denote the full subcategory of POS_d consisting of all local dcpos.

For each poset P , let $BSpec(P)$ denote the set of all upper bounded co-primes of $C(P)$. The poset $BSpec(P)$ is a local dcpo [12]. Note that if L is a complete lattice, then every co-prime element of $C(L)$ is upper bounded, thus $BSpec(P) = Spec(C(L))$. If the object map $P \mapsto BSpec(P)$ extends to a functor which is left adjoint to the forgetful functor $LD \rightarrow POS$, then $Spec(C(P)) = BSpec(L) = \{\downarrow x : x \in L\}$. This may not be true if the Scott space of L is not a sober space. In [9], Isbell constructed a complete lattice whose Scott space is not sober. Thus the map $P \mapsto BSpec(P)$ may not define a left adjoint to the forgetful functor.

For a poset P , let $BE(P) = \{F \in E(P) : F \text{ is upper bounded}\}$. Then obviously $\Psi(P) = \{\downarrow x : x \in P\} \subseteq BE(P)$.

Lemma 7 *For every poset P , $BE(P)$ is a local dcpo.*

Lemma 8 *If P is a local dcpo, then $BE(P) = \{\downarrow x : x \in P\}$.*

Theorem 5 *The category LD is fully reflexive in POS_d .*

4. EQUIVALENCE BETWEEN CATEGORIES OF COMPLETELY DISTRIBUTIVE LATTICES AND DOMAINS

A mapping $f : P \rightarrow Q$ between two dcpos is said to preserve the way-below relation if for any $x, y \in P$, $x \ll y$ implies $f(x) \ll f(y)$. Let **CONP** be the category of continuous dcpos and Scott continuous mappings that preserve the way-below relation. A mapping $f : L \rightarrow M$ between complete lattices is said to preserve the relation $\triangleleft$ if for any $x, y \in L$, $x \triangleleft y$ implies $f(x) \triangleleft f(y)$. Let **CLAT** be the category of all complete lattices and mappings that preserve arbitrary sups and $\triangleleft$.

Recall that for a complete lattice L , $Spec(L)$ denotes the set of all co-prime elements of L .

Lemma 9 *If P is a continuous poset and $F \in C(P)$, then $F \in Spec(\Gamma(P))$ if and only if $F = cl(D)$ for some directed subset D of P .*

Proof If D is directed, then by Corollary 1 $cl(D)$ is co-prime.

Conversely, suppose that $F \in Spec(C(P))$. Since P is continuous, by Theorem 4 $C(P)$ is a completely distributive lattice. Thus it follows that $F = \bigvee \{\downarrow x : \downarrow x \triangleleft F\}$. By Lemma 1 of [18], $\{\downarrow x : \downarrow x \triangleleft F\}$ is a directed set, therefore $D = \{x : \downarrow x \triangleleft F\}$ is directed and, obviously $cl(D) = F$.

Corollary 3 *If P is a continuous dcpo, then $\text{Spec}(C(P)) = \{\downarrow x : x \in P\}$.*

For any morphism $f : P \rightarrow Q$ in **CONP**, define the mapping $C(f) : C(P) \rightarrow C(Q)$ be the mapping such that $C(f)(A) = cl(f(A))$ for all $A \in C(P)$. Obviously $C(f)$ is lower adjoint to the mapping that sends $B \in C(Q)$ to $f^{-1}(B)$.

Lemma 10 *For any morphism $f : P \rightarrow Q$ in **CONP**, $C(f) : C(P) \rightarrow C(Q)$ is a morphism in **CLAT**.*

Proof Suppose $A, B \in C(P)$ and $A \triangleleft B$. Since P is continuous, $C(P)$ is completely distributive. So there exist $x, y \in B$ such that $A \subseteq \downarrow x \triangleleft \downarrow y \subseteq B$. Then by Lemma 1, $x \ll y$ holds in P . Thus $f(x) \ll f(y)$. Then $C(A) = cl(f(A)) \subseteq \downarrow f(x) \subseteq \downarrow f(y) \subseteq cl(f(B))$. Again, by Lemma 1, $\downarrow f(x) \triangleleft \downarrow f(y)$. Hence $C(f)(A) \triangleleft C(f)(B)$. Since $C(f)$ is a lower adjoint, it preserves arbitrary joins, thus it is a morphism in **CLAT**.

Lemma 11 [18] *Let L be a completely distributive lattice. If $A \subseteq \{x \in L : x \triangleleft a\}$ such that $\sup A = a$, then $a \in \text{Spec}(L)$ if and only if A is a directed set.*

Lemma 12 *Let L be a completely distributive lattice. For any two co-primes a and b , $a \ll b$ holds in $\text{Spec}(L)$ iff $a \triangleleft b$ holds in L .*

Proof Since $\text{Spec}(L)$ is a sub dcpo of L , thus $a \triangleleft b$ implies $a \ll b$. Conversely, let $a \ll b$ in $\text{Spec}(L)$. By 11, the set $\{y \in \text{Spec}(L) : y \triangleleft b\}$ is directed and has a as its supremum. Thus $b = \sup\{y \in \text{Spec}(L) : y \triangleleft b\}$ holds in $\text{Spec}(L)$. Then there is $y \in \text{Spec}(L)$, $y \triangleleft b$ such that $a \leq y$. It thus follows that $a \triangleleft b$.

Lemma 13 *For any completely distributive lattice L , $\text{Spec}(L)$ is a continuous dcpo with respect to the inherited order.*

Lemma 14 *For any morphism $g : L \rightarrow M$ in **CLAT**, the restriction $g|_{\text{Spec}(L)} : \text{Spec}(L) \rightarrow \text{Spec}(M)$ is a morphism in **ConPOSET**.*

Proof Let $a \in \text{Spec}(L)$. By 11, $\{x \in L : x \triangleleft a\}$ is directed and has a as its supremum. Thus $g(a) = \sup\{g(y) : y \triangleleft a\}$. The set $\{g(y) : y \triangleleft a\}$ is directed and is contained in $\{u \in Q : u \triangleleft g(a)\}$ because g preserves $\triangleleft$. Thus by 11, $g(a) \in \text{Spec}(Q)$. Also $\text{Spec}(L)$ is closed under direct supremum, $g|_{\text{Spec}(L)}$ preserves sups of directed sets, therefore it is a morphism in **ConPOSET**.

It follows from the above lemmas that we have two functors:

$$F : \text{CONP} \rightarrow \text{CDLAT} \text{ and } G : \text{CDLAT} \rightarrow \text{CONP},$$

where $F(P) = C(P)$, $F(f)(A) = cl(f(A))$ for any continuous dcpo P and Scott continuous $f : P \rightarrow Q$, and $G(L) = \text{Spec}(L)$ and $G(g)$ is the restriction of g on $\text{Spec}(L)$ for any homomorphism $g : L \rightarrow M$ in **CDLAT**.

Intuitively, completely distributive lattices form a much smaller part of the class of continuous dcpo's. However the two categories formed by them are equivalent. This result was first proved independently by R. Hoffmann and J. Lawson (see [8], [11]).

Theorem 6 *The functor $F : \text{CONP} \rightarrow \text{CDLAT}$ establishes an equivalence between the two categories CONP and CDLAT .*

This result can be extended to more general structure [4].

5. THE HOARE POWERDOMAIN

In mathematics, one often needs to consider "power structure" from a given structure. For example, the power set $\mathcal{P}(X)$ of a set X , the lattice $\text{Id}(P)$ of all ideals of a poset P , the exponential space $C(X)$ of a topological space X (the set $C(X)$ of all closed sets of X with the Vietoris topology), the lattice $\text{Sub}(H)$ of all closed subspaces of a Hilbert space H .

In domain theory, one can construct the power structures–powerdomains, in several ways.

An signature of finitary algebras is a set Σ (of operation symbols) together with an arity function assigning to each operation symbol α in Σ a non-negative integer– the arity of α .

Given a signature Σ of finitary algebras, A Σ -algebra is a set A such that for each $\sigma \in \Sigma$, there is a mapping

$$\alpha : A^n \longrightarrow A,$$

where n is the arity of α .

Definition 4 [6] *A directed complete partially ordered Σ -algebra, or a dcpo-algebra, is a dcpo that is also a Σ -algebra for which all the operations are Scott continuous (from the appropriate products equipped with the Scott topology). A homomorphism is a function between dcpo-algebras of the same signature that is Scott continuous and a homomorphism for each of the operations.*

Given any set X and any signature Σ , there is a free Σ -algebra over X , $T_\Sigma(X)$, consisting of terms that can be built up recursively from X by formally applying the various operations in Σ . Every mapping $f : X \rightarrow A$ from X to a Σ -algebra extends uniquely to an algebra homomorphism from $T_\Sigma(X)$ into A .

An equation (inequality) is of the form $\tau_1 = \tau_2$ ($\tau_1 \leq \tau_2$), where τ_1, τ_2 are terms in $T_\Sigma(X)$.

A dcpo-algebra with signature Σ is said to satisfy an equation $\tau_1 = \tau_2$ (in equality $\tau_1 \leq \tau_2$) if for any mapping $f : X \rightarrow A$, if the extension homomorphism of f is $\bar{f} : T_\Sigma(X) \rightarrow A$, then $\bar{f}(\tau_1) = \bar{f}(\tau_2)$ ($\bar{f}(\tau_1) \leq \bar{f}(\tau_2)$) holds in A .

Given a family $\mathcal{E}$ of equations and inequalities, denote by $\text{DCPO}(\Sigma, \mathcal{E})$ the category of dcpo-algebras with signature Σ that satisfy all equations and inequalities in $\mathcal{E}$.

Let Σ be the signature consists of a single binary operation, denoted by $\uplus$. Let $\mathcal{E}$ be the inequality

- (i) $x \leq x \uplus y$, and equations
- (ii) $x \uplus y = y \uplus x$,
- (iii) $x \uplus (y \uplus z) = (x \uplus y) \uplus z$,
- (iv) $x \uplus x = x$.

Then an object of $DCPO(\Sigma, \mathcal{E})$ is called inflationary semilattice[6].

Let X be a dcpo. A pair (A, p) is the free algebra over X with respect to $DCPO(\Sigma, \mathcal{E})$, if A is a inflationary semilattice and $p : X \rightarrow A$ is a Scott continuous mapping such that every Scott-continuous $g : X \rightarrow B$ to an inflationary semilattice B extends uniquely to a dcpo-morphism $k : A \rightarrow B$. The free inflationary semilattice over a domain P is called the Hoare powerdomain of P and denoted by $\mathbf{P}^H(P)$.

Proposition 1 [6] *Let X be a topological space, and $\Psi(X)$ be the poset of all nonempty closed sets of X ordered by the inclusion relation, and $p : X \rightarrow \Psi(X)$ be defined by $p(x) = cl(\{x\})$. For any continuous $f : X \rightarrow L$ into an inflationary semilattice (with the Scott topology), there is a unique dcpo-morphism $k : \Psi(X) \rightarrow L$ such that $f = k \circ p$.*

Now if P is a dcpo. Then P is a topological space with the Scott topology. By the above proposition we obtain

Theorem 7 *Let P be a dcpo. Then the free inflationary semilattice over P consists of $C_0(P)$ of all nonempty Scott closed sets of P with binary union as the operation, inclusion relation as the order and the embedding of P given by $p(x) = \downarrow x$.*

Corollary 4 [6] *If L is a domain, then its Hoare powerdomain $\mathbf{P}^H(L)$ is again a domain.*

For other kinds of powerdomains see Chapter IV of [6].

6. SCOTT CLOSED LATTICES OF COMPLETE SEMILATTICES

A dcpo P is a continuous poset if and only if its Scott closed set lattice $C(P)$ is a complete distributive lattice. In addition, every completely distributive lattice L is isomorphic to $C(P)$ for some continuous dcpo P . A natural question is: what are the lattices $C(P)$ for noncontinuous poset P ? In this section we give the characterizations of $C(P)$ for complete lattice and complete semilattice P . For details on this part, see [7].

Definition 5 *Let L be a complete lattice and $x, y \in L$. Define $x \prec y$, if for every nonempty Scott-closed set $E \subseteq P$, the relation $\bigvee C > y$ always implies that $x \in E$.*

Lemma 15 *Let L be a complete lattice.*

- (i) $x \prec y$ implies $x \leq y$ for any $x, y \in L$.
- (ii) $x \leq y \prec z \leq w$ implies $x \prec w$ for any x, y, z and w in L .
- (iii) For any $a \in L$, $\{x \in L : x \prec a\}$ is a Scott closed subset of L .

Definition 6 [7] A complete lattice L is called C -continuous if for any $a \in L$,

$$a = \bigvee \{x \in L : x \prec a\}.$$

Lemma 16 Let P be a dcpo. Then for any Scott closed set $\mathcal{F}$ of the complete lattice $C(P)$ we have

$$\sup_{C(P)} \mathcal{F} = \bigcup \mathcal{F}.$$

Let $x \in P$ be an element of a dcpo P . For any $\mathcal{F} \in C(C(P))$, if $\downarrow x \subseteq \sup_{C(P)} \mathcal{F}$, then by the above lemma, $\downarrow x \subseteq \bigcup \mathcal{F}$, thus $x \in F$ for some $F \in \mathcal{F}$. This then implies $\downarrow x \subseteq F$, and so $\downarrow x \in \mathcal{F}$ because $\mathcal{F}$ is a lower set of $C(P)$.

Corollary 5 For any $x \in P$, where P is a dcpo, $\downarrow x \prec \downarrow x$ holds in $C(P)$.

Definition 7 An element x of a complete lattice is called C -algebraic if $x \prec x$. The set of all C -algebraic elements of L will be denoted by $K_C(L)$.

Definition 8 A complete lattice L is called C -prealgebraic if for any $a \in L$, $a = \bigvee \{x \in L : x \in K_C(L), x \leq a\}$.

L is called a C -algebraic lattice if it is C -prealgebraic and, for any $x \in L$, $\downarrow x \cap K_C(L)$ is a Scott closed set of L .

Corollary 6 For any dcpo P , $C(P)$ is C -prealgebraic.

A complete semilattice L is a dcpo such that every upper bounded subset of L has a supremum in L .

Example 5 (1) $(\Gamma_0([0, 1]), \subseteq)$ is a complete semilattice which is not a lattice, where $\Gamma_0([0, 1])$ is the set of all nonempty closed subsets of $[0, 1]$ and $A \leq B$ iff $B \subseteq A$.

(2) Let X be a set and $\text{End}(X)$ be the set of all mappings $f : X \rightarrow X$. Define $f \leq g$ for $f, g \in \text{End}(X)$ if for any $x \in X$, $f(x) \neq x$ implies $f(x) = g(x)$. Then $(\text{End}(X), \leq)$ is a complete semilattice which is not a lattice provided that X is a singleton set (see [17]).

(3) If M is a subset of a complete semilattice L such that M is closed under existing sups, then M is a complete semilattice under the order inherited from L . For instance, let $\text{End}_0(N)$ be the set of all $f : N \rightarrow N$ such that $N - \text{Fix}(f)$ is an infinite set, where $\text{Fix}(f)$ is the set of fixed points of f . Then $\text{End}_0(N)$ is a complete semilattice under the inherited order from $\text{End}(N)$.

Theorem 8 For any complete semilattice L , $K_C(C(L)) = \{\downarrow x : x \in L\}$.

Since the sub set $\{\downarrow x : x \in L\}$ of $C(L)$ is isomorphic to L , it follows that we can recover L from $C(L)$, up to isomorphism, as the set of all C -compact elements.

Corollary 7 Let P and Q be two complete semilattices such that $C(P)$ and $C(Q)$ are isomorphic as posets. Then P and Q are isomorphic as posets.

Remark 3 Let P be a poset that is not a dcpo. Then By Theorem 2, $C(P)$ and $C(E(P))$ are isomorphic, where $E(P)$ is the D-completion of P . However P and $E(P)$ are not isomorphic because $E(P)$ is a dcpo and P is not.

Definition 9 A complete lattice L is called *C-stable* if

- (i) $1_L \prec 1_L$, and
- (ii) for any $x \in L$ and a Scott closed set $D \subseteq L$ such that $x \prec y$ for all $y \in D$, then $x \prec \bigwedge D$.

A complete lattice satisfies only condition (ii) is called a *weakly C-stable lattice*.

The following theorem gives a complete characterization of Scott lattices of complete lattices and complete semilattices.

Theorem 9 Let M be a complete lattice.

- (i) M is isomorphic to $C(L)$ for a complete lattice L iff M is a C-stable and C-algebraic lattice.
- (ii) M is isomorphic to $C(L)$ for a complete semilattice L iff M is a weakly C-Stable and C-algebraic lattice.

7. SOME PROBLEMS

In section 3, we presented a dcpo completion of posets via Scott closed set lattices. It would be meaningful to describe the dcpo completion of some of the concrete posets. One such example is the poset of all continuous functions on a topological space. Let X be a topological space and $\mathbf{R}^* = \mathbf{R} \cup \{+\infty\}$ be the extended real line. A function $f : X \rightarrow \mathbf{R}^*$ is lower semicontinuous if $f^{-1}(r, +\infty)$ is open for any $r \in \mathbf{R}$. Let $LSC(X, \mathbf{R}^*)$ be the set of all lower semicontinuous functions and $C(X, \mathbf{R}^*)$ be the set of all continuous functions from X to $\mathbf{R}^*$. By [6], $SCL(X, \mathbf{R}^*)$ is a continuous lattice. Also by Proposition I-1.22.1.(iii) [6], for any $f, g \in LSC(X, \mathbf{R}^*)$, $f << g$, if and only if there is a $h \in C(X, \mathbf{R}^*)$ such that $f(x) \leq h(x) << g$ for all $x \in X$. It then follows that $C(X, \mathbf{R}^*)$ is a continuous poset that is not a dcpo.

Problem 1. Study the D-completion $E(C(X, \mathbf{R}^*))$ of $C(X, \mathbf{R}^*)$.

Is $LSC(X, \mathbf{R}^*)$ a D-completion of $C(X, \mathbf{R}^*)$?

Let P be a continuous dcpo. If Q is a dcpo such that $C(Q)$ is isomorphic to $C(P)$, then $C(Q)$ is completely distributive, thus Q is also continuous. It then follows that P is isomorphic to the $Spec(C(P))$ and, similarly, Q is isomorphic to $Spec(C(Q))$. Therefore P is isomorphic to Q . It is natural to ask if the converse of the above is true.

Problem 2. Let P be a dcpo such that for any dcpo Q , $C(P) \cong C(Q)$ implies $P \cong Q$. Must P be a continuous dcpo?

Let **SCLAT** be the class of all Scott closed set lattices. To have a better understanding of Scott closed set lattices, it is useful to investigate the collective properties of **SCLAT**.

Problem 3 Is **SCLAT** closed under cartesian product?

REFERENCES

1. S. Abramsky and A. Jung, *Domain Theory*, Volume 3 of Handbook for Logic in Computer Science, Clarendon Press 1994.
2. K. V. Adaricheva, V. A. Gorbunov and M. V. Semenova, On noncomplete continuous lattices, *Algebra Universalis*, (2001), 1-2:215-230.
3. B. Banaschewski, On the topologies of Injective spaces, Continuous lattices and their applications, Lecture Notes in Pure and Applied Mathematics, 101(1985), 1-8.
4. B. Ern  and D. Zhao, Z-join spectra of Z-supercompactly generated lattices, *Applied categorical Structures*, 9(2001), 41-63.
5. R. C. Flagg and P. Slunderhauf, The essence of ideal completion in quantitative form, *Theoretical Computer Science*, (2002), 278(1-2): 14-158.
6. G. Gierz, K.H. Hoffmann, K. Keimel, J.D. Lawson, M.W. Mislove, and D.S. Scott, *Continuous Lattices and Domains*, Cambridge University Press, 2003.
7. W. Ho, and D. Zhao, On the characterization of Scott-closed set lattices, (2007) (Preprint)
8. R. E. Hoffmann, Continuous posets-prime spectra of completely distributive lattices, and Hausdorff compactification, Lecture Note in Mathematics, 871(1981), 159-208.
9. J. R. Isbell, Completion of a construction of Johnstone, *Proc. Amer. Math. Soc.*, 85(1982), 333-334.
10. P. T. Johnstone, Scott is not always sober, In Continuous Lattices, Springer LNM 871(1981), 282-283.
11. J. D. Lawson, The duality of continuous posets, *Houston Journal of Mathematics*, 5(1979), 357-394.
12. M. W. Mislove, Local DCPOs, Local CPOs and Local completions, *Electronic Notes in Theoretical Computer Science*, 20(1999).
13. G. N. Raney, Completely distributive lattices, *Proc. Amer. Math. Soc.*, 3(1952), 677-680.
14. S. Papert, Which distributive lattices are lattices of closed sets? *Proc. Cam. Philos. Soc.*, **55**(1959), 172-6.
15. D. Zhao and T. Fan, Dcpo-completion of posets, Preprint(2007).
16. B. Zhao and D. Zhao, The lim-inf convergence on partially ordered sets, *J. Mathematical Analysis and its applications*, 309(2005), 701-708.

17. D. Zhao, On reflexive subobject lattices and reflexive endomorphism algebras, *Comment. Math. Univ. Carolinae*, 44(2003), 23-32.
18. D. Zhao, N-compactness in L-fuzzy topological spaces, *J. Mathematical Analysis and its applications*, 128(1987), 64-79.

ZHAO DONGSHENG: Mathematics and Mathematics Education, National Institute of Education, Nanyang Technological University, Singapore.
E-mail: dongsheng.zhao@nie.edu.sg.

Multicriteria Dynamic Optimization Problems and Cooperative Dynamic Games

JACOB ENGWERDA

Abstract. We survey some recent research results in the field of dynamic cooperative differential games with non-transferable utilities. Problems which fit into this framework occur for instance if a person has more than one objective he likes to optimize or if several persons decide to combine efforts in trying to realize their individual goals. We assume that all persons act in a dynamic environment and that no side-payments take place. For these kind of problems the notion of Pareto efficiency plays a fundamental role. In economic terms, an allocation in which no one can be made better-off without someone else becoming worse-off is called Pareto efficient. In this paper we present as well necessary as sufficient conditions for existence of a Pareto optimum for general non-convex games. These results are elaborated for the special case that the environment can be modeled by a set of linear differential equations and the objectives can be modeled as functions containing just affine quadratic terms. Furthermore we will consider for these games the convex case. In general there exists a continuum of Pareto solutions and the question arises which of these solutions will be chosen by the participating persons. We will flash some ideas from the axiomatic theory of bargaining, which was initiated by Nash [16, 17], to predict the compromise the persons will reach.

Key words and Phrases: Dynamic Optimization, Pareto Efficiency, Cooperative Differential Games, LQ Theory, Riccati Equations, Bargaining

INTRODUCTION

In this paper we consider the problem to find an "optimal" strategy in case either there is one individual who has multiple objectives, or, there is more than one person affecting a dynamic system and in order to minimize their cost these persons decide to coordinate their actions. As an example one can think of a government who likes to realize an acceptable income level, full-fledged employment for everyone, a just income distribution, a stable price level, a stable currency

2000 Mathematics Subject Classification:

exchange rate, a sustainable balance of payment, and a lasting economic growth. One can imagine that the realization of all these goals simultaneously will be a tough or, more probably, impossible task. From that perspective it is more realistic to classify all simultaneously feasible realizations of the objectives and next let the government choose one of these options. An example illustrating a situation where more than one player is involved in a game and the players cooperate in order to realize their objectives is for instance the next advertising game (see [13]). Consider two competing divisions in a conglomerate company which wish to maximize their individual profits by choosing an optimal advertising policy. Being divisions of a parent company, they are not out to "hurt" each other. Therefore it seems reasonable to assume that they will coordinate their policies in order to maximize their profits.

The above examples are characterized by the fact that there is either one or more player who can affect the outcome of the state of a system. We will assume that the evolution of this system over time is described by a differential equation. Since players simultaneously affect the outcome of the system the performance of the players depends on the actions taken by the other players. To realize an outcome which is as good as possible for all of them simultaneously, we assume in this paper that players decide to coordinate their actions and engage in a cooperative game. It is assumed that all players can communicate and enter into binding agreements. However, no side-payments take place. Moreover, it is assumed that every player has all information on as well the state dynamics as the cost functions of his opponents, and all players are able to implement their decisions. In literature this information structure is known as the open-loop information game. Concerning the strategies used by the players we assume that there are no restrictions. That is, every control trajectory chosen by player i , $u_i(\cdot)$, may be chosen arbitrarily from a set $\mathcal{U}$ (which depends on the problem setting and will be specified later) in order to have a well-posed problem. More formally, we consider the minimization of the performance criteria:

$$J_i(t_0, u_1, u_2) := \int_{t_0}^T g_i(t, x(t), u_1(t), u_2(t)) dt + h_i(x(T)), \quad i = 1, 2, \quad (1)$$

where $x(t)$, is the solution of the differential equation

$$\dot{x}(t) = f(t, x(t), u_1(t), u_2(t)), \quad x(t_0) = x_0. \quad (2)$$

For the moment we do not explicitly state conditions on the functions g_i and f and the set of control actions $\mathcal{U}$. We just assume that they are such that the above integrals are well-defined and the differential equation has a solution on $[t_0, T]$ in the extended sense (see e.g. [4], [12] or [5]). Later on, when needed, additional assumptions will be made on these functions and set. Furthermore we make the convention that in case we consider this problem for the initial time $t_0 = 0$, we use the shorthand notation $J(u_1, u_2)$ instead of (1).

By cooperation, in general, the cost a specific player incurs is not uniquely determined anymore. If, within the above two-player context, both players decide

for instance to use their control variables to minimize the performance of player 1 as much as possible, a different minimum is attained for player 1 than in case both players agree to help collectively player 2 to minimize his performance. So, depending on how the players choose to "divide" their control efforts, a player incurs different "minima". Therefore, in general, each player is confronted with a whole set of possible outcomes from which somehow one outcome (which in general does not coincide with a player's overall lowest cost) is cooperatively selected. Now, if there are two control strategies u_1 and u_2 such that every player has a lower cost if strategy u_1 is played, then it seems reasonable to assume that all players will prefer this strategy. We say that the solution induced by strategy u_1 *dominates* the solution induced by the strategy u_2 . So, dominance means that the outcome is better for all players. Proceeding in this line of thinking it seems reasonable to consider only those cooperative outcomes which have the property that, if a different strategy than the one corresponding with this cooperative outcome is chosen then at least one of the players has higher costs. Or, stated differently, to consider only solutions that are such that they can not be improved upon by all players simultaneously. This motivates the concept of Pareto efficiency.

Definition 0.1 Let $\mathcal{U}$ denote the set of admissible strategies. A set of control actions $\hat{u}$ (in the sequel: a control) is called *Pareto efficient* if the set of inequalities

$$J_i(u) \leq J_i(\hat{u}), \quad i = 1, \dots, N,$$

where at least one of the inequalities is strict, does not allow for any solution $u \in \mathcal{U}$. The corresponding point $(J_1(\hat{u}), \dots, J_N(\hat{u})) \in \mathbb{R}^N$ is called a *Pareto solution*. The set of all Pareto solutions is called the *Pareto frontier*. $\square$

A Pareto solution is therefore never dominated, and for that reason called an *undominated* solution. Usually there is more than one Pareto solution, because dominance is a property which generally does not provide a total ordering. Lemma 0.2, below, gives a sufficient condition under which one can conclude that a control is Pareto efficient. This result is well-known in literature (see e.g. [13], [5] or [31]). This lemma holds without using any convexity conditions on the J'_i 's nor any convexity assumptions regarding the strategy space. Its converse, Theorem 0.13, was proved by Fan et al in [9] under some convexity assumptions on the cost functions. This result was exploited in [7] to solve the regular cooperative indefinite linear quadratic game. In this paper we present in Section 2 for general non-convex problems necessary conditions for a control to be Pareto efficient and next discuss additional conditions from which one can conclude that these necessary conditions are sufficient too. As far as we know these conditions have not been stated explicitly in the literature. The obtained results resemble the corresponding results for the static optimization problem as reported, e.g., in [22]. Section 3 deals with the regular, non-convex, indefinite finite-planning horizon linear quadratic differential game. Most results are presented here for a finite planning horizon. In particular we present here an algorithm to compute all Pareto efficient solutions when the system is scalar. Unfortunately, it remains unclear whether this algorithm can also

be used for non-scalar systems to find all Pareto efficient solutions. In Section 4 we consider the special case that both the set of admissible control functions and cost functions are convex. We show that in that case the algorithm presented in Section 3 can be used too to calculate all Pareto efficient solutions as well in case the planning horizon is finite as in case it is infinite.

Assuming an open-loop information structure it seems reasonable that players will also make a decision at the start of the planning horizon which Pareto efficient control they will use. In section 5 we present a number of candidate solutions that are motivated from the axiomatic bargaining literature. Section 6 concludes and mentions a number of open problems. This overview is to a large extent based on results reported in [5], [6], [7] and [8]. This paper contains some more detailed results and proofs.

THE GENERAL CASE

In this section we first present some preliminary results and properties about Pareto efficient solutions (frontier) which hold in general. After this, we present then the necessary conditions followed by a discussion of some sufficient conditions. In the subsequent analysis the following set of parameters, $\mathcal{A}$ (the "unit-simplex"), plays a crucial role.

$$\mathcal{A} := \{\alpha = (\alpha_1, \dots, \alpha_N) \mid \alpha_i \geq 0 \text{ and } \sum_{i=1}^N \alpha_i = 1\}.$$

The following two lemmas provide a characterization of Pareto efficient controls. To emphasize the role of the assumptions, we include of both results a proof which is a straightforward copy of the proof for the finite dimensional case (see e.g. [5] and [22]).

Lemma 0.2 *Let $\alpha_i \in (0, 1)$, with $\sum_{i=1}^N \alpha_i = 1$. Assume $\hat{u} \in \mathcal{U}$ is such that*

$$\hat{u} \in \arg \min_{u \in \mathcal{U}} \left\{ \sum_{i=1}^N \alpha_i J_i(u) \right\}. \quad (3)$$

Then $\hat{u}$ is Pareto efficient. □

Proof. Let $\alpha_i \in (0, 1)$, with $\sum_{i=1}^N \alpha_i = 1$ and $\hat{u} \in \arg \min_{u \in \mathcal{U}} \left\{ \sum_{i=1}^N \alpha_i J_i(u) \right\}$. Assume $\hat{u}$ is not Pareto efficient. Then, there exists an N -tuple of strategies $\bar{u}$ such that

$$J_i(\bar{u}) \leq J_i(\hat{u}), \quad i = 1, \dots, N,$$

where at least one of the inequalities is strict. But then

$$\sum_{i=1}^N \alpha_i J_i(\bar{u}) < \sum_{i=1}^N \alpha_i J_i(\hat{u}),$$

which contradicts the fact that $\hat{u}$ is minimizing. $\square$

Lemma 0.3 *$\hat{u} \in \mathcal{U}$ is Pareto efficient if and only if for each i $\hat{u}(\cdot)$ minimizes J_i on the constrained set*

$$\mathcal{U}_i := \{u \mid J_j(u) \leq J_j(\hat{u}), j = 1, \dots, N, j \neq i\}, \text{ for } i = 1, \dots, N. \quad (4)$$

Proof. " $\Rightarrow$ " Suppose $\hat{u}$ is Pareto efficient. If $\hat{u}$ does not minimize J_k on the constrained set $\mathcal{U}_k$ for some k , then there exists a u such that $J_j(u) \leq J_j(\hat{u})$ for all $j \neq k$ and $J_k(u) < J_k(\hat{u})$. This contradicts the Pareto efficiency of $\hat{u}$.
 " $\Leftarrow$ " Suppose $\hat{u}$ minimizes each J_k on $\mathcal{U}_k$. If $\hat{u}$ does not provide a Pareto optimum, then there exists a $u(\cdot) \in \mathcal{U}$ and an index k such that $J_i(u) \leq J_i(\hat{u})$ for all i and $J_k(u) < J_k(\hat{u})$. This contradicts the minimality of $\hat{u}$ for J_k on $\mathcal{C}_k$. $\square$

A direct consequence of this lemma is that in the two-player case the Pareto frontier can be, loosely speaking, visualized as a decreasing function.

Corollary 0.4 ($N = 2$) *Assume that the Pareto frontier consists of more than one point. If both $\hat{u}$ and $\tilde{u}$ are two Pareto efficient controls with $J_1(\hat{u}) \leq J_1(\tilde{u})$, then $J_2(\hat{u}) \geq J_2(\tilde{u})$.*

Proof. Assume that $\hat{u}$ and $\tilde{u}$ are two Pareto efficient controls with $J_1(\hat{u}) \leq J_1(\tilde{u})$. Let $\tilde{\mathcal{U}}_1 := \{u \mid J_1(u) \leq J_1(\hat{u})\}$ and $\tilde{\mathcal{U}}_1 := \{u \mid J_1(u) \leq J_1(\tilde{u})\}$. Then, according Lemma 0.3, $J_2(u) \geq J_2(\hat{u})$, $\forall u \in \tilde{\mathcal{U}}_1$ and, (i), $J_2(u) \geq J_2(\tilde{u})$, $\forall u \in \tilde{\mathcal{U}}_1$. Next notice that $\hat{u} \in \tilde{\mathcal{U}}_1$. From (i) it follows then directly that in particular $J_2(\hat{u}) \geq J_2(\tilde{u})$. $\square$

Corollary 0.5 *Assume $\mathcal{U}$ is such that with $u_i(\cdot) \in \mathcal{U}$, $i = 1, 2$ also any concatenation of $u_i(\cdot)$ belongs to $\mathcal{U}$. That is for every t_0 we have that with $u[0, t_0] := u_1[0, t_0]$, $u[t_0, T] := u_2[t_0, T]$, $u \in \mathcal{U}$.*

If $\hat{u}[0, T]$ is a Pareto efficient control for $x(0) = x_0$ in (1,2), then for any $t_0 > 0$, $\hat{u}[t_0, T]$ is a Pareto efficient control for $x(t_0) = \hat{x}(t_0)$ in (1,2). Here $\hat{x}(t_0) = x(t_0, 0, \hat{u}[0, t_0])$ is the value of the state at time t_0 induced by $\hat{u}[0, t_0]$.

Proof. Consider for $x(0) = x_0$ (see (4))

$$\mathcal{U}_1(0) := \{u \mid J_2(u) \leq J_2(\hat{u}[0, T])\}. \quad (5)$$

Let $t_0 > 0$. We will next show (see Lemma 0.3) that $\hat{u}[t_0, T]$ minimizes $J_1(t_0, u)$ on the constrained set

$$\tilde{\mathcal{U}}_1(t_0) := \{u \mid J_2(u) \leq J_2(t_0, \hat{u}[t_0, T])\}, \text{ subject to (30) with } x(t_0) = \hat{x}(t_0). \quad (6)$$

To that end we first note that $\hat{u}[t_0, T] \in \tilde{\mathcal{U}}_1(t_0)$.

Next we show that every element $u \in \mathcal{U}_1(t_0)$ can be viewed as an element $u^e \in \mathcal{U}_1(0)$ restricted to the time interval $[t_0, T]$. That is, $\forall u \in \tilde{\mathcal{U}}_1(t_0)$ there exists $u^e \in \mathcal{U}_1(0)$ such that $u^e[t_0, T] = u$. For, let $u^e[0, T]$ be the concatenation of $\hat{u}[0, t_0]$ with $u[t_0, T]$. Then, clearly, u^e is such that $x(t_0) = \hat{x}(t_0)$. Furthermore,

$$\begin{aligned} J_2(u^e) &= \int_0^T g_2(t, x(t), u^e(t))dt + h_2(x(T)) \\ &= \int_0^{t_0} g_2(t, \hat{x}(t), \hat{u}(t))dt + \int_{t_0}^T g_2(t, x(t), u(t))dt + h_2(x(T)) \\ &\leq \int_0^{t_0} g_2(t, \hat{x}(t), \hat{u}(t))dt + \int_{t_0}^T g_2(t, \hat{x}(t), \hat{u}(t))dt + h_2(\hat{x}(T)) \\ &= J_2(\hat{u}). \end{aligned}$$

So, by definition, $u^e \in \mathcal{U}_1(0)$.

From the dynamic programming principle it follows now directly that $\hat{u}[t_0, T]$ has to minimize $J_1(t_0, u)$ on $\tilde{\mathcal{U}}_1(t_0)$.

In the same way one can show that $\hat{u}[t_0, T]$ also minimizes $J_2(t_0, u)$ on the corresponding constrained set $\tilde{\mathcal{U}}_2(t_0)$, which proves the claim. $\square$

Corollary 0.6 *Assume $J_1(u)$ has a minimum which is uniquely attained at $\hat{u}$. Then $(J_1(\hat{u}), \dots, J_N(\hat{u}))$ is a Pareto solution.*

Proof. Using the notation of Lemma 0.3 it follows directly from the fact that the minimum of $J_1(u)$ is uniquely attained at $\hat{u}$ that $\mathcal{U}_i = \{\hat{u}\}$, for $i = 2, \dots, N$. Furthermore, it is clear that $\hat{u} \in \mathcal{U}_1$.

Consequently, $\min_{u \in \mathcal{U}_i} J_i = J_i(\hat{u})$, $i = 2, \dots, N$ and, since $J_1(u) \geq J_1(\hat{u})$ for arbitrary u , $\min_{u \in \mathcal{U}_1} J_1(u) = J_1(\hat{u})$ too. Lemma 0.3 yields then the advertized result. $\square$

Remark 0.7 From Lemma 0.2 it follows now that in case $J_i(u)$ has a unique minimum location for all i , then every control such that for some $\alpha_i \in [0, 1]$, with $\sum_{i=1}^N \alpha_i = 1$, $\hat{u} \in \arg \min_{u \in \mathcal{U}} \{\sum_{i=1}^N \alpha_i J_i(u)\}$ is Pareto efficient. $\square$

Lemma 0.8 ($N = 2$) *Assume $\alpha_1 < \alpha_2$ and $\hat{u}_i \in \mathcal{U}$ is such that*

$$\hat{u}_i \in \arg \min_{u \in \mathcal{U}} \{\alpha_i J_1(u) + (1 - \alpha_i) J_2(u)\}, \quad i = 1, 2. \quad (7)$$

Then, for all $\alpha \in (\alpha_1, \alpha_2)$ there exists a $\bar{J}(\alpha)$ such that for all u

$$\alpha J_1(u) + (1 - \alpha) J_2(u) \geq \bar{J}(\alpha). \quad (8)$$

Proof. Note that $\alpha J_1(u) + (1 - \alpha)J_2(u) =$

$$\begin{aligned} & \frac{\alpha - \alpha_1}{\alpha_2 - \alpha_1}(\alpha_2 J_1(u) + (1 - \alpha_2)J_2(u)) + \frac{\alpha_2 - \alpha}{\alpha_2 - \alpha_1}(\alpha_1 J_1(u) + (1 - \alpha_1)J_2(u)) \geq \\ & \frac{\alpha - \alpha_1}{\alpha_2 - \alpha_1}(\alpha_2 J_1(\hat{u}_2) + (1 - \alpha_2)J_2(\hat{u}_2)) + \frac{\alpha_2 - \alpha}{\alpha_2 - \alpha_1}(\alpha_1 J_1(\hat{u}_1) + (1 - \alpha_1)J_2(\hat{u}_1)). \end{aligned}$$

□

Using Lemma 0.3 we derive next our main result on necessary conditions for a Pareto solution. For that purpose we impose some additional smoothness conditions on our functions. Furthermore we will for simplicity assume that $\mathcal{U}$ consists of the set of piecewise continuous functions¹.

Assumption 0.9 *Assume that*

- (i) *$f(t, x, u)$ and $g_i(t, x, u)$ are continuous functions on $\mathbb{R}^{1+n+m}$. Moreover, for both f and g_i all partial derivatives w.r.t. x and u exist and are continuous.*
- (ii) *$h_i(x)$ is continuously differentiable.*

□

Theorem 0.10 *Let Assumptions 0.9 be satisfied. Assume $(J_1(\hat{u}), J_2(\hat{u}))$ is a Pareto solution for problem (1,2). Then, there exists an $\alpha \in [0, 1]$, a costate function $\lambda^T(t) : [0, T] \rightarrow \mathbb{R}^n$ (which is continuous and piecewise continuously differentiable) such that, with $H(t, x, u, \lambda) := \alpha g_1(t, x, u) + (1 - \alpha)g_2(t, x, u) + \lambda f(t, x, u)$, $\hat{u}$ satisfies*

$$H(t, \hat{x}(t), \hat{u}(t), \lambda(t)) \leq H(t, \hat{x}(t), u(t), \lambda(t)), \text{ at each } t \in [0, T], \quad (9)$$

$$\dot{\lambda}(t) = -[\alpha \frac{\partial g_1}{\partial x} + (1 - \alpha) \frac{\partial g_2}{\partial x} + \lambda(t) \frac{\partial f}{\partial x}]; \quad \lambda(T) = \frac{\partial(\alpha h_1 + (1 - \alpha)h_2)}{\partial x}, \quad (10)$$

$$\dot{\hat{x}}(t) = f(t, \hat{x}(t), \hat{u}_1(t), \hat{u}_2(t)), \quad \hat{x}(0) = x_0. \quad (11)$$

Proof. Introduce the state variable $\dot{x}_2 := g_2(t, x, u)$, $x_2(0) = 0$, and $x_2^* := \int_0^T g_2(t, \hat{x}, \hat{u})dt + h_2(\hat{x}(T))$, where $\hat{x}$ solves $\dot{\hat{x}} = f(t, \hat{x}, \hat{u})$, $\hat{x}(0) = x_0$. From Lemma 0.3 we have then that in particular the optimization problem

$$\min_u \int_0^T g_1(t, x, u)dt + h_1(x(T)) \text{ such that } x_2(T) + h_2(x(T)) - x_2^* \geq 0$$

subject to

$$\begin{bmatrix} \dot{x} \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} f(t, x, u) \\ g_2(t, x, u) \end{bmatrix}, \quad \begin{bmatrix} x(0) \\ x_2(0) \end{bmatrix} = \begin{bmatrix} x_0 \\ 0 \end{bmatrix},$$

has a solution. From the maximum principle we conclude next (see e.g. [20] and [21, p.61]) that there exist (continuous, piecewise continuously differentiable) costate

¹see e.g. [12] (or [5, p.134]) for a generalization.

functions $\lambda_i(t)$ and a constant $p_1 \in \{0, 1\}$ (with $[p_1, \lambda_1(t), \lambda_2(t)] \neq [0, 0, 0]$ for all $t \in [0, T]$ (i)) such that the Hamiltonian $H^1 := p_1 g_1(t, x, u) + \lambda_1(t) f(t, x, u) + \lambda_2(t) g_2(t, x, u)$ is minimized at $u = \hat{u}$. Furthermore the costate variables satisfy the set of differential equations:

$$\begin{bmatrix} \dot{\lambda}_1 \\ \dot{\lambda}_2 \end{bmatrix} = - \begin{bmatrix} H_x^1 \\ H_{x_2}^1 \end{bmatrix}; \quad \begin{bmatrix} \lambda_1(T) \\ \lambda_2(T) \end{bmatrix} = \begin{bmatrix} p_1 h_1'(x(T)) + \alpha_1 h_2'(x(T)) \\ \alpha_1 \end{bmatrix}, \quad (12)$$

where $\alpha_1 \geq 0$, and $\alpha_1(x_2(T) + h_2(x(T)) - x_2^*) = 0$.

Since $H_{x_2}^1 = 0$, it follows that $\lambda_2(t) = \alpha_1$, for all $t \in [0, T]$. Substitution of this into the "first-order" condition of H^1 , gives then that for all $t \in [0, T]$

$$\begin{aligned} p_1 g_1(t, \hat{x}, \hat{u}) + \lambda_1(t) f(t, \hat{x}, \hat{u}) + \alpha_1 g_2(t, \hat{x}, \hat{u}) &\leq p_1 g_1(t, \hat{x}, u) + \lambda_1(t) f(t, \hat{x}, u) \\ &\quad + \alpha_1 g_2(t, \hat{x}, u). \end{aligned} \quad (13)$$

In a similar way it follows also from Lemma 0.3 that the optimization problem

$$\min_u \int_0^T g_2(t, x, u) dt + h_2(x(T)) \text{ such that } x_1(T) + h_1(x(T)) - x_1^* \geq 0$$

subject to

$$\begin{bmatrix} \dot{x} \\ \dot{x}_1 \end{bmatrix} = \begin{bmatrix} f(t, x, u) \\ g_1(t, x, u) \end{bmatrix}, \quad \begin{bmatrix} x(0) \\ x_1(0) \end{bmatrix} = \begin{bmatrix} x_0 \\ 0 \end{bmatrix},$$

has a solution, where $x_1^* := \int_0^T g_1(t, \hat{x}, \hat{u}) dt + h_1(\hat{x}(T))$. Analogously as before one obtains then from the necessary conditions the existence of an $\alpha_2 \geq 0$, a costate function $\mu_1(t)$ and a constant $p_2 \in \{0, 1\}$ that satisfy:

$$\begin{aligned} \alpha_2 g_1(t, \hat{x}, \hat{u}) + p_2 g_2(t, \hat{x}, \hat{u}) + \mu_1(t) f(t, \hat{x}, \hat{u}) &\leq \alpha_2 g_1(t, \hat{x}, u) + p_2 g_2(t, \hat{x}, u) \\ &\quad + \mu_1(t) f(t, \hat{x}, u), \quad t \in [0, T], \end{aligned} \quad (14)$$

$$\dot{\mu}_1 = -(\alpha_2 \frac{\partial g_1}{\partial x} + p_2 \frac{\partial g_2}{\partial x} + \mu_1(t) \frac{\partial f}{\partial x}); \quad \mu_1(T) = \alpha_2 h_1'(x(T)) + p_2 h_2'(x(T)). \quad (15)$$

Adding (13) and (14) yields

$$\begin{aligned} (p_1 + \alpha_2) g_1(t, \hat{x}, \hat{u}) + (p_2 + \alpha_1) g_2(t, \hat{x}, \hat{u}) + (\lambda_1(t) + \mu_1(t)) f(t, \hat{x}, \hat{u}) &\leq \\ (p_1 + \alpha_2) g_1(t, \hat{x}, u) + (p_2 + \alpha_1) g_2(t, \hat{x}, u) + (\lambda_1(t) + \mu_1(t)) f(t, \hat{x}, u). \end{aligned} \quad (16)$$

Addition of $\lambda_1(t)$ from (12) to $\mu_1(t)$ from (15) shows that $\tilde{\lambda}(t) := \lambda_1(t) + \mu_1(t)$ satisfies

$$\begin{aligned} \dot{\tilde{\lambda}}(t) &= -[(p_1 + \alpha_2) \frac{\partial g_1}{\partial x} + (p_2 + \alpha_1) \frac{\partial g_2}{\partial x} + \tilde{\lambda}(t) \frac{\partial f}{\partial x}]; \\ \tilde{\lambda}(T) &= (p_1 + \alpha_2) h_1' + (p_2 + \alpha_1) h_2'. \end{aligned} \quad (17)$$

Next notice that if $p_1 = \alpha_1 = 0$ it follows straightforwardly from (12) that $\lambda_1(t) = \lambda_2(t) = p_1 = 0$, which violates the maximum principle condition (i).

So either p_1 or α_1 differs from zero (and similarly either p_2 or α_2 is larger than zero too). Therefore $p_1 + p_2 + \alpha_1 + \alpha_2 > 0$. Introducing finally $\alpha := \frac{p_1 + \alpha_2}{p_1 + p_2 + \alpha_1 + \alpha_2}$ and $\lambda(t) := \frac{\tilde{\lambda}(t)}{p_1 + p_2 + \alpha_1 + \alpha_2}$, one obtains directly by division of (16) and (17) by $p_1 + p_2 + \alpha_1 + \alpha_2$ the conditions (9-11). $\square$

Remark 0.11 Notice that the above necessary conditions in Theorem 0.10 are closely related to the minimization of

$$\alpha J_1 + (1 - \alpha) J_2 \text{ subject to (30).} \quad (18)$$

By considering the Hamiltonian for this problem $H := \alpha J_1 + (1 - \alpha) J_2 + \lambda f$, we obtain from the maximum principle in particular the conditions stated in Theorem 0.10. Unfortunately the maximum principle conditions just provide necessary conditions. So, in case all conditions from Theorem 0.10 are met, we still can not conclude (in general) that problem (18) will have a solution. $\square$

Remark 0.12 From the proof of Theorem 0.10 it is clear that without any complications one can also consider problem (1,2) subject to inequality constraints $m_j(x(t), u_1(t), u_2(t)) \leq 0$, where m_j is continuously differentiable in all its arguments, $j = 1, \dots, k$. By restricting the set of admissible controls to those for which these inequalities are satisfied and under the assumption that the constraint qualification² is met, the following necessary conditions result.

Assume $(J_1(\hat{u}), J_2(\hat{u}))$ is a Pareto solution for this problem. Then, there exists an $\alpha \in [0, 1]$, a (continuous and piecewise continuously differentiable) costate function $\lambda^T(t) : [0, T] \rightarrow \mathbb{R}^n$ and (continuous) nonnegative Lagrange parameters $\mu_i(t)$ such that, with the Hamiltonian H as in Theorem 0.10, $\hat{u}$ satisfies

$$H(t, \hat{x}(t), \hat{u}(t), \lambda(t)) \leq H(t, \hat{x}(t), u(t), \lambda(t)), \text{ at each } t \in [0, T] \text{ for all } u \text{ satisfying } m_j(\hat{x}(t), u_1(t), u_2(t)) \leq 0; \quad (19)$$

$$0 = \alpha \frac{\partial g_1}{\partial u} + (1 - \alpha) \frac{\partial g_2}{\partial u} + \lambda(t) \frac{\partial f}{\partial u} + \sum_{j=1}^k \mu_j(t) \frac{\partial m_j}{\partial u}; \quad (20)$$

$$\dot{\lambda}(t) = -[\alpha \frac{\partial g_1}{\partial x} + (1 - \alpha) \frac{\partial g_2}{\partial x} + \lambda(t) \frac{\partial f}{\partial x} + \sum_{j=1}^k \mu_j(t) \frac{\partial m_j}{\partial x}], \quad (21)$$

$$\lambda(T) = \frac{\partial(\alpha h_1 + (1 - \alpha) h_2)}{\partial x};$$

$$\dot{\hat{x}}(t) = f(t, \hat{x}(t), \hat{u}_1(t), \hat{u}_2(t)), \quad x(0) = x_0. \quad (22)$$

$$m_j(\hat{x}(t), \hat{u}_1(t), \hat{u}_2(t)) \leq 0; \quad \mu_j(t) m_j(\hat{x}(t), \hat{u}_1(t), \hat{u}_2(t)) = 0; \quad \mu_j(t) \geq 0. \quad (23)$$

$\square$

²The most convenient constraint qualification is the following rank condition. If p of the inequalities are satisfied with equality then the matrix of partial derivatives of these p constraints w.r.t. $u(t)$ must have rank p .

Next we proceed with the derivation of some sufficient conditions for a control function to be Pareto efficient. The first well-known result (see e.g. [5, Theorem 6.4]) states that under convexity assumptions on the performance functions one can derive all Pareto efficient controls from the minimization of the parametrized optimal control problem (18). This property will be exploited in Section 4 to obtain for a specific class of linear quadratic differential games both necessary and sufficient conditions for existence of Pareto efficient controls.

Theorem 0.13 *If $\mathcal{U}$ is convex³ and $J_i(u)$ is convex for all $i = 1, \dots, N$, then for all Pareto efficient $\hat{u}$ there exist $\alpha \in \mathcal{A}$, such that*

$$\hat{u} \in \arg \min_{u \in \mathcal{U}} \left\{ \sum_{i=1}^N \alpha_i J_i(u) \right\}. \quad \square$$

The next result gives sufficient conditions under which one can conclude from a solution of (9-11) that it will be Pareto efficient. The conditions and proof are inspired by Arrow's theorem.

Theorem 0.14 *Let Assumptions 0.9 be satisfied. Assume there exist an $\alpha \in (0, 1)$, a costate function $\lambda^{*T}(t) : [0, T] \rightarrow \mathbb{R}^n$, u^* and x^* that satisfy (9-11). Introduce the Hamiltonian $H(t, x, u, \lambda^*) := \alpha g_1 + (1 - \alpha)g_2 + \lambda^* f$. Assume that $H(t, x, u, \lambda^*)$ has a minimum w.r.t. u for all x . Let $H^0(t, x, \lambda^*) := \min_u H(t, x, u, \lambda^*)$. Then, if both $H^0(t, x, \lambda^*)$ and $h(x) := \alpha h_1(x) + (1 - \alpha)h_2(x)$ are convex in x , u^* is Pareto efficient.*

Proof. Let α be as in Theorem 0.10 and $J(u) := \alpha J_1(u) + (1 - \alpha)J_2(u)$. Then,

$$\begin{aligned} J(u^*) - J(u) &= \int_0^T \{H(t, x^*, u^*, \lambda^*) - \lambda^*(t)\dot{x}^*(t) - \\ &\quad (H(t, x, u, \lambda^*) - \lambda^*(t)\dot{x}(t))\} dt + h(x^*(T)) - h(x(T)) \\ &= \int_0^T \{H(t, x^*, u^*, \lambda^*) + \dot{\lambda}^*(t)x^*(t) - H(t, x, u, \lambda^*) + \\ &\quad \dot{\lambda}^*(t)x(t)\} dt + h(x^*(T)) - h(x(T)) + \lambda^*(T)(x(T) - x^*(T)), \quad (\text{i}) \end{aligned}$$

where the last equality follows by integration by parts. From the convexity assumption on H^0 and the fact that H^0 is differentiable (see [5, Lemma 4.7]) we get

$$H^0(t, x, \lambda^*) - H^0(t, x^*, \lambda^*) \geq \frac{\partial H^0}{\partial x}(t, x^*, \lambda^*)(x - x^*).$$

³Note that if $\mathcal{U}_i$ is convex also the Cartesian product $\mathcal{U} := \mathcal{U}_1 \times \mathcal{U}_2$ is a convex set.

Since $H^0(t, x^*, \lambda^*) = H(t, x^*, u^*, \lambda^*)$, using (10), we can rewrite the above inequality as

$$\begin{aligned} 0 &\leq H^0(t, x, \lambda^*) - H(t, x^*, u^*, \lambda^*) + \dot{\lambda}^*(x - x^*) \\ &\leq H(t, x, u, \lambda^*) - H(t, x^*, u^*, \lambda^*) + \dot{\lambda}^*(x - x^*). \end{aligned} \quad (\text{ii})$$

The last inequality is due to the fact that by definition of H^0 , $H^0(t, x, \lambda^*) \leq H(t, x, u, \lambda^*)$ for every choice of u .

On the other hand it follows from (10) and the convexity of h

$$\begin{aligned} h(x^*(T)) - h(x(T)) + \lambda^*(T)(x(T) - x^*(T)) &= h(x^*(T)) - h(x(T)) + \\ &\quad \frac{\partial h}{\partial x}(x^*(T))(x(T) - x^*(T)) \leq 0. \end{aligned} \quad (\text{iii})$$

From the inequalities (ii) and (iii) it is then obvious that $J(u^*) - J(u) \leq 0$ in (i). So, $\alpha J_1 + (1 - \alpha)J_2$ is minimized at u^* . So according to Lemma 0.2 u^* is Pareto efficient. $\square$

Remark 0.15 For the constrained problem considered in Remark 0.12 the conditions (20-23) are also sufficient to conclude that $\hat{u}$ is Pareto efficient if $h(x)$ is convex and either i) $H^0(x, \lambda, t) := \min_{\{u | m_j(x, u) \leq 0\}} H(t, x, u, \lambda)$ (as defined in Theorem 0.14) exists and is convex on the convex hull of the set $\mathcal{B} := \{x \mid \exists u, m_j(x, u) \leq 0, j = 1, \dots, k\}$ or ii) the Hamiltonian $H(t, x, u, \lambda^*)$ is simultaneously convex in (x, u) and the constraints $m_j(x, u)$ are simultaneously (quasi-)convex in (x, u) . Details and extensions on this point can be found in [20, Chapter 4.3] or [21]. $\square$

We illustrate some of the theory presented above in an example that was presented in [13, Example 3.5].

Example 0.16 Consider the following advertising game of two competing divisions in a conglomerate company which wish to maximize their individual profits by choosing an optimal advertising policy. Being divisions of a parent company, they are not out to "hurt" each other; thus, a cooperative game solution seems reasonable. With x_i the gross revenue of the i^{th} division and u_i the rate of expenditure for advertising it is assumed that the changes in the gross revenues are given by

$$\dot{x}_1(t) = 12u_1(t) - 2u_1^2(t) - x_1(t) - u_2(t) \quad (24)$$

$$\dot{x}_2(t) = 12u_2(t) - 2u_2^2(t) - x_2(t) - u_1(t) \quad (25)$$

with $x_i(0) = x_{i0}$ and $u_i(t) \geq 0$.

The profits are

$$J_i = \int_0^1 \left\{ \frac{1}{3} x_i(t) - u_i(t) \right\} dt, \quad i = 1, 2,$$

and these are to be maximized for both players.

The with this problem corresponding Hamiltonian is

$$H := -\alpha\left(\frac{x_1}{3} - u_1\right) - (1 - \alpha)\left(\frac{x_2}{3} - u_2\right) + \lambda_1(12u_1 - 2u_1^2 - x_1 - u_2) + \lambda_2(12u_2 - 2u_2^2 - x_2 - u_1).$$

By Theorem 0.10 every Pareto efficient control $\hat{u} \geq 0$ satisfies for some $\alpha \in [0, 1]$

$$H(t, \hat{x}(t), \hat{u}(t), \lambda_1(t), \lambda_2(t)) \leq H(t, \hat{x}(t), u(t), \lambda_1(t), \lambda_2(t)), \text{ at each } t \in [0, 1], \quad (26)$$

$$\dot{\lambda}_1(t) = \lambda_1(t) + \frac{\alpha}{3}; \quad \lambda_1(1) = 0, \quad (27)$$

$$\dot{\lambda}_2(t) = \lambda_2(t) + \frac{1 - \alpha}{3}; \quad \lambda_2(1) = 0, \quad (28)$$

and the system equations (24,25).

It is straightforwardly verified that for $\alpha \in (0, 1)$ we obtain the actions from which Leitmann showed in [13, Example 3.5] that they are Pareto efficient. That is:

$$\begin{aligned} \text{if } \alpha > \alpha_1, \quad \hat{u}_1 &= \begin{cases} \frac{3}{4}(e^{t-1} - 1)^{-1} + 3 - \frac{1-\alpha}{4\alpha} & \text{for } t \in [0, \tau_1] \\ 0 & \text{for } t \in (\tau_1, 1] \end{cases} \\ \text{if } 0 < \alpha \leq \alpha_1, \quad \hat{u}_1 &= 0, \end{aligned}$$

whereas

$$\begin{aligned} \text{if } 0 < \alpha < \alpha_2, \quad \hat{u}_2 &= \begin{cases} \frac{3}{4}(e^{t-1} - 1)^{-1} + 3 - \frac{\alpha}{4(1-\alpha)} & \text{for } t \in [0, \tau_2] \\ 0 & \text{for } t \in (\tau_2, 1] \end{cases} \\ \text{if } \alpha \geq \alpha_2, \quad \hat{u}_2 &= 0. \end{aligned}$$

Here $\alpha_1 = 1/(3(e^{-1} - 1)^{-1} + 13)$, $\alpha_2 = 1 - \alpha_1$, $\tau_1 = 1 + \ln(1 + 3(\frac{1-\alpha}{\alpha} - 12)^{-1})$ and $\tau_2 = 1 + \ln(1 + 3(\frac{\alpha}{1-\alpha} - 12)^{-1})$.

On the other hand it is easily verified that if $\alpha = 0$, $\lambda_1(t) = 0$ and $\lambda_2(t) = \frac{1-e^{t-1}}{3}$. So H is minimized for $\hat{u}_1(t) = 0$ and $\hat{u}_2(t) = \max\{0, \frac{3}{4}(e^{t-1} - 1)^{-1} + 3\}$. Obviously, the minimized Hamiltonian H^0 is linear in the state variables x_i . So in particular H^0 is a convex function of these variables. Therefore we conclude from Theorem 0.14 that the case $\alpha = 0$ yields an appropriate solution too. Similarly one can also show that for $\alpha = 1$ we obtain the Pareto efficient solution $\hat{u}_1(t) = \max\{0, \frac{3}{4}(e^{t-1} - 1)^{-1} + 3\}$ and $\hat{u}_2(t) = 0$. From Theorem 0.10 again we conclude finally that there are no additional Pareto efficient controls than those advertized here. $\square$

THE GENERAL LINEAR QUADRATIC CASE

In this section we consider the linear quadratic differential game. That is

$$J_i(u_1, u_2) := \int_0^T [x^T(t), u_1^T(t), u_2^T(t)] M_i \begin{bmatrix} x(t) \\ u_1(t) \\ u_2(t) \end{bmatrix} dt + x^T(T) Q_{iT} x(T), \quad i = 1, 2, \quad (29)$$

where $M_i = \begin{bmatrix} Q_i & V_i & W_i \\ V_i^T & R_{1i} & N_i \\ W_i^T & N_i^T & R_{2i} \end{bmatrix}$ is symmetric, $R_i := \begin{bmatrix} R_{1i} & N_i \\ N_i^T & R_{2i} \end{bmatrix} \geq 0$, $i = 1, 2$, and $x(t)$ is the solution of the linear differential equation

$$\dot{x}(t) = Ax(t) + B_1 u_1(t) + B_2 u_2(t), \quad x(0) = x_0. \quad (30)$$

Notice that we make no definiteness assumptions w.r.t. matrix Q_i . In this section we will assume that the planning horizon, T , is finite.

For notational convenience we introduce for $\alpha \in [0, 1]$ the next matrices and vectors $M := \alpha M_1 + (1 - \alpha) M_2$, $Q := \alpha Q_1 + (1 - \alpha) Q_2$, $R := \alpha R_1 + (1 - \alpha) R_2$, $V := \alpha V_1 + (1 - \alpha) V_2$, $W := \alpha W_1 + (1 - \alpha) W_2$, $Q_T := \alpha Q_{1T} + (1 - \alpha) Q_{2T}$, $B := [B_1 \ B_2]$,

$$u^T := [u_1^T \ u_2^T] \text{ and } G := \begin{bmatrix} A - BR^{-1} \begin{bmatrix} V^T \\ W^T \end{bmatrix} & -BR^{-1} B^T \\ -(Q + [V \ W] R^{-1} \begin{bmatrix} V^T \\ W^T \end{bmatrix}) & -A^T \end{bmatrix}.$$

For this special case Theorem 0.10 can be specialized as follows.

Corollary 0.17 *If $(J_1(\hat{u}), J_2(\hat{u}))$ is a Pareto solution for problem (29,30) then there exists an $\alpha \in [0, 1]$ such that*

$$[\hat{x}^T(t), \hat{u}^T(t)] M \begin{bmatrix} \hat{x}(t) \\ \hat{u}(t) \end{bmatrix} + \lambda B \hat{u} \leq [\hat{x}^T(t), u^T(t)] M \begin{bmatrix} \hat{x}(t) \\ u(t) \end{bmatrix} + \lambda B u \quad (31)$$

$$\dot{\lambda}(t) = -2\hat{x}^T(t)Q - 2\hat{u}^T \begin{bmatrix} V^T \\ W^T \end{bmatrix} - \lambda A; \quad \lambda(T) = 2\hat{x}^T(T)Q_T; \quad (32)$$

$$\dot{\hat{x}}(t) = A\hat{x}(t) + B\hat{u}(t), \quad \hat{x}(0) = x_0. \quad (33)$$

In case α is such that $R_i > 0$, $i = 1, 2$, (i.e. R_i is positive definite), the above formulae can be equivalently rephrased as that every Pareto efficient control satisfies $\hat{u}(t) = -R^{-1} \left(\begin{bmatrix} V^T \\ W^T \end{bmatrix} x(t) + B^T \tilde{\lambda}(t) \right)$, where $\tilde{\lambda}(t)$ is the solution of the set of linear differential equations:

$$\begin{bmatrix} \dot{\hat{x}}(t) \\ \dot{\tilde{\lambda}}^T(t) \end{bmatrix} = G \begin{bmatrix} x(t) \\ \tilde{\lambda}^T(t) \end{bmatrix}, \quad \begin{bmatrix} x(0) \\ \tilde{\lambda}^T(T) \end{bmatrix} = \begin{bmatrix} x_0 \\ Q_T x(T) \end{bmatrix}. \quad (34)$$

□

In case $R > 0$, for all $\alpha \in [0, 1]$, it follows directly from this Corollary 0.33 that $(0, 0)$ is the only potential Pareto solution if $x_0 = 0$. For in that case $(x(\cdot), \lambda(\cdot)) = (0, 0)$

always solves the above set of differential equations (34).

Another property which readily follows from (34) is that if for some α for both the initial states x_0 and x_1 there exists a $\lambda_i(t)$ satisfying (34) then also for every linear combination of x_0 and x_1 (34) has a solution.

The above property hints to the property that if for both the initial states x_0 and x_1 there exists a Pareto solution also for every linear combination of these initial states a Pareto solution will exist. Unfortunately this property does not hold in general, as will be demonstrated in Example 0.28. However, the following linearity property does hold.

Lemma 0.18 *Assume $\hat{u}$ is a Pareto efficient control for (29,30). Then $\lambda\hat{u}$ is a Pareto efficient solution for (29,30) with $x(0) = \lambda x_0$.*

Proof. Let $x(t, x_0, u)$ denote the solution of (30). Then elementary calculations show that $x(t, \lambda x_0, \lambda u) = \lambda x(t, x_0, u)$ and, consequently, $J_i(\lambda x_0, \lambda u) = \lambda^2 J_i(x_0, u)$. From Lemma 0.3 we know that $\hat{u}$ is Pareto efficient if and only if $\hat{u}$ minimizes $J_i(x_0, u)$ on the constrained set

$$\mathcal{U}_i(x_0, \hat{u}) := \{u \mid J_j(x_0, u) \leq J_j(x_0, \hat{u}), j = 1, 2, j \neq i\}, \text{ for } i = 1, 2. \quad (35)$$

We will next show that $\lambda\hat{u}$ minimizes $J_1(\lambda x_0, u)$ on the constrained set

$$\tilde{\mathcal{U}}_1 := \{u \mid J_2(\lambda x_0, u) \leq J_2(\lambda x_0, \lambda\hat{u})\}. \quad (36)$$

Let $u \in \tilde{\mathcal{U}}_1$. Then from (36) $J_2(\lambda x_0, u) \leq J_2(\lambda x_0, \lambda\hat{u})$ or, equivalently, $J_2(x_0, \frac{1}{\lambda}u) \leq J_2(x_0, \hat{u})$. So, by (35), $\frac{1}{\lambda}u \in \mathcal{U}_1$. But this implies that $J_1(x_0, \frac{1}{\lambda}u) \geq J_1(x_0, \hat{u})$ or, equivalently, $J_1(\lambda x_0, u) \geq J_1(\lambda x_0, \lambda\hat{u})$.

In a similar way one can show that $\lambda\hat{u}$ also minimizes $J_2(\lambda x_0, u)$ on the corresponding constrained set $\tilde{\mathcal{U}}_2$, from which the claim is obvious then. $\square$

It is well-known that existence of a solution of the linear quadratic control problem for an arbitrary initial state is equivalent to the existence of a solution of an associated Riccati equation. For that reason we consider below the problem under which conditions for an arbitrary initial state (29,30) has a Pareto solution. We have the following two preliminary results.

Lemma 0.19 *Consider the two-point boundary value problem*

$$\begin{bmatrix} \dot{x}(t) \\ \dot{\lambda}(t) \end{bmatrix} = G \begin{bmatrix} x(t) \\ \lambda(t) \end{bmatrix}, \quad \begin{bmatrix} x(0) \\ \lambda(T) \end{bmatrix} = \begin{bmatrix} x_0 \\ Q_T x(T) \end{bmatrix}. \quad (37)$$

Let $[W_1 \ W_2] := [I \ 0]e^{-GT}$. Then, (37) has a solution for every x_0 if and only if $U := W_1 + W_2 Q_T$ is invertible.

Proof. Obviously, (37) has a solution for every x_0 if and only if $\forall x_0$ the following differential equation has a solution

$$\dot{\tilde{x}} = G\tilde{x}, \text{ with } \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \tilde{x}(0) + \begin{bmatrix} 0 & 0 \\ -Q_T & I \end{bmatrix} \tilde{x}(T) = \begin{bmatrix} x_0 \\ 0 \end{bmatrix}.$$

Clearly, this is the case if and only if $\forall x_0$ the following equation has a solution

$$\left(\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ -Q_T & I \end{bmatrix} e^{GT} \right) \tilde{x}(0) = \begin{bmatrix} x_0 \\ 0 \end{bmatrix}.$$

Or, equivalently, $\forall x_0$ the equation below is solvable

$$\left(\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} e^{-GT} + \begin{bmatrix} 0 & 0 \\ -Q_T & I \end{bmatrix} \right) e^{GT} \tilde{x}(0) = \begin{bmatrix} x_0 \\ 0 \end{bmatrix}.$$

Or, stated differently, with $z := e^{GT} \tilde{x}(0)$ the following equation has a solution $\forall x_0$

$$\begin{bmatrix} W_1 & W_2 \\ -Q_T & I \end{bmatrix} z = \begin{bmatrix} x_0 \\ 0 \end{bmatrix}.$$

Premultiplication of both sides of this equation by matrix $\begin{bmatrix} I & -W_2 \\ 0 & I \end{bmatrix}$ shows then that the above equation is solvable if and only if matrix U is invertible. $\square$

Whereas from [5, Corollary 5.13] we recall the following result.

Lemma 0.20 *Let G be as introduced before and $U(t) := [I \ 0]e^{G(T-t)} \begin{bmatrix} I \\ Q_T \end{bmatrix}$. Consider the with G corresponding Riccati differential equation*

$$\begin{aligned} \dot{K}(t) &= -A^T K(t) - K(t)A + (K(t)B + [V \ W])R^{-1}(K(t) + [V \ W]^T) - Q, \\ K(T) &= Q_T. \end{aligned} \tag{38}$$

Then, (38) has a solution on $[0, T]$ if and only if $U(t)$ is invertible on $[0, T]$. $\square$

Remark 0.21 According to the fundamental existence-uniqueness theorem of differential equations there exists a maximum time interval $[0, T_1)$ where equation (38) has a unique solution. So we conclude from 0.2 and [5, Theorem 5.1] that for a planning horizon $T < T_1$ for every initial state x_0 the game has a Pareto solution. $\square$

Using the above lemmas we obtain for the scalar case, that is the case that the dimension of the state variable $x(t)$ is one, the following result.

Theorem 0.22 *Consider the scalar system with $R_i > 0$, $i = 1, 2$. Let $x_0 \neq 0$. Then,*

1. *if (29,30) has a Pareto efficient control $\hat{u}(x_0)$ there exists an $\alpha \in [0, 1]$ such that $\min \alpha J_1 + (1 - \alpha) J_2$ subject to (30) has a solution for all initial states $x_0 \in \mathbb{R}$.*
2. *if J_i , $i = 1, 2$, attains a minimum then*
 - i) *for all $\alpha \in [0, 1]$, $\min \alpha J_1 + (1 - \alpha) J_2$ exists for all x_0 ;*
 - ii) *all Pareto efficient controls of (29,30) are given by $\{\hat{u} \mid \hat{u} = \arg \min \alpha J_1 + (1 - \alpha) J_2, \alpha \in [0, 1]\}$.*

Proof. "1." Assume $(J_1(\hat{u}), J_2(\hat{u}))$ is a Pareto solution for problem (29,30) for some $x_0 \neq 0$. Then, according to Lemma 0.18, for every initial state problem (29,30) has a Pareto solution. Let $x_0 \neq 0$ be fixed and α , $x(t)$ and $\lambda(t)$ a with $\hat{u}(x_0)$ corresponding solution that satisfies (34). Then it is easily verified that for the initial state μx_0 , α , $\mu x(t)$ and $\mu \lambda(t)$ satisfy (34). In other words, for this α , $\forall x_0$ (34) has a solution. So, by Lemma 0.19, $U(t) := [I \ 0] e^{G(T-t)} \begin{bmatrix} I \\ Q_T \end{bmatrix}$ is invertible at $t = 0$. Furthermore, by Corollary 0.33, $\hat{u}[t_0, T]$ is Pareto efficient for problem (29,30) if we consider J_i on the interval $[t_0, T]$, instead of $[0, T]$, with initial state $x(t_0) = e^{G t_0} x_0$. Again it is easily verified that also in this case α , $x(t)$ and $\lambda(t)$ satisfy (34). So similar as before it follows that $U(t_0)$ is invertible too. Since $t_0 \in [0, T]$ was chosen arbitrarily we conclude from Lemma 0.20 that the Riccati differential equation (38) has a solution on $[0, T]$. But this implies (see e.g. [5, Theorem 5.1]) that the optimization problem $\min \alpha J_1 + (1 - \alpha) J_2$ subject to (30) has a solution for all initial states x_0 . It is easily verified that this solution is actually attained by $\hat{u}(x_0)$.

"2." Since $J_i(\lambda x_0, \lambda u) = \lambda^2 J_i(x_0, u)$ and by assumption J_1 and J_2 have a minimum, by Lemma 0.8 for all $\alpha \in [0, 1]$ there exists a $\bar{J}(x_0, \alpha)$ such that

$$\begin{aligned} \alpha J_1(\lambda x_0, \lambda u) + (1 - \alpha) J_2(\lambda x_0, \lambda u) &= \lambda^2 (\alpha J_1(x_0, u) + (1 - \alpha) J_2(x_0, u)) \\ &\geq \lambda^2 \bar{J}(x_0, \alpha). \end{aligned}$$

From this it follows that for all $\alpha \in [0, 1]$, for all x_0 $\inf \alpha J_1(x_0, u) + (1 - \alpha) J_2(x_0, u)$ exists. But this implies (see e.g. [5, p.182,183]) that actually the optimization problem $\min \alpha J_1 + (1 - \alpha) J_2$ subject to (30) has a unique solution for all x_0 .

Part ii) follows directly from Part i) and Remark 0.7. $\square$

Remark 0.23

1. As already noticed in the proof of Theorem 0.22.1 it follows directly from Lemma 0.18 that, in case the scalar game has a Pareto solution for some initial state different from zero, the game (29,30) has a Pareto solution for every initial state.

2. From the proof of Theorem 0.22 we can in fact conclude the following result. If for $x_0 \neq 0$ there exists a Pareto solution and for $\alpha_0 \in [0, 1]$ (34) has a solution with $x(0) = x_0$, then $\min \alpha_0 J_1 + (1 - \alpha_0) J_2$ subject to (30) has a solution for all initial states $x_0 \in \mathbb{R}$.
3. The assumption that the minimum of J_i exists in Theorem 0.22.2 implies that J_i is convex (see [7, Theorems 2.8, 2.9]). Therefore, this part of the theorem is a special case of those dealt with in Section 4. $\square$

The next example illustrates some of the subtleties of Theorem 0.22.

Example 0.24 Consider the minimization of

$$J_i = \int_0^{\pi/2} \{-x^2 + \beta_i u^2\} dt \quad \text{subject to } \dot{x} = u, \quad x(0) = x_0, \quad i = 1, 2.$$

First consider the minimization of

$$J_1 \quad \text{subject to } \dot{x} = u, \quad x(0) = x_0. \quad (39)$$

In case $\beta_1 = 1$, it is well-known (see e.g. [5, Example 5.1]) that this problem (39) has a solution if and only if $x_0 = 0$. In case $x_0 = 0$, $u(\cdot) = \gamma \cos(t)$ yields for every γ the optimal value 0.

Furthermore, if $\beta_1 > 1$, the with this problem corresponding Riccati differential equation (38), $\dot{k}(t) = \frac{1}{\beta_1} k^2(t) + 1$; $k(\frac{\pi}{2}) = 0$, has a solution on $[0, \frac{\pi}{2}]$. So in that case problem (39) has for every initial state x_0 a solution.

Finally, in case $\beta_1 < 1$ one can use e.g. the control sequence from [3, Remark 3.1.4] (which was used to show that for $\beta_1 = 1$ this example has no solution in case $x_0 \neq 0$) to construct also for $x_0 = 0$ a control sequence for which J_1 becomes negative (implying that $\inf J_1$ does not exist), yielding the conclusion that for $\beta_1 < 1$ for all initial states problem (39) has no solution.

Next we consider a consequence of these conclusions for cooperative games in relationship with Theorem 0.22.

Consider the case $\beta_i = 1$, $i = 1, 2$. Then, obviously, for every $\alpha \in [0, 1]$ $\alpha J_1 + (1 - \alpha) J_2 = J_1$. From the above consideration we have that for all $x_0 \neq 0$, for all $\alpha \in [0, 1]$, this problem has no solution. So from Theorem 0.22.1 we conclude that this cooperative game has no Pareto solution if $x_0 \neq 0$. On the other hand we conclude from Lemma 0.2 and the above considerations that for $x_0 = 0$ there does exist a Pareto solution, i.e. $(0, 0)$. $\square$

A direct consequence of the above Theorem 0.22 is the following theorem.

Theorem 0.25 *Consider the scalar system with $R_i > 0$, $i = 1, 2$. Then for every x_0 (29,30) has a Pareto efficient control $\hat{u}(x_0)$ if and only if there exists an $\alpha \in [0, 1]$ such that for every x_0 $\min \alpha J_1 + (1 - \alpha)J_2$ subject to (30) has a solution.*

Proof. " $\Rightarrow$ " In particular it follows that (29,30) has a Pareto efficient solution for some $x_0 \neq 0$. Theorem 0.22 yields then the advertized result.

" $\Leftarrow$ " Since the minimum exists for all x_0 it is well-known that the argument at which this minimum is attained is unique (see e.g. [5, Theorem 5.1]). Consequently Remark 0.7 yields then directly the conclusion. $\square$

Using Theorem 0.25 we arrive then at the next procedure to find all Pareto efficient solutions for the scalar game.

Algorithm 0.26 Using the notation introduced in the beginning of this section, consider the scalar system with $R_i > 0$, $i = 1, 2$.

Let $\hat{\mathcal{U}}$ be the set of all Pareto efficient controls for which for every x_0 (29,30) has a Pareto solution. Next consider the Riccati differential equation

$$\dot{K}(t) = -A^T K(t) - K(t)A - Q(\alpha) + (K(t)B + [V(\alpha) \ W(\alpha)])R^{-1}(\alpha)(B^T K(t) + [V(\alpha) \ W(\alpha)]^T); \ K(T) = Q_T(\alpha). \quad (40)$$

Then all Pareto efficient controls $\hat{u}(\alpha) \in \hat{\mathcal{U}}$ are obtained by determining all $\alpha \in [0, 1]$ for which (40) has a solution $K_\alpha(t)$ on $[0, T]$. More in particular, if (40) has a solution $K_\alpha(t)$ on $[0, T]$ then

$$\hat{u}(\alpha)(t) := -R^{-1}(\alpha)([V(\alpha) \ W(\alpha)]^T + B^T K_\alpha(t))\hat{x}(t),$$

where $\hat{x}(t)$ solves the differential equation $\dot{x}(t) = (A - BR^{-1}(\alpha)([V(\alpha) \ W(\alpha)]^T + B^T K_\alpha(t)))x(t)$, $x(0) = x_0$, yields a Pareto efficient control. The corresponding Pareto solution is obtained by determining $(J_1(\hat{u}(\alpha)), J_2(\hat{u}(\alpha)))$. $\square$

Example 0.27 Consider the cooperative game with

$$J_1 = \int_0^{\frac{\pi}{2}} \{-x^2 + \frac{9}{10}u_1^2 + \frac{1}{10}u_2^2\}dt \text{ and } J_2 = \int_0^{\frac{\pi}{2}} \{-x^2 + \frac{1}{10}u_1^2 + \frac{9}{10}u_2^2\}dt$$

subject to the system

$$\dot{x} = \frac{4}{10}(u_1 + u_2), \ x(0) = x_0.$$

Here player i controls u_i . According to Algorithm 0.26 this game has for every initial state a Pareto efficient solution precisely for those $\alpha \in [0, 1]$ for which the following Riccati differential equation has a solution on $[0, \frac{\pi}{2}]$:

$$\dot{k} = sk^2 + 1, \ k(\frac{\pi}{2}) = 0, \text{ where } s := \frac{16}{(1 + 8\alpha)(9 - 8\alpha)}.$$

The solution of this differential equation is $k(t) = \frac{1}{\sqrt{s}} \tan \sqrt{s}(t - \frac{\pi}{2})$. So, $k(t)$ exists on $[0, \frac{\pi}{2}]$ if and only if $\sqrt{s} < 1$. It is easily verified that this is equivalent to the condition $\alpha \in (\frac{1}{8}, \frac{7}{8})$.

So, the set of all Pareto efficient controls is given by

$$\hat{u}(t) = -\frac{4k(t)}{10} \begin{bmatrix} \frac{10}{1+8\alpha} & 0 \\ 0 & \frac{10}{9-8\alpha} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \hat{x}(t),$$

where $\hat{x}(t)$ solves $\dot{x}(t) = -sk(t)x(t)$, $x(0) = x_0$, and $\alpha \in (\frac{1}{8}, \frac{7}{8})$.

To complete the picture we will finally analyze the case $\alpha \in I := [0, \frac{1}{8}] \cup [\frac{7}{8}, 1]$. From Remark 0.23.2 and Theorem 0.25 it follows that for all $\alpha \in I$ there does not exist a $x_0 \neq 0$ such that the game has a Pareto solution for this initial state. So, only for $x_0 = 0$ for these $\alpha \in I$ an additional Pareto solution might occur. However, as already noticed directly after Corollary 0.33, this leads to the only candidate control $(0, 0)$, from which we know already that it is Pareto efficient. $\square$

The following example illustrates a two-dimensional state game where for all initial states in the interior of the second and fourth quadrant and the point $(0, 0)$ a Pareto solution exists whereas for all other initial states there exists no Pareto solution.

Example 0.28 Consider the cooperative minimization for $i = 1, 2$ of

$$J_i = \int_0^{\pi/2} \{x^T(t)M_i x(t) + u^2(t)\}dt, \text{ with } M_1 = \begin{bmatrix} -1 & 1 \\ 1 & -2 \end{bmatrix}, M_2 = \begin{bmatrix} -3 & 2 \\ 2 & -2 \end{bmatrix};$$

subject to

$$\dot{x}_1 = u, \quad x_1(0) = p, \quad \dot{x}_2 = u, \quad x_2(0) = q.$$

To solve this problem, consider the minimization of

$$\alpha J_1 + (1 - \alpha)J_2 = \int_0^{\pi/2} \{x^T(t) \begin{bmatrix} -3 + 2\alpha & 2 - \alpha \\ 2 - \alpha & -2 \end{bmatrix} x(t) + u^2(t)\}dt \quad (41)$$

subject to

$$\dot{x}(t) = \begin{bmatrix} 1 \\ 1 \end{bmatrix} u(t); \quad x(0) = \begin{bmatrix} p \\ q \end{bmatrix}. \quad (42)$$

From the above set of state differential equations it follows directly that $x_1(t) = x_2(t) + p - q$. Substitution of this into (41) shows that the minimization of (41) subject to (42) is equivalent to the minimization of $\alpha J_1 + (1 - \alpha)J_2 =$

$$\alpha J_1 + (1 - \alpha)J_2 = \int_0^{\pi/2} \{-(x_2(t) + (1 - \alpha)(p - q))^2 + u^2(t) + (-2 + \alpha^2)(p - q)^2\}dt$$

subject to

$$\dot{x}_2(t) = u(t); \quad x_2(0) = q.$$

From Example 0.24 it follows that this problem has a solution if and only if the initial state satisfies $-(1 - \alpha)(p - q) = q$ or, equivalently, $(1 - \alpha)p = -\alpha q$. From this it is easily verified that for every initial state (p, q) of (42) there is exact one $\alpha \in [0, 1]$ such that the minimization of (41) has a solution if (p, q) is located either in the second quadrant, the fourth quadrant, on the line $p = 0$ (which corresponds with $\alpha = 0$) or on the line $q = 0$ (which corresponds with $\alpha = 1$). So we conclude from Lemma 0.2 that for all the initial states (p, q) which are located in the interior of the second and fourth quadrant there is a Pareto solution.

Next consider the initial state $(p, 0)$. Then $x_1(t) = x_2(t) + p$ and we can rewrite J_1 as follows:

$$\begin{aligned} J_1 &= \int_0^{\frac{\pi}{2}} \{-(x_1(t) - x_2(t))^2 - x_2^2(t) + u^2(t)\} dt \\ &= \int_0^{\frac{\pi}{2}} \{-p^2 - x_2^2(t) + u^2(t)\} dt. \end{aligned}$$

From Example 0.24 we recall again that the minimum value of J_1 is 0. Furthermore, $u_2(t) := \gamma \cos(t)$ belongs to the with this problem corresponding constrained set $\mathcal{U}_2$ (see Lemma 0.3). Using this control we see that

$$\begin{aligned} J_2 &= \int_0^{\frac{\pi}{2}} \{-2(x_1(t) - x_2(t))^2 - x_1^2(t) + u_2^2(t)\} dt \\ &= \int_0^{\frac{\pi}{2}} \{-2p^2 - (\gamma \sin(t) + p)^2 + \cos^2(t)\} dt \\ &= -\frac{3p^2\pi}{2} + 2p\gamma. \end{aligned}$$

Since γ is an arbitrary number it is clear from this that $\min_{u \in \mathcal{U}_2} J_2$ does not exist if $p \neq 0$. So, by Lemma 0.3, for the initial state $(p, 0)$, there exists no Pareto efficient solution. Similarly it follows that also for the initial state $(0, q)$ the problem has no Pareto efficient solution. Finally, it is easily verified that for the initial state $(0, 0)$, for every γ , $\hat{u} = \gamma \cos(t)$ is a Pareto efficient control yielding the same Pareto solution $(0, 0)$.

Obviously, the with problem (41,42) corresponding Riccati differential equation has for every $\alpha \in [0, 1]$ not a solution on $[0, \frac{\pi}{2}]$.

Finally, consider matrix

$$G := \begin{bmatrix} A & -BR^{-1}B^T \\ -Q & -A^T \end{bmatrix},$$

from (34). It is easily verified that

$$G = SJS^{-1}, \text{ with } S = \begin{bmatrix} 0 & 1/\alpha & 0 & \frac{-\alpha}{2-\alpha^2} \\ 0 & 1/\alpha & 0 & \frac{1-\alpha}{2-\alpha^2} \\ \frac{1-\alpha}{\alpha} & 0 & -1 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix} \text{ and } J = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

So for the initial state $x_0 = [p; q]^T$ there exists an $\alpha \in [0, 1]$, $x(\cdot)$, and $\lambda(\cdot)$ satisfying (34) if and only if there exists an $\alpha \in [0, 1]$ and $\bar{\lambda}_1, \bar{\lambda}_2$ such that the equation

$$e^{G\frac{\pi}{2}}[p; q; \bar{\lambda}_1; \bar{\lambda}_2]^T = [x_1(\frac{\pi}{2}); x_2(\frac{\pi}{2}); 0; 0]^T$$

has a solution. An elementary spelling of these equations shows that there exists a solution if and only if $p(1 - \alpha) = -\alpha q$. So, in particular we conclude from this that for all initial states located in either the interior of the first or third quadrant there exists no Pareto solution. $\square$

THE CONVEX LINEAR QUADRATIC CASE

Inspired by the results of Lemma 0.2 and Theorem 0.13 we consider in this section in detail the case when the linear quadratic cost functions are convex. The performance criteria of the players is again given by (29). An additional assumption will be that $R_i > 0$ throughout this section. On the other hand we consider in this section a slightly more general system, i.e.

$$\dot{x}(t) = Ax(t) + B_1u_1(t) + B_2u_2(t) + c(t), \quad x(0) = x_0. \quad (43)$$

The variable $c(\cdot) \in L_2$ is some given trajectory. The planning horizon, T , may be either finite or infinite.

Note, that whenever J_i are convex also $\sum_{i=1}^N \alpha_i J_i(\gamma)$ is convex for an arbitrary $\alpha \in \mathcal{A}$.

Next we consider some sets of control functions that are relevant in our problem setting. It can be easily shown that each of these (nonempty⁴) sets is convex.

Lemma 0.29 *Let $\mathcal{U}$ be given by either:*

- 1) $L_2[0, T] := \{(u_1, u_2) \mid u_i(\cdot) \text{ are square integrable functions on } [0, T]\}$.
- 2) $\{u \in L_2[0, T] \mid x(0) = x(T) = 0 \text{ in } \dot{x}(t) = Ax(t) + Bu(t)\}$.
- 3) $L_{2,e}^+ := \{(u_1, u_2) \mid u_i(\cdot) \in L_{2,loc} \text{ and } \lim_{T \rightarrow \infty} J_i(x_0, u) \text{ exists in } \mathbb{R} \cup \{-\infty, \infty\}, \forall x_0\}$, where $L_{2,loc}$ is the set of locally square-integrable functions, i.e.,

$$L_{2,loc} = \{u \mid \forall T > 0, \int_0^T u^T(s)u(s)ds < \infty\}.$$

⁴assuming for case 2) and 3) that $(A, [B_1 \ B_2])$ is stabilizable

4) $L_{2,e,s}^+ := \{(u_1, u_2) \mid (u_1, u_2) \in L_{2,e}^+ \text{ and } \lim_{t \rightarrow \infty} x(t) = 0 \text{ in (43)}\}$.
Then each of these set of control functions is convex. $\square$

Next we consider the question under which conditions the cost functions J_i in (29) are convex. To that end we first derive some preliminary results.

Lemma 0.30 Assume $\mathcal{U}$ is convex. Consider the cost function $J(s, T, u, x_0) =$

$$\int_s^T \{[x^T(t) \ u^T(t)]M \begin{bmatrix} x(t) \\ u(t) \end{bmatrix} + 2[p_1^T(t) \ p_2^T(t)] \begin{bmatrix} x(t) \\ u(t) \end{bmatrix} + c_2(t)\}dt + x^T(T)Q_Tx(T) \quad (44)$$

$$\text{subject to the state dynamics: } \dot{x}(t) = Ax(t) + Bu(t) + c_1(t), \ x(s) = x_0, \quad (45)$$

where u , p_i and c_i are such that (44) and (45) have a solution; $M = \begin{bmatrix} Q & V \\ V^T & R \end{bmatrix}$.
Let $x_0 \in \mathbb{R}^N$. Then, $J(0, T, u, x_0)$ is convex as a function of u if and only if $\bar{J}(0, T, v, 0) \geq 0$ for all v , where

$$\bar{J}(s, T, v, 0) = \int_s^T [z^T(t) \ v^T(t)]M \begin{bmatrix} z(t) \\ v(t) \end{bmatrix} dt + z^T(T)Q_Tz(T) \\ \text{with } \dot{z}(t) = Az(t) + Bv(t), \ z(s) = 0. \quad (46)$$

Proof: Let $x_u(t)$ denote the state trajectory of (45) in case the control $u(\cdot)$ is used. Then it is well-known that due to the linearity of the system

$$x_{\lambda u + (1-\lambda)w}(t) = \lambda x_u(t) + (1-\lambda)x_w(t). \quad (47)$$

From the definition of convexity (see (49)) it follows then that $J(0, T, u, x_0)$ is convex if and only if

$$\begin{aligned} & \int_0^T \{[x_{\lambda u + (1-\lambda)w}^T(t) \ (\lambda u + (1-\lambda)w)^T(t)]M \begin{bmatrix} x_{\lambda u + (1-\lambda)w}(t) \\ (\lambda u + (1-\lambda)w)(t) \end{bmatrix} + \\ & \quad 2[p_1^T(t) \ p_2^T(t)] \begin{bmatrix} x_{\lambda u + (1-\lambda)w}(t) \\ (\lambda u + (1-\lambda)w)(t) \end{bmatrix} + c_2(t)\}dt x_{\lambda u + (1-\lambda)w}^T(t) \\ & \leq \int_0^T \{\lambda[x_u^T(t) \ u(t)]M \begin{bmatrix} x_u(t) \\ u(t) \end{bmatrix} + 2[p_1^T(t) \ p_2^T(t)] \begin{bmatrix} x_u(t) \\ u(t) \end{bmatrix} + c_2(t) + \\ & \quad (1-\lambda)([x_w^T(t) \ w(t)]M \begin{bmatrix} x_w(t) \\ w(t) \end{bmatrix} + 2[p_1^T(t) \ p_2^T(t)] \begin{bmatrix} x_w(t) \\ w(t) \end{bmatrix} + c_2(t))\}dt + \\ & \quad \lambda x_u^T(T)Q_Tx_u(T) + (1-\lambda)x_w^T(T)Q_Tx_w(T). \end{aligned}$$

Or equivalently, using (47), we obtain after some rewriting

$$\begin{aligned} & \lambda(1-\lambda) \left(\int_0^T \{[x_u^T(t) - x_w^T(t) \ u(t) - w(t)]M \begin{bmatrix} x_u(t) - x_w(t) \\ u(t) - w(t) \end{bmatrix} \}dt + \right. \\ & \quad \left. (x_u^T(T) - x_w^T(T))Q_T(x_u^T(T) - x_w^T(T)) \right) \geq 0, \end{aligned}$$

where $\dot{x}_u(t) = Ax_u(t) + Bu(t) + c(t)$, $x_u(0) = x_0$ and $\dot{x}_w(t) = Ax_w(t) + Bw(t) + c(t)$, $x_w(0) = x_0$.

With $z := x_u - x_w$ and $v := u - w$ the stated result then follows immediately. $\square$

Remark 0.31

1. Note that the second part of the equivalence does not depend on x_0 . In particular it follows from this that if J is convex for one x_0 then J is convex for all x_0 . This property can also be verified by a direct elaboration of the convexity definition using the linearity property of the system again.
2. Consider a linear system in its state controllable canonical form

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} A_1 & A_2 \\ 0 & A_3 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} B_1 \\ 0 \end{bmatrix} u(t), \quad x(0) = x_0,$$

with (A_1, B_1) controllable and cost $J =$

$$\int_0^T \{x^T(t) \begin{bmatrix} Q_1 & Q_2 \\ Q_2^T & Q_3 \end{bmatrix} x(t) + u^T(t) Ru(t)\} dt + x^T(T) \begin{bmatrix} Q_{1T} & Q_{2T} \\ Q_{2T}^T & Q_{3T} \end{bmatrix} x(T).$$

Then it follows immediately from Lemma 0.30 that J is convex if and only if the controllable part of this system is convex. That is if, with

$$\begin{aligned} \dot{x}_1(t) &= A_1 x_1(t) + B_1 u(t), \quad x_1(0) = x_{10}, \text{ and} \\ J_1 &= \int_0^T \{x_1^T(t) Q_1 x_1(t) + u^T(t) Ru(t)\} dt + x_1^T(T) Q_{1T} x_1(T), \end{aligned}$$

J_1 is convex. $\square$

Lemma 0.32 *Assume $\mathcal{U}$ is convex. Consider the linear quadratic cost function (44) and (45). Then, if $J(0, T, u, x_0)$ is convex for some x_0 , $J(s, T, u, x_0)$ is convex for all $s \geq 0$ and for all x_0 .*

Proof: Let $s > 0$. Let $u_i(t) = v(t) + w_i(t)$, where $v(t) = 0$, $t \geq s$ and $w_i(t) = 0$, $0 \leq t \leq s$. Then,

$$J(0, T, u_i, x_0) = J(0, s, v, x_0) + J(s, T, w_i, x(s, x_0)). \quad (48)$$

Since $J(0, T, u, x_0)$ is convex, by definition, for an arbitrary u and w and $\lambda \in (0, 1)$

$$J(0, T, \lambda u + (1 - \lambda)w, x_0) \leq \lambda J(0, T, u, x_0) + (1 - \lambda)J(0, T, w, x_0). \quad (49)$$

So in particular if we choose $u = u_1$ and $w = u_2$ as above we get, using (48),

$$\begin{aligned} J(0, s, v, x_0) &+ J(s, T, \lambda w_1 + (1 - \lambda)w_2, x(s, x_0)) = J(0, T, \lambda u_1 + (1 - \lambda)u_2, x_0) \\ &\leq \lambda J(0, T, u_1, x_0) + (1 - \lambda)J(0, T, u_2, x_0) \\ &= \lambda(J(0, s, v, x_0) + J(s, T, w_1, x(s, x_0))) + \\ &\quad (1 - \lambda)(J(0, s, v, x_0) + J(s, T, w_2, x(s, x_0))) \\ &= J(0, s, v, x_0) + \lambda J(s, T, w_1, x(s, x_0)) + (1 - \lambda)J(s, T, w_2, x(s, x_0)). \end{aligned}$$

Comparing both sides of this inequality shows then that $J(s, T, w, x(s, x_0))$ is convex. The rest of the statement follows then directly from Remark 0.31, item 1. $\square$

Corollary 0.33 Consider the linear quadratic cost function (44) and (45). Then, $J(s, T, u, x_0)$ is convex for all $s \geq 0$ and x_0 if and only if $J(0, T, u, x_0)$ is convex for all/an x_0 , which holds if and only if $\bar{J}(0, T, v, 0) \geq 0$ for all v , where $\bar{J}$ is given by (46). $\square$

Theorem 0.34 Let either T be finite or (A, B) be stabilizable. Then $J(0, T, u, x_0)$ is convex for all/an x_0 if and only if $\inf \bar{J}(0, T, v, 0)$ exists.

Proof:

" $\Rightarrow$ " From Corollary 0.33 it follows that if $J(0, T, u, x_0)$ is convex $\bar{J}(0, T, v, 0) \geq 0$ for all v . In case T is finite obviously $\inf \bar{J}(0, T, v, 0) \leq J(0, T, 0, 0)$ whereas in case (A, B) is stabilizable, $\inf \bar{J}(0, \infty, v, 0)$ is bounded from above too. Consequently, $\inf \bar{J}(0, T, v, 0)$ exists.

" $\Leftarrow$ " Assume $\inf \bar{J}(0, T, v, 0) = m$. Then $m \geq 0$. For if $m < 0$ there would exist a $\bar{v}$ such that $\bar{J}(0, T, \bar{v}, 0) < 0$. But from this it follows directly from the linearity of the system that $\bar{J}(0, T, \lambda \bar{v}, 0) = \lambda^2 \bar{J}(0, T, \bar{v}, 0)$. From which it is clear that $\inf \bar{J}(0, T, v, 0)$ would not exist. So $\bar{J}(0, T, v, 0) \geq 0$ for all v . This implies, see Corollary 0.33 again, that $J(0, T, u, x_0)$ is convex for all/an x_0 . $\square$

Next we consider the case $R > 0$ in M . It is well-known that then the next Riccati equations play an important role

$$\dot{K}(t) = A^T K(t) + K(t)A - (K(t)B + V)R^{-1}(B^T K(t) + V^T) + Q, \quad K(T) = Q_T; \quad (\text{DRE})$$

$$0 = A^T K + KA - (KB + V)R^{-1}(B^T K + V^T) + Q. \quad (\text{ARE})$$

Let Γ denote the set of all symmetric solutions of (ARE). From, e.g., [5] and [27] (see also [18], [11], [30], [14] [15] and [28]) we recall the next results. Notice that these results are formulated in terms of existence of the performance function for an arbitrary initial state. In the above cited references one can also find conditions for the existence of $\inf \bar{J}(0, T, v, 0)$. Since the formulation of these conditions is more involved and is somewhat outside the main scope of this note they are not presented here.

Theorem 0.35 Consider the minimization of

$$J := \int_0^T [x^T(t) \ u^T(t)] M \begin{bmatrix} x(t) \\ u(t) \end{bmatrix} dt + x^T(T) Q_T x(T)$$

subject to $\dot{x}(t) = Ax(t) + Bu(t)$, $x(0) = x_0$. Then the following holds:

1. Let $T < \infty$ and $\mathcal{U} = L_2[0, T]$. Then, $\inf_u J$ exists for all x_0 if and only if (DRE) has a solution on $[0, T]$. Moreover, if this infimum exists it is in fact a minimum. This $\min_u J(u) = x_0^T K(0)x_0$ is attained uniquely by $u^*(t) = -R^{-1}(B^T K(t) + V)x^*(t)$, where $K(t)$ is the unique symmetric solution of (DRE) and $x^*(\cdot)$ solves $\dot{x}(t) = (A - BR^{-1}(B^T K(t) + V))x(t)$, $x^*(0) = x_0$.
2. Let $T = \infty$, $\mathcal{U} = L_{2,e,s}^+$ and (A, B) be stabilizable. Then, $\inf_u J$ exists for all x_0 if and only if $\Gamma \neq \emptyset$. Moreover, there exists a u^* attaining this infimum if and only if (ARE) has a stabilizing solution K^+ . Under this condition $\min_u J(u) = x_0^T K^+ x_0$ is attained uniquely by $u^*(t) = -R^{-1}(B^T K^+ + V)x^*(t)$, where $x^*(\cdot)$ solves $\dot{x}(t) = (A - BR^{-1}(B^T K^+ + V))x(t)$, $x^*(0) = x_0$.
3. Let $Q_T = 0$, $T = \infty$, $\mathcal{U} = L_{2,e}^+$ and (A, B) be controllable. Then, $\inf_u J$ exists for all x_0 if (ARE) has a symmetric solution $K \leq 0$. Moreover, if this condition is satisfied there exists a u^* attaining this infimum if and only if $\Delta \subset K^-$. Here K^- is the smallest solution of Γ and $\Delta := K^+ - K^-$, where K^+ is the largest solution of Γ . In that case $\min_u J(u) = x_0^T K_f x_0$ is attained uniquely by $u^*(t) = -R^{-1}(B^T K_f + V)x^*(t)$, where $x^*(\cdot)$ solves $\dot{x}(t) = (A - BR^{-1}(B^T K_f + V))x(t)$, $x^*(0) = x_0$ and $K_f = K^- P_N + K^+(I - P_N)$. Here P_N is the projector onto the subspace N along $\Delta^{-1}N^\perp$ with N the undetectable (w.r.t. $\mathbb{C}^- \cup \mathbb{C}^0$) subspace of $(K^-, (A - BR^{-1}(B^T K^- + V)))$. $\square$

Remark 0.36 If (ARE) has a solution $K \geq 0$ then K_f in Theorem 0.35.3) is the smallest positive semi-definite solution of (ARE).

If one merely assumes (A, B) to be stabilizable instead of controllable in Theorem 0.35.3) (ARE) has in general not a smallest solution anymore. It can still be shown under the same condition as in 3) (that (ARE) has a solution $K \leq 0$) that the infimum exists and equals $x_0^T K_f x_0$ for some $K_f \in \Gamma$. However there does not exist a nice characterization like in 3) of K_f (see [10]). $\square$

Next introduce $B := [B_1 \ B_2]$. Combining the results of Lemmas 0.2, 0.13 and Theorems 0.34, 0.35 one can derive now straightforwardly existence and computational algorithms for both the finite and infinite horizon game problems. We will just state the result for the infinite planning horizon case where the state converges to zero. The formulation of the other results is left to the reader.

Corollary 0.37 Consider the cooperative game (29,43) with $T = \infty$, $\mathcal{U} = L_{2,e,s}^+$ and (A, B) stabilizable.

Assume (ARE), with $V := [V_i \ W_i]$ and $R := \begin{bmatrix} R_{1i} & N_i \\ N_i^T & R_{2i} \end{bmatrix}$ has a stabilizing⁵

⁵That is, $\sigma(A - BR^{-1}V^T - SK_i) \subset \mathbb{C}^-$.

solution K_i , for $i = 1, 2$, respectively.

Then all Pareto efficient solutions are obtained by determining for $\alpha \in \mathcal{A}$

$$u^*(\alpha) := \arg \min_{u \in \mathcal{U}} \alpha_1 J_1 + \alpha_2 J_2, \text{ subject to (43).} \quad (50)$$

The corresponding Pareto solutions are

$$(J_1(u^*(\alpha)), J_2(u^*(\alpha))). \quad \square$$

Notice that, since by assumption both cost functions are strict convex, the minimization problem (50) has for all $\alpha \in \mathcal{A}$ a unique solution. From this result one obtains then the next procedure to calculate all Pareto efficient outcomes for this game.

Theorem 0.38 (Solution Cooperative Game)

Consider the cooperative game (29,43) with $T = \infty$, $\mathcal{U} = L_{2,e,s}^+$ and (A, B) stabilizable. For $\alpha \in \mathcal{A}$ let

$$M(\alpha) := \alpha_1 M_1 + \alpha_2 M_2 =: \begin{bmatrix} \tilde{Q} & \tilde{V} \\ \tilde{V}^T & \tilde{R} \end{bmatrix} \text{ where } \tilde{V} = [\alpha_1 V_1 + \alpha_2 V_2, \alpha_1 W_1 + \alpha_2 W_2],$$

$$\tilde{Q} = \alpha_1 Q_1 + \alpha_2 Q_2, \text{ and } \tilde{R} = \alpha_1 \begin{bmatrix} R_{11} & N_1 \\ N_1^T & R_{21} \end{bmatrix} + \alpha_2 \begin{bmatrix} R_{12} & N_2 \\ N_2^T & R_{22} \end{bmatrix}.$$

Furthermore, let $\tilde{S} := B\tilde{R}^{-1}B^T$.

Assume that (51) below has a stabilizing solution X_i for $\alpha_i = 1$, $i = 1, 2$, respectively.

$$A^T X + XA - (XB + \tilde{V})\tilde{R}^{-1}(B^T X + \tilde{V}^T) + \tilde{Q} = 0. \quad (51)$$

Then the set of all cooperative Pareto solutions is given by

$$\{(J_1(u^*(\alpha)), J_2(u^*(\alpha))) \mid \alpha \in \mathcal{A}\}.$$

Here

$$u^*(t) = -\tilde{R}^{-1}(B^T X_s + \tilde{V}^T)x(t) - \tilde{R}^{-1}B^T \int_t^\infty e^{-A_{cl}^T(t-s)} X_s c(s) ds$$

where X_s is the stabilizing solution of (51) and, with $A_{cl} := A - B\tilde{R}^{-1}\tilde{V}^T - \tilde{S}X_s$, the closed-loop system is $\dot{x}(t) = A_{cl}x(t) - Sm(t) + c(t)$, $x(0) = x_0$. In case $c(\cdot) = 0$ the with these actions corresponding cost are $J_i(x_0, u^*) = x_0^T \tilde{M}_i x_0$, where $\tilde{M}_i$ is the unique solution of the Lyapunov equation

$$A_{cl}^T \tilde{M}_i + \tilde{M}_i A_{cl} = -[I, -(X_s B + \tilde{V})\tilde{R}^{-1}] M_i \begin{bmatrix} I \\ -\tilde{R}^{-1}(B^T X_s + \tilde{V}^T) \end{bmatrix}. \quad \square$$

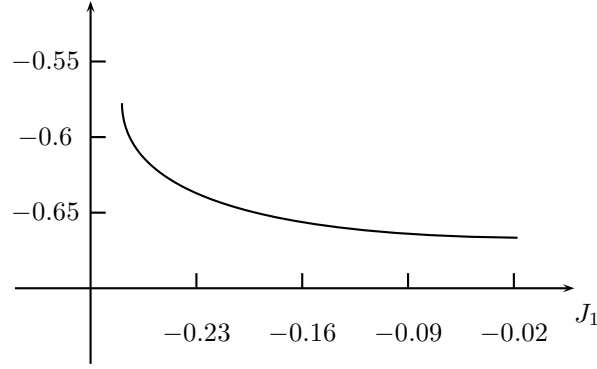


Figure 1: Pareto frontier Example 0.39.

Example 0.39 To illustrate the above result consider the next simplistic environmental example. Consider a fishery management game, where two industries pollute a lake and their income depends on the fish-stock of the lake. Assume that both the revenues and cost depend quadratically on the fish-stock and the produced pollution. Within this context we consider the next model:

$$\dot{x}(t) = -2x(t) - u_1(t) - u_2(t), \quad x(0) = 1, \quad (52)$$

with cost functions

$$J_1 = \int_0^\infty -x^2(t) + 2u_1^2(t) + u_2^2(t)dt \quad \text{and} \quad J_2 = \int_0^\infty -2x^2(t) + u_1^2(t) + 2u_2^2(t)dt.$$

Here x is the fish-stock above the supposed natural equilibrium level in the lake and u_i is the pollution produced by industry i . The factor -2 in (52) is the resultant of the harvesting and growth rate of the fish-stock. The fact that industry 1 gives more weight to the pollution it produces than to the quantity produced by the other industry can be interpreted as that industry 1 is really concerned about the pollution it produces itself (and which it is able to control in contrast to the pollution produced by the other industry). In figure 1 we plotted the Pareto frontier if both players would cooperate in this game. $\square$

THE COOPERATIVE GAME

As we motivated in the introduction the set of Pareto solution(s) constitute a natural set of outcomes to consider when players cooperate. As we illustrated in the previous sections in a number of examples in a two-player game usually the Pareto frontier will be a one-dimensional surface in $\mathbb{R}^2$. However, this does not always have to be the case. This can, e.g., already be illustrated by a simple two-player scalar linear quadratic game. Assuming both players have the same cost function (with positive weights) it is easily verified that in that case the Pareto

frontier reduces to a single point in $\mathbb{R}^2$.

Assuming that in a cooperative setting more than one solution exists, a natural question that arises is which solution of all these possibilities will be chosen by the players. Since we assume an open-loop information structure of the game we can use concepts developed within the area of static game theory to motivate the choice for a certain solution. Within the area of static cooperative game theory one usually distinguishes two types of games: the transferable utility (TU) games and the nontransferable utility (NTU) games⁶.

By working together in coalitions, players can generate benefits. A typical example is that of a number of firms cooperating in order to save costs. Not only is it interesting to know how players can cooperate in an optimal way, but also the problem of allocation arises. TU-games were introduced by [29] and can be seen as an allocation problem in which an amount of money is to be divided and can be freely transferred between the players. Here one abstracts from the fact that the involved players may value the obtained payoffs differently.

Generally, the analysis of a TU-game focuses on how to allocate the joint benefits/costs of the grand coalition of all players, where the values of subcoalitions serve as a benchmark. In order to allocate the value of the grand coalition several one-point solutions, each with its own appealing properties, have been proposed in the literature. As examples we mention the Shapley value [23], the compromise value [26] and the nucleolus [19]. An important concept in the allocation of the value of a grand coalition is core stability. The core of a (N)TU-game consists of those allocation vectors such that no subcoalition has an incentive to split off. Hence, if the TU-game measures the benefits of each coalition, an allocation is an element of the core if for each coalition its members receive together at least as much as they could achieve when they would act as a separate coalition.

NTU games, introduced by [2], on the other hand assume that there is no common measure of utility. That is, the objects to be divided are not valued in the same way by all the players. Consequently, no side-payments take place. Within the NTU-games one discerns the so-called bargaining problems. For these games one assumes that only the set of feasible outcomes for the individual players and a so-called threatpoint is specified. This threatpoint (or disagreement point) is a value every individual player can obtain by playing noncooperatively. Below we will elaborate three bargaining solutions which have been proposed in literature. Bargaining theory has its origin in two papers by Nash [16] and [17]. In these papers a bargaining problem is defined as a situation in which two (or more) individuals or organizations have to agree on the choice of one specific alternative from a set of alternatives available to them, while having conflicting interests over this set of alternatives. Nash proposes in [17] two different approaches to the bargaining problem, namely the *axiomatic* and the *strategic* approach. The axiomatic approach lists a number of desirable properties the solution must have, called the *axioms*. The strategic approach on the other hand, sets out a particular bargaining procedure

⁶Note that from a formal point of view a TU-game can be viewed as a special case of a NTU-game.

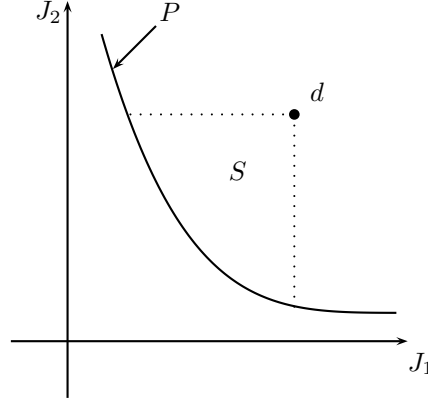


Figure 2: The bargaining game.

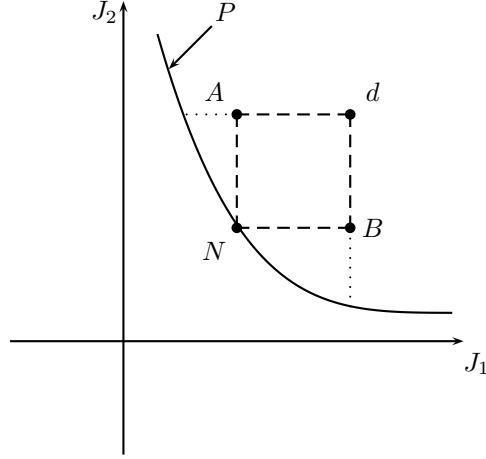
and asks what outcomes would result from rational behavior by the individual players.

So, bargaining theory deals with the situation in which players can realize -through cooperation- other (and better) outcomes than the one which becomes effective when they do not cooperate. This non-cooperative outcome is called the *threatpoint*. The question is to which outcome the players may possibly agree.

In Figure 2 a typical bargaining game is sketched (see also Figure 1). The inner part of the ellipsoid marks out the set of possible outcomes, the *feasible set* S , of the game. The point d is the threatpoint. The edge P is the set of individually rational Pareto-optimal outcomes.

We assume that if the agents unanimously agree on a point $x = (J_1, J_2) \in S$, they obtain x . Otherwise they obtain d . This presupposes that each player can enforce the threatpoint, when he does not agree with a proposal. The outcome x the players will finally agree on is called the solution of the bargaining problem. Since the solution depends on as well the feasible set S as the threatpoint d , it will be written as $F(S, d)$. Notice that the difference for player i between the solution and the threatpoint, $J_i - d_i$, is the reduction in cost player i incurs by accepting the solution. In the sequel we will call this difference the utility gain for player i . We will use the notation $J := (J_1, J_2)$ to denote a point in S and $x \succ y$ ($x \prec y$) to denote the *vector inequality*, i.e. $x_i > y_i$ ($x_i < y_i$), $i = 1, 2$. In axiomatic bargaining theory a number of solutions have been proposed. In Thomson [25] a survey is given on this theory. We will present here the three most commonly used solutions: the Nash bargaining solution, the Kalai-Smorodinsky solution and the Egalitarian solution.

The *Nash bargaining solution*, $N(S, d)$, selects the point of S at which the

Figure 3: The Nash Bargaining solution $N(S, d)$.

product of utility gains from d is maximal. That is,

$$N(S, d) = \arg \max_{J \in S} \prod_{i=1}^N (J_i - d_i), \text{ for } J \in S \text{ with } J \preceq d.$$

In Figure 3 we sketched the N solution. Geometrically, the Nash Bargaining solution is the point on the edge of S (that is a part of the Pareto frontier) which yields the largest rectangle (N, A, B, d) .

The Kalai-Smorodinsky solution, $K(S, d)$, sets utility gains from the threatpoint proportional to the player's most optimistic expectations. For each agent, the most optimistic expectation is defined as the lowest cost he can attain in the feasible set subject to the constraint that no agent incurs a cost higher than his coordinate of the threatpoint. Defining the *ideal point* as

$$I(S, d) := \max\{J_i \mid J \in S, J \preceq d\},$$

the *Kalai-Smorodinsky solution* is then

$$K(S, d) := \text{maximal point of } S \text{ on the segment connecting } d \text{ to } I(S, d).$$

In Figure 4 the Kalai-Smorodinsky solution is sketched for the two-player case. Geometrically, it is the intersection of the Pareto frontier P with the line which connects the threatpoint and the ideal point. The components of the ideal point are the minima each player can reach when the other player is fully altruistic under cooperation.

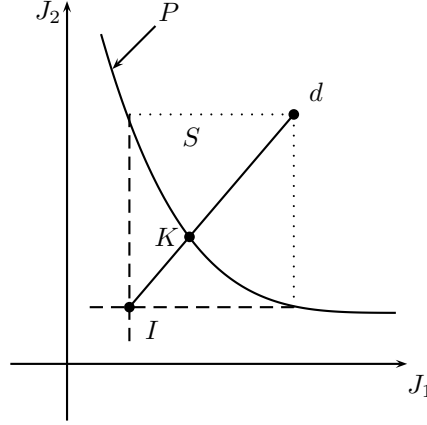


Figure 4: The Kalai-Smorodinsky solution $K(S, d)$.

Finally, the *Egalitarian solution*, $E(S, d)$, represents the idea that gains should be equal divided between the players. Formal, $E(S, d) :=$

maximal point in S for which $E_i(S, d) - d_i = E_j(S, d) - d_j$, $i, j = 1, \dots, N$.

Again, we sketched this solution for the two-player case. In Figure 5 we observe that geometrically this Egalitarian solution is obtained as the intersection point of the 45°-line through the threatpoint d with the Pareto frontier P .

Notice that in particular in contexts where interpersonal comparisons of utility is inappropriate or impossible, the first two bargaining solutions still make sense.

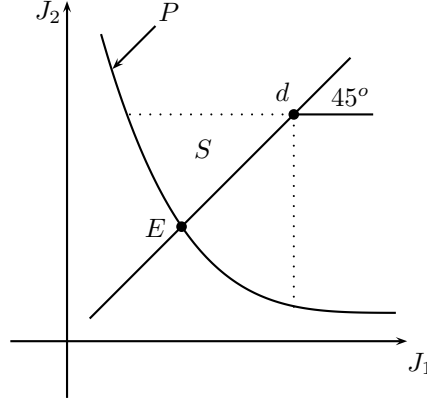
As already mentioned above these bargaining solutions can be motivated using an "axiomatic approach". In this case some people prefer to speak of an arbitration scheme instead of a bargaining game. An arbiter draws up the reasonable axioms and depending on these axioms, a solution results.

Algorithms to calculate the first two solutions numerically are outlined in [5]. The calculation of the Egalitarian solution requires the solution of one non-linear constrained equations problem. The involved computer time to calculate this E -solution approximately equals that of calculating the N -solution.

Example 0.40 Consider the following differential game on government debt stabilization (see van Aarle et al. [1]). Assume that government debt accumulation, $\dot{d}(t)$, is the sum of interest payments on government debt, $rd(t)$, and primary fiscal deficits, $f(t)$, minus the seignorage (i.e. the issue of base money) $m(t)$. So,

$$\dot{d}(t) = rd(t) + f(t) - m(t), d(0) = d_0.$$

Here $d(t)$, $f(t)$ and $m(t)$ are expressed as fractions of GDP and r represents the rate of interest on outstanding government debt minus the growth rate of output.

Figure 5: The Egalitarian solution $E(S, d)$.

The interest rate $r > 0$ is assumed to be exogenous. Assume that fiscal and monetary policies are controlled by different institutions, the fiscal authority and the monetary authority, respectively, which have different objectives. The objective of the fiscal authority is to minimize a sum of time profiles of the primary fiscal deficit, base-money growth and government debt

$$J_1 = \int_0^\infty e^{-\delta t} \{f^2(t) + \eta m^2(t) + \lambda d^2(t)\} dt.$$

The parameters, η and λ express the relative priority attached to base-money growth and government debt by the fiscal authority. The monetary authorities are assumed to choose the growth of base money such that a sum of time profiles of base-money growth and government debt is minimized. It is assumed that they (almost) do not care about the fiscal deficit. That is

$$J_2 = \int_0^\infty e^{-\delta t} \{\epsilon f^2(t) + m^2(t) + \kappa d^2(t)\} dt.$$

Here $1/\kappa$ can be interpreted as a measure for the conservatism of the central bank w.r.t. the money growth and ϵ is almost zero. Furthermore all variables are normalized such that their targets are zero, and all parameters are positive.

Introducing $\tilde{d}(t) := e^{-\frac{1}{2}\delta t} d(t)$, $\tilde{m} := e^{-\frac{1}{2}\delta t} m(t)$ and $\tilde{f} := e^{-\frac{1}{2}\delta t} f(t)$ the above model can be rewritten as

$$\dot{\tilde{d}}(t) = (r - \frac{1}{2}\delta)\tilde{d}(t) + \tilde{f}(t) - \tilde{m}(t), \quad \tilde{d}(0) = d_0.$$

Where the cost functions of both players are

$$J_1 = \int_0^\infty \{\tilde{f}^2(t) + \eta \tilde{m}^2(t) + \lambda \tilde{d}^2(t)\} dt$$

and

$$J_2 = \int_0^\infty \{\epsilon \tilde{f}^2(t) + \tilde{m}^2(t) + \kappa \tilde{d}^2(t)\} dt.$$

If both the monetary and fiscal authority agree to cooperate in order to reach their goals, by Theorem 0.13 the set of all Pareto solutions is obtained by considering the simultaneous minimization of

$$\begin{aligned} J_c(\alpha) &:= \alpha J_1 + (1 - \alpha) J_2 \\ &= \int_0^\infty \{\alpha_1 \tilde{f}^2(t) + \beta_1 \tilde{m}^2(t) + \beta_2 \tilde{d}^2(t)\} dt, \end{aligned}$$

where $\alpha_1 = \epsilon + \alpha(1 - \epsilon)$, $\beta_1 = 1 + \alpha(-1 + \eta)$ and $\beta_2 = \kappa + \alpha(\lambda - \kappa)$. This cooperative game problem can be reformulated as the minimization of

$$J_c(\alpha) = \int_0^\infty \{\beta_2 \tilde{d}^2(t) + [\tilde{f} \ \tilde{m}] \begin{bmatrix} \alpha & 0 \\ 0 & \beta_1 \end{bmatrix} \begin{bmatrix} \tilde{f} \\ \tilde{m} \end{bmatrix}\} dt,$$

subject to

$$\dot{\tilde{d}}(t) = (r - \frac{1}{2}\delta)\tilde{d}(t) + [1 \ -1] \begin{bmatrix} \tilde{f} \\ \tilde{m} \end{bmatrix}, \quad \tilde{d}(0) = d_0.$$

In Figure we plotted the set of Pareto solutions in case $\eta = 0.1$; $\lambda = 0.6$; $\epsilon = 0.00001$; $\kappa = 0.5$; $r = 0.06$, $\delta = 0.04$ and $d(0) = 1$. Furthermore we plotted the corresponding N , K and E solution if the threatpoint is $(2, 1)$. For these parameters we also calculated the corresponding non-cooperative open-loop Nash equilibrium (see e.g. [5, Section 7.7]). This solution is given by $(J_1, J_2) = (0.4658, 0.7056)$. Taking this as the threatpoint, we illustrated in Figure the new (more realistic) bargaining solutions. Furthermore we tabulated in Table 1 below the percentage gains for the fiscal and monetary authority if they would cooperate instead of playing non-cooperative. Assuming that real society's cost are measured by the average of the cost of the fiscal and monetary authority we observe from Table 1 that cooperation leads to a reduction of society's cost with approximately 30%.

Solution	Coordinates	% gain J_1	% gain J_2
N	(0.3383, 0.471)	27.4	33.2
K	(0.3445, 0.4582)	26.0	35.1
E	(0.3755, 0.42)	19.4	40.5

Table 1: Percentage gains of cooperation

□

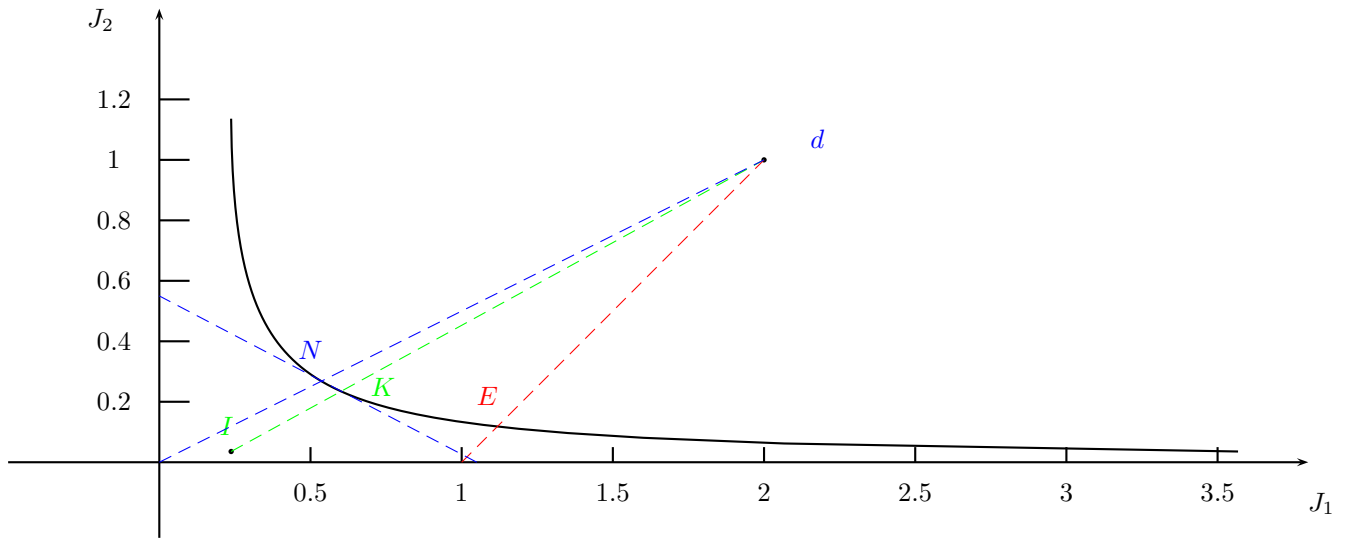


Figure 6: Pareto frontier Example 0.40 if $\eta = 0.1$; $\lambda = 0.6$; $\kappa = 0.5$; $r = 0.06$ and $\delta = 0.04$.

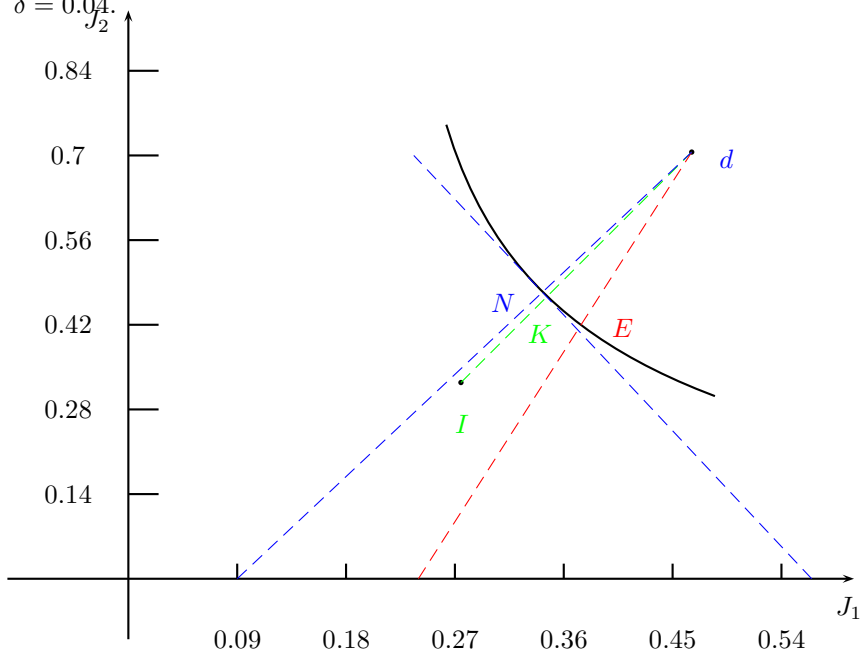


Figure 7: Bargaining solutions Example 0.40 if threatpoint is $(0.4658, 0.7056)$.

CONCLUDING REMARKS

In this paper we considered both necessary and sufficient conditions for the existence of a Pareto solution in a cooperative dynamic differential game with an open-loop information structure. This problem setting naturally occurs if either one person has more than one objectives he likes to optimize, or, the dynamics of the considered system is affected by more than one person and these persons agree to cooperate in order to optimize their individual objective.

The presented necessary condition is a mixture of the maximum principle conditions and the necessary conditions for existence of a Pareto solution in a static game. In line with Arrow's conditions, we also gave conditions under which the necessary conditions are sufficient too. Finally we indicated the effects on the necessary conditions if the problem setting includes inequality constraints.

In the second part of this paper we considered the special case of regular indefinite linear quadratic control problems. First we elaborated the necessary conditions for the general case and next we considered the special case when the game is convex.

For the general case we showed how one can obtain all Pareto efficient outcomes for scalar systems if the game has Pareto solutions for every initial state. Unfortunately, the generalization of these results to non-scalar systems remains an open problem. We illustrated in an example some peculiarities that may occur.

For the convex game we showed how one can obtain all Pareto efficient outcomes. Using the theory of convex analysis we were able to treat both the finite and infinite planning horizon problems in a uniform way. For the infinite planning horizon we considered both the fixed endpoint problem and the free endpoint problem.

We believe that without too many complications the presented theory on the general case can be extended for an infinite planning horizon. The elaboration of the herewith involved technical problems remains however a point for future research. Another open problem that remains to be solved within this area is the case that the performance criteria of both players are not directly affected by the control efforts used by the other player (giving rise to singular control problems). Open problems that remain to be solved for the convex case are, e.g., the free endpoint infinite planning horizon problem under the assumption that the system is merely stabilizable.

Finally, in the third part of this paper, we considered the question which cooperative solution will result in a multi-player context. Given our open-loop information structure we argued that it makes sense to use solution concepts developed in static game theory to answer this question. In particular we reviewed some axiomatic bargaining solutions which have been developed for non-transferable utility games. We illustrated them in an example.

*

REFERENCES

- [1] Aarle B. van, Bovenberg L. and Raith M., 1995, Monetary and fiscal policy interaction and government debt stabilization, *Journal of Economics*, Vol.62, pp.111-140.
- [2] Aumann R. and Peleg B., 1960, Von Neumann-Morgenstern solutions to cooperative games without side payments, *Bulletin of the American Mathematical Society*, vol.66, pp.173-179.
- [3] Clements D.J. and Anderson B.D.O., 1978, *Singular Optimal Control: The Linear-Quadratic Problem*, Springer-Verlag, Berlin.
- [4] Coddington E.A. and Levinson N., 1955, *Theory of Ordinary Differential Equations*, McGraw Hill, New York.
- [5] Engwerda J.C., 2005, *LQ Dynamic Optimization and Differential Games*, Wiley, Chichester.
- [6] Engwerda J.C., 2006, Linear Quadratic Differential Games: an overview, CD-Rom Proceedings 12th International Symposium on Dynamic Games and Applications, 3-6th July 2006, Sophia Antipolis, (also: CentER Discussion Paper DP 2006-110, online available via: <http://greywww.kub.nl:2080/greyfiles/center/2006/110.html>, Tilburg University, The Netherlands), to appear in *Annals of Dynamic Games*.
- [7] Engwerda J.C., 2007, A note on cooperative linear quadratic control, CentER Discussion Paper 2007-15, Tilburg University, The Netherlands. (Revised version will appear in *Automatica* as: The regular convex cooperative linear quadratic control problem).
- [8] Engwerda J.C., 2007, Necessary and sufficient conditions for solving cooperative differential games, Internal Report, Tilburg University, The Netherlands.
- [9] Fan K., Glicksberg I. and Hoffman A.J., 1957, Systems of inequalities involving convex functions, *American Mathematical Society Proceedings*, Vol.8.
- [10] Geerts T., 1991, A priori results in linear-quadratic optimal control theory, *Kybernetika*, vol.27, pp.446-457.
- [11] Jacobson D.H., 1971, A general sufficiency theorem for the second variation, *J. Math. Anal. Appl.*, vol.34, pp.578-589.
- [12] Lee E.B. and Markus L., 1967, *Foundations of Optimal Control Theory*, John Wiley & Sons, Inc., New York.
- [13] Leitmann G., 1974, *Cooperative and Non-cooperative Many Players Differential Games*, Springer Verlag, Berlin.

- [14] Molinari B.P., 1975, Nonnegativity of a quadratic functional, *SIAM J. Control and Optimization*, vol.13, pp.792-806.
- [15] Molinari B.P., 1977, The time-invariant linear-quadratic optimal control problem, *Automatica*, vol.13, pp.347-357.
- [16] Nash J., 1950, The bargaining problem, *Econometrica*, Vol.18, pp.155-162.
- [17] Nash J., 1953, Two-person cooperative games, *Econometrica*, Vol.21, pp.128-140.
- [18] Pars L.A., 1962, *An Introduction to the Calculus of Variations*, Heinemann, London.
- [19] Schmeidler D., 1969, The nucleolus of a characteristic function game, *SIAM Journal on Applied Mathematics*, vol.17, pp.1163-1170.
- [20] Seierstad A. and Sydsaeter K., 1987, *Optimal Control Theory with Economic Applications*, North-Holland, Amsterdam.
- [21] Sethi S.P. and Thompson G.L., 2000, *Optimal Control Theory*, Kluwer Academic Publishers, London.
- [22] Simon C.P. and Blume L., 1994, *Mathematics for Economists*, Norton & Company, New York.
- [23] Shapley L., 1953, A value for n -person games, *Annals of Mathematics Studies*, vol.28, pp.307-317.
- [24] Takayama A., 1985, *Mathematical Economics*, Cambridge University Press, Cambridge.
- [25] Thomson W., 1994, Cooperative models of bargaining, In Aumann and Hart (Eds.), *Handbook of Game Theory*, Vol.2, Elsevier Science, pp.1238-1277.
- [26] Tijs S., 1981, *Bounds for the core and the τ -value*. In O. Moeschlin and D. Pallaschke (Eds.), *Game Theory and Mathematical Economics*, North-Holland, Amsterdam, pp.123-132.
- [27] Trentelman H.L., 1989, The regular free-endpoint linear quadratic problem with indefinite cost, *SIAM J. Control and Optimization*, vol.27, pp.27-42.
- [28] Trentelman H.L. and Willems J.C., 1991, The dissipation inequality and the algebraic Riccati equation. In S. Bittanti, A.J. Laub and J.C. Willems (Eds.), *The Riccati Equation*, Springer Verlag, Berlin, pp.197-242.
- [29] Von Neumann J. and Morgenstern O., 1944, *Theory of Games and Economic Behavior*, Princeton University Press, Princeton.

- [30] Willems J.C., 1971, Least squares stationary optimal control and the algebraic Riccati equation, IEEE Transactions on Automatic Control, vol.16, pp.641-634.
- [31] Yeung D.W.K and Petrosyan L.A., 2005, Subgame consistent solutions of a cooperative stochastic differential game with nontransferable payoffs, Journal of Optimization Theory and Applications, vol.124, pp.701-724.

JACOB ENGWERDA: Tilburg University; Dept. of Econometrics and O.R.; P.O. Box: 90153;
5000 LE Tilburg; The Netherlands
E-mail: engwerda@uvt.nl

ON THE BOOTSTRAP ACCURACY FOR THE TRIMMED MEAN: A SURVEY OF SOME RECENT DEVELOPMENTS

N.V. GRIBKOVA AND R. HELMERS

Abstract. A survey of some new results on empirical Edgeworth expansions and the M out of N bootstrap accuracy for trimmed means and studentized trimmed means will be presented.

We discuss the second order correctness of the M out of N (N being the real data sample size, M denotes the size of the bootstrap resample) bootstrap, provided the trimmed mean is properly defined and a local smoothness condition on the underlying distribution is satisfied. We also assume a relation between M and N , as $\min(M, N)$ gets large.

In practical applications, the M out of N bootstrap requires extensive Monte Carlo simulations. To reduce the computation time one can take M of smaller order than N , i.e. M/N approaches zero as $\min(M, N)$ gets large, and apply the bootstrap in conjunction with extrapolation.

Key words and Phrases: bootstrap accuracy, Empirical Edgeworth expansions, Studentized trimmed means.

1. INTRODUCTION

The trimmed mean is a well known estimator of a location parameter. Its asymptotic properties (and in more general setting the properties of the truncated L -statistics) were studied by many authors (see Chernoff et. al [7], Tukey & McLaughlin [27], Shorack [25], Stigler [26], Bickel [2], Bjerre [5]–[6], De Wet [28], Helmers [18]–[20], van Zwet [30], Deheuvels et. al [9], Csörgő et. al [8], Griffin & Pruitt [14].

2000 Mathematics Subject Classification: Primary 62E20, 62G30, 62G09; Secondary 60F05.

The main reason for applying the trimmed mean is robustness (see Huber [22], Hampel et. al [17]).

The centre of a distribution is often estimated by the sample mean or the sample median. However, it is well known that the sample mean is sensitive to outliers and thus not robust. On the other hand, the sample median is robust against outliers but not very efficient if the underlying distribution is, for instance, normal. An estimator showing intermediate behavior, and which includes both the sample mean and sample median, is the trimmed (sample) mean. Compared with robust M -estimates of maximum likelihood type, the trimmed mean not only has the same asymptotic variance but also is easy to compute.

We also want to mention a historical note by van Zwet [29], in which he describes unpublished work done in the Netherlands during the Second World War by van de Hulst – later a famous astronomer – on the trimmed mean and its close relation to an appropriate M -estimator.

The limit distribution of the trimmed mean for an arbitrary population distribution was found by Stigler [26]. Specifically he has shown that in order for the trimmed mean to be asymptotically normal, it is necessary and sufficient that the sample is trimmed at sample quantiles for which the corresponding population quantiles are uniquely defined.

In this paper we briefly survey some recent results on the second order approximations to the distribution function of the trimmed mean and Studentized trimmed mean: the Edgeworth expansion (EE), empirical EE and the bootstrap. Our focus in this paper is on the classical case of trimming a fixed fraction (percentage) of the observations. The case of slightly trimmed means, i.e. trimming a vanishing fraction of the data is also of much interest. We hope to pursue this matter elsewhere.

Consider a sequence $X_1, X_2, \dots$ of independent and identically distributed (i.i.d.) real-valued random variables (r.v.) with common distribution function (df) F , and let $X_{1:N} \leq \dots \leq X_{N:N}$ ($N = 1, 2, \dots$) be the corresponding order statistics.

Let $F^{-1}(u) = \inf\{x : F(x) \geq u\}$, $0 < u \leq 1$, denotes the left-continuous inverse function of the df F and put F_N^{-1} to be the inverse of F_N .

Define two versions of the trimmed mean by

$$T_{N(1)} = \frac{1}{([\beta N] - [\alpha N])} \sum_{i=[\alpha N]+1}^{[\beta N]} X_{i:N}, \quad T_{N(2)} = \frac{1}{\beta - \alpha} \int_{\alpha}^{\beta} F_N^{-1}(u) du, \quad (1)$$

where $0 < \alpha < \beta < 1$ are any fixed numbers, $[\cdot]$ represents the greatest integer function. Note that $T_{N(1)}$ is the natural definition of a trimmed mean; and $T_{N(2)}$ (the 'plug-in version' of a trimmed mean) — only slightly different from the classical $T_{N(1)}$ — is very useful in a bootstrap context, in particular, it enables us to obtain the second order correctness property for the M out of N bootstrap for a (Studentized) trimmed mean $T_{N(2)}$ (cf. Gribkova & Helmers [12]).

Let $X_1^*, \dots, X_M^*$ be a bootstrap resample of the size $M = M(N)$ from the empirical df F_N based on the first N original observations $X_1, \dots, X_N$; F_M^* denotes

the bootstrap empirical df , and $X_{1:M}^* \leq \dots \leq X_{M:M}^*$ — the corresponding order statistics. Here and throughout this article we use the shorthand notation M , omitting its argument N .

Define the M out of N bootstrap counterparts of the trimmed means by:

$$T_{N,M(1)}^* = \frac{1}{([\beta M] - [\alpha M])} \sum_{i=[\alpha M]+1}^{[\beta M]} X_{i:M}^*, \quad T_{N,M(2)}^* = \frac{1}{\beta - \alpha} \int_{\alpha}^{\beta} (F_M^*)^{-1}(u) du, \quad (2)$$

where α and β are same as in (1).

It is well known that both versions of trimmed mean (1) serve as statistical estimators of the location parameter

$$\mu(\alpha, \beta) = \frac{1}{\beta - \alpha} \int_{\alpha}^{\beta} F^{-1}(u) du, \quad (3)$$

the population trimmed mean. The bootstrap counterpart of the parameter $\mu(\alpha, \beta)$ is given by

$$\mu_N(\alpha, \beta) = \frac{1}{\beta - \alpha} \int_{\alpha}^{\beta} F_N^{-1}(u) du = T_{N(2)}. \quad (4)$$

Let us introduce the ν -th ($0 < \nu < 1$) quantile of F , F_N and (F_M^*) :

$$\xi_{\nu} = F^{-1}(\nu), \quad \xi_{\nu N:N} = F_N^{-1}(\nu), \quad \xi_{\nu M:M} = (F_M^*)^{-1}(\nu). \quad (5)$$

First we note that the following simple relations are valid:

$$\begin{aligned} T_{N(2)} &= c_{\alpha,N} \left(T_{N(1)} - \xi_{\alpha N:N} \right) + T_{N(1)} - c_{\beta,N} \left(T_{N(1)} - \xi_{\beta N:N} \right), \\ T_{N,M(2)}^* &= c_{\alpha,M} \left(T_{N,M(1)}^* - \xi_{\alpha M:M}^* \right) + T_{N,M(1)}^* - c_{\beta,M} \left(T_{N,M(1)}^* - \xi_{\beta M:M}^* \right), \end{aligned} \quad (6)$$

with $c_{\nu,L} = \frac{\nu L - [\nu L]}{(\beta - \alpha)L} = O(1/L)$, where $L = N, M$ and $\nu = \alpha, \beta$. Since $\xi_{\nu N:N}$ and $T_{N(j)}$, $j = 1, 2$, are bounded uniformly in N with probability one (a.s.) as $N \rightarrow \infty$ and in view of (6)

$$N^{1/2}(T_{N(2)} - T_{N(1)}) \rightarrow 0, \quad \text{a.s., as } N \rightarrow \infty. \quad (7)$$

Moreover, since $\xi_{\nu M:M}^*$, $\nu = \alpha, \beta$, and $T_{N,M(j)}^*$, $j = 1, 2$, are bounded uniformly with P^* -probability of the order $1 - O(M^{-c})$, for every $c > 0$, a.s. [P], (cf. Gribkova & Helmers [12]),

$$M^{1/2}(T_{N,M(2)}^* - T_{N,M(1)}^*) \rightarrow_{P^*} 0, \quad \text{a.s., as } \min(N, M) \rightarrow \infty. \quad (8)$$

Here and throughout this paper P^* denotes bootstrap probability measure having discrete mass points $X_{i:N}$ with atoms $1/N$, and E^* denotes the corresponding expectation.

The relations (7) – (8) imply that the limit distributions of both versions of the trimmed mean, given by (1), are identical; the same remark is valid for the corresponding bootstrap counterparts. Therefore, we can use in the sequel with impunity the following shorthand notations:

$$T_N = T_{N_{(j)}}, \quad T_{N,M}^* = T_{N,M_{(j)}}^*, \quad j = 1, 2,$$

unless otherwise stipulated. However, the second order asymptotic properties (one-term Edgeworth expansions) of two versions of the trimmed mean (cf. (1)) as well as two corresponding bootstrap counterparts (cf. (2)) are different (cf. Gribkova & Helmers [12]), and $T_{N_{(2)}}$ is the more appropriate definition of a trimmed mean in a bootstrap context (cf. Section 3).

2. EDGEWORTH EXPANSIONS

The existence of an Edgeworth expansion (EE) for a Studentized trimmed mean was obtained by Hall & Padmanabhan [16], but these authors did not obtain explicit formula for the one-term of EE. In papers by Gribkova & Helmers [11]-[12] the simple explicit formulas for the $N^{-1/2}$ - term (correcting for skewness and bias) of the Edgeworth expansion for a Studentized trimmed mean (in the 'real world') and the $M^{-1/2}$ - term for its bootstrap counterpart (in the 'bootstrap world') were established. The proof of these results is based on a U -statistic type approximation and also uses a version of Bahadur's [1] representation for sample quantiles and its bootstrap version. These EE's allow to prove that the M out of N bootstrap approximation to the df of the trimmed mean and Studentized trimmed mean is more accurate than normal one, provided a natural smoothness condition and a condition on the size of the bootstrap resample M in relation to size N are satisfied.

The main purpose of this short survey is to describe the recent results of Gribkova & Helmers [11]-[12] on the Edgeworth expansion and the M out of N bootstrap accuracy.

Let us introduce r.v.'s X_i Winsorized outside $(\xi_\alpha, \xi_\beta]$, in other words

$$W_i = \xi_\alpha \vee (X_i \wedge \xi_\beta), \quad i = 1, \dots, N, \quad (9)$$

where $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$, the corresponding quantile function is

$$Q(u) = \xi_\alpha \vee (F^{-1}(u) \wedge \xi_\beta).$$

Define

$$\mu_W = \int_0^1 Q(u) du, \quad \sigma_W^2 = \int_0^1 (Q(u) - \mu_W)^2 du, \quad (10)$$

the first two central moments of W_i .

We suppose throughout this paper that F^{-1} is continuous at the points α and β . This is the 'iff' condition for asymptotic normality of Stigler [26]. In this

case the limit distribution of the trimmed mean is the normal one with asymptotic variance $(\beta - \alpha)^2 \sigma_W^2$ (cf. (10)), it is the limit distribution of $\bar{W}_N$ – the average of W_i (cf. Gribkova & Helmers [11]–[12]), where it is shown that the first linear term of the Hoeffding decomposition of the trimmed mean is very close to $\bar{W}_N$.

We note in passing that if Stigler's 'iff' condition for asymptotic normality does not hold, the limit df becomes non normal. We refer to [26] and also to our forthcoming paper [13] for more details about the non normal case. It is shown in [13] that the M fewer than N bootstrap always works in this case as well.

To Studentize a trimmed mean suppose that $\xi_\alpha \neq \xi_\beta$ (that is ξ_α is not an atom of the df F with mass at least $(\beta - \alpha)$), then $\sigma_W^2 > 0$, and use a plug-in estimator of the asymptotic variance. Define empirical quantile function

$$Q_N(u) = \xi_{\alpha N:N} \vee (F_N^{-1}(u) \wedge \xi_{\beta N:N}), \quad 0 < u \leq 1, \quad (11)$$

and plug-in estimates

$$\mu_{W,N} = \int_0^1 Q_N(u) du, \quad \sigma_{W,N}^2 = \int_0^1 (Q_N(u) - \mu_{W,N})^2 du. \quad (12)$$

Define the distribution functions of the normalized and Studentized trimmed mean by

$$F_{T_N}(x) = P\left(\frac{N^{1/2}(T_N - \mu(\alpha, \beta))}{(\beta - \alpha)^{-1} \sigma_W} \leq x\right), \quad F_{T_N,S}(x) = P\left(\frac{N^{1/2}(T_N - \mu(\alpha, \beta))}{(\beta - \alpha)^{-1} \sigma_{W,N}} \leq x\right) \quad (13)$$

Define the bootstrap version Q_M^* of the quantile function Q_N by

$$Q_M^*(u) = \xi_{\alpha M:M}^* \vee ((F_M^*)^{-1}(u) \wedge \xi_{\beta M:M}^*), \quad 0 < u \leq 1. \quad (14)$$

The expectation and the variance of W_1^* are given by $\mu_{W,N}$ and $\sigma_{W,N}^2$ (cf. (12)).

Define the df 's of the normalized and Studentized M out of N bootstrapped trimmed mean by

$$\begin{aligned} F_{T_{N,M}^*}^*(x) &= P^*\left(\frac{M^{1/2}(T_{N,M}^* - T_{N(2)})}{(\beta - \alpha)^{-1} \sigma_{W,N}} \leq x\right), \\ F_{T_{N,M}^*,S}^*(x) &= P^*\left(\frac{M^{1/2}(T_{N,M}^* - T_{N(2)})}{(\beta - \alpha)^{-1} \sigma_{W,M}^*} \leq x\right), \end{aligned} \quad (15)$$

respectively, where

$$(\sigma_{W,M}^*)^2 = \int_0^1 (Q_M^*(u) - \mu_{W,M}^*)^2 du, \quad \mu_{W,M}^* = \int_0^1 Q_M^*(u) du,$$

with Q_M^* as in (14). We remind that $T_{N(2)}$ is equal to $\mu_N(\alpha, \beta)$ – the empirical counterpart of the parameter $\mu(\alpha, \beta)$ (cf. (3), (4)).

Note that instead of plug-in estimator for the asymptotic variance of the trimmed mean the jackknife estimator, and the bootstrap estimator can be used, which are also consistency, we will make this in details somewhere else.

2.1. The EE's in the real world. In the addition to the first two moments of the r.v. W_1 (cf, (10)) let us introduce its third central moment:

$$\gamma_W = \int_0^1 (Q(u) - \mu_W)^3 du. \quad (16)$$

To state the results on EE we will assume that following smoothness condition is satisfied:

(C.1) *Suppose that the distribution function F has a density $f = F'$ in neighborhoods of the points ξ_ν , with $f(\xi_\nu) > 0$, $\nu = \alpha, \beta$. In addition, we assume that the density f satisfies a Hölder's condition of the order $a > 0$ at the points ξ_α, ξ_β , i.e. there exists a constant C such that $|f(x) - f(\xi_\nu)| \leq C|x - \xi_\nu|^a$, for all x in the neighborhood of ξ_ν , $\nu = \alpha, \beta$.*

Define a quantity

$$\delta_W = -\alpha^2 \frac{1}{f(\xi_\alpha)} [\mu_W - \xi_\alpha]^2 + (1 - \beta)^2 \frac{1}{f(\xi_\beta)} [\mu_W - \xi_\beta]^2. \quad (17)$$

Define real numbers λ_1 and λ_2 by

$$\lambda_1 = \gamma_W / \sigma_W^3, \quad \lambda_2 = \delta_W / \sigma_W^3. \quad (18)$$

It was established in Gribkova & Helmers [11]-[12] that EE's for the df of the normalized trimmed means: $T_{N(j)}$, $j = 1, 2$, are given by

$$G_N(x) = \Phi(x) - \frac{\phi(x)}{6\sqrt{N}} \left((\lambda_1 + 3\lambda_2)(x^2 - 1) + 6\sqrt{N} \frac{b_N}{\sigma_W} \right), \quad (19)$$

where Φ is the standard normal distribution function, $\phi = \Phi'$ its density, b_N is the bias term. Moreover, the skewness terms (i.e. terms involving λ_1 and λ_2) for both versions of trimmed mean: $T_N = T_{N(j)}$, $j = 1, 2$, are identical, whereas the bias terms for $j = 1, 2$ are different. The bias term for $T_N = T_{N(2)}$ is given by

$$b_N = b_{N(2)} = \frac{1}{2\sqrt{N}} \left\{ -\alpha(1 - \alpha) \frac{1}{f(\xi_\alpha)} + \beta(1 - \beta) \frac{1}{f(\xi_\beta)} \right\}. \quad (20)$$

In the case $T_N = T_{N(1)}$ the bias term is equal to

$$b_N = b_{N(1)} = b_{N(2)} + b_{[,N]}, \quad (21)$$

with $b_{N(2)}$ as in (20) and

$$b_{[,N]} = \frac{1}{\sqrt{N}} \left\{ -(\alpha N - [\alpha N]) \left(\mu(\alpha, \beta) - \xi_\alpha \right) + (\beta N - [\beta N]) \left(\mu(\alpha, \beta) - \xi_\beta \right) \right\} \quad (22)$$

The EE's for the df of the Studentized trimmed means $T_{N(j)}$, $j = 1, 2$, are given by

$$H_N(x) = \Phi(x) + \frac{\phi(x)}{6\sqrt{N}} \left((2x^2 + 1)\lambda_1 + 3(x^2 + 1)\lambda_2 - 6\sqrt{N} \frac{b_N}{\sigma_W} \right), \quad (23)$$

where b_N is as in (19).

Following result means the validity of the one-term EE for the normalized and Studentized trimmed means $T_N = T_{N(j)}$, $j = 1, 2$.

Theorem 1. *Suppose that condition (C.1) holds true. Then*

$$\begin{aligned} \sup_{x \in \mathbb{R}} |F_{T_N}(x) - G_N(x)| &= o(N^{-1/2-p}), \\ \sup_{x \in \mathbb{R}} |F_{T_N,S}(x) - H_N(x)| &= o(N^{-1/2-p}), \end{aligned} \quad (24)$$

for every $p < \min(a/2, 1/4)$, as $N \rightarrow \infty$.

Note that if $a \geq 1/2$ (cf. the parameter appearing in the Hölder condition (C.1)), the order of remainders in (24) is $o(N^{-3/4+\epsilon})$ for every $\epsilon > 0$, i.e. the classical Bahadur's order.

To conclude this subsection, we note that a completely different way of approximating F_{T_N} and $F_{T_N,S}$ is to use saddlepoint approximations. These approximations will typically work better in the tail of the distribution in comparison with the Edgeworth expansions and bootstrap approximations considered in the present paper. Only very recently these saddlepoint expansions were established for the trimmed mean and the Studentized trimmed mean. We refer to Helmers et. al [20] for more details.

2.2. The EE's in the bootstrap world. Let us introduce third central moment of the r.v. W_1^* , i.e. the bootstrap counterpart of γ_W (cf. (16)):

$$\gamma_{W,N} = \int_0^1 (Q_N(u) - \mu_{W,N})^3 du,$$

where $\mu_{W,N}$ is defined as in (12).

For the validity of the EE's in the bootstrap world we will need a slightly stronger smoothness condition than (C.1) :

(C.1') *In addition to (C.1) assume that the density f satisfies a uniform Hölder condition of the order $a > 0$ in neighborhoods of the points ξ_α, ξ_β , i.e. there exists a constant C such that $|f(x) - f(y)| \leq C|x - y|^a$, for all x, y from the neighborhood of ξ_ν , $\nu = \alpha, \beta$.*

Moreover, the following condition on the rate at which M as a function of N is tending to infinity will be required:

$$(C.2) \quad M = O(N), \quad \text{as} \quad N \wedge M \rightarrow \infty.$$

In Gribkova & Helmers [12] the validity of the one-term EE in the bootstrap world was proved under a more general condition $M = O(N^{2-d})$ for some $0 < d < 2$. However now we prefer to formulate the results in their most simple and short form.

Define the M out of N bootstrap counterparts of δ_W , λ_1 and λ_2 (cf. (17)-(18):

$$\begin{aligned} \delta_{W,N} &= -\alpha^2 \frac{1}{f(\xi_{\alpha N:N})} [\mu_{W,N} - \xi_{\alpha N:N}]^2 + (1-\beta)^2 \frac{1}{f(\xi_{\beta N:N})} [\mu_{W,N} - \xi_{\beta N:N}]^2, \\ \lambda_{1,N} &= \gamma_{W,N} / \sigma_{W,N}^3, \quad \lambda_{2,N} = \delta_{W,N} / \sigma_{W,N}^3. \end{aligned} \quad (25)$$

It was proved in Gribkova & Helmers [12] that if conditions (C.1') and (C.2) hold true, then the Edgeworth expansion for the $d\mathbf{f}$ of a normalized version of the bootstrapped trimmed means $T_{N,M(j)}^*$, $j = 1, 2$, are given by

$$G_{N,M}(x) = \Phi(x) - \frac{\phi(x)}{6\sqrt{M}} \left((\lambda_{1,N} + 3\lambda_{2,N})(x^2 - 1) + 6\sqrt{M} \frac{b_{N,M}}{\sigma_{W,N}} \right). \quad (26)$$

As well as in real world the asymptotic expansions for $T_{N,M(2)}^*$ and $T_{N,M(1)}^*$ are identical in their skewness terms, but differ in the bias terms. The bias term in the case $T_{N,M}^* = T_{N,M(2)}^*$ is given by

$$b_{N,M} = b_{N,M(2)} = \frac{1}{2\sqrt{M}} \left\{ -\alpha(1-\alpha) \frac{1}{f(\xi_{\alpha N:N})} + \beta(1-\beta) \frac{1}{f(\xi_{\beta N:N})} \right\}, \quad (27)$$

and in the case $T_{N,M}^* = T_{N,M(1)}^*$ it is equal to

$$b_{N,M} = b_{N,M(1)} = b_{N,M(2)} + b_{[,N,M]},$$

where $b_{N,M(2)}$ as in (27), and

$$\begin{aligned} b_{[,N,M]} &= \frac{1}{\sqrt{M}} \left\{ -(\alpha M - [\alpha M]) \left(T_{N(2)} - \xi_{\alpha N:N} \right) \right. \\ &\quad \left. + (\beta M - [\beta M]) \left(T_{N(2)} - \xi_{\beta N:N} \right) \right\}. \end{aligned}$$

We remind that $T_{N(2)}$ is the bootstrap counterpart of the parameter $\mu(\alpha, \beta)$ (cf. (4), (22)).

The Edgeworth expansion for the bootstrapped Studentized trimmed mean is given by

$$H_{N,M}(x) = \Phi(x) + \frac{\phi(x)}{6\sqrt{M}} \left((2x^2 + 1)\lambda_{1,N} + 3(x^2 + 1)\lambda_{2,N} - 6\sqrt{M} \frac{b_{N,M}}{\sigma_{W,N}} \right). \quad (28)$$

where $b_{N,M}$ is given as in (26).

The following result is the M out of N bootstrap version of the Theorem 1.

Theorem 2. *Suppose that the conditions (C.1') and (C.2) hold true. Then*

$$\begin{aligned} \sup_{x \in \mathbb{R}} |F_{T_{N,M}}^*(x) - G_{N,M}(x)| &= o(M^{-1/2-p}), \\ \sup_{x \in \mathbb{R}} |F_{T_{N,M},S}^*(x) - H_{N,M}(x)| &= o(M^{-1/2-p}), \end{aligned}$$

for every $p < \min(a/2, 1/4)$, a.s. $[P]$, as $N \wedge M \rightarrow \infty$.

3. THE M OUT OF N BOOTSTRAP ACCURACY

The results of this subsection are direct consequences of Theorems 1 and 2 and bounds of the order the consistency (cf. Gribkova & Helmers [11]-[12] of the parameters appearing in formulas for one-term EE (26), (28) being the plug-in estimates of the corresponding parameters appearing in formulas (19), (23).

As second order properties of the trimmed means $T_{N(j)}$, $j = 1, 2$, are different in general, it will be convenient to use following notations: let now $F_{T_{N(j)}}(x)$, $j = 1, 2$, denote the df given by (13), where T_N is replaced by $T_{N(j)}$, and $F_{T_{N,M(j)}}^*(x)$ – the df defined by (15), where $T_{N,M}^*$ is replaced by $T_{N,M(j)}^*$, $j = 1, 2$.

It is evident from our results on EE that the second order correctness of the M out of N bootstrap approximations (i.e. the bound of the order $o(N^{-1/2})$ for the difference of df 's of the trimmed means in the real world and in the bootstrap world) can be achieved if and only if $M/N \rightarrow 1$, as $N \rightarrow \infty$, that is when $M = N + o(N)$ (if we do not extrapolation, see below). Moreover, if we assume a slightly stronger condition: $M = N + O(N^r)$ with some $0 \leq r < 1$, then $\frac{1}{\sqrt{M}} - \frac{1}{\sqrt{N}} = O(N^{-1/2-(1-r)})$. This bound together with Theorems 1 and 2 directly implies the following result on the M out of N bootstrap for the trimmed mean, given by $T_{N(2)}$ (cf.(1)).

Theorem 3. *Suppose that condition (C.1') is satisfied and let $M = N + O(N^r)$, $0 \leq r < 1$. Then*

$$\sup_{x \in \mathbb{R}} |F_{T_{N,M(2)}}^*(x) - F_{T_{N(2)}}(x)| = o(N^{-1/2-p}), \quad (29)$$

and

$$\sup_{x \in \mathbb{R}} |F_{T_{N,M(2)},S}^*(x) - F_{T_{N(2)},S}(x)| = o(N^{-1/2-p}), \quad (30)$$

for every $p < \min(a/2, 1 - r, 1/4)$, a.s. $[P]$, as $N \rightarrow \infty$.

This result means that the M out of N bootstrap approximation to the trimmed mean (given as $T_{N(2)}$, cf. (1)) is more accurate than the standard normal approximation. In addition, we want to emphasize the fact, useful in statistical practice, that bootstrap approximation is more accurate than the normal approximation not only in the case when $M = N$ (the naive bootstrap), but also in a more

general situation when size M of the bootstrap sample differs from size of the real data sample (but satisfies the condition $M = N + O(N^r)$, $r < 1$). Note that if condition (C.1') is satisfied with $a \geq 1/2$ and $r \leq 3/4$, then (29) – (30) provide the bounds of the order $o(N^{-3/4+\epsilon})$ for every $\epsilon > 0$.

However, we cannot achieve the same result for the trimmed mean given as $T_{N_{(1)}}$ (cf. [12]).

4. THE NAIVE BOOTSTRAP

Consider the important special case $M = N$ (the naive bootstrap). In this case the condition (C.1') implies now that the term $b_{[\cdot],N} - b_{[\cdot],N,N}$ (a part of the difference of EE's in the real world and in the bootstrap world for the trimmed mean given as (1), cf. Section 3) is of negligible order and contributes to a remainder term (cf. Gribkova & Helmers [12]). Therefore the one-term EE's from Section 2 imply the second order correctness of the naive bootstrap for $T_{N_{(1)}}$ as well as for $T_{N_{(2)}}$, and using again our joining notations T_N and $T_{N,M}^*$, we come to the following result, which is valid for both versus of the trimmed mean.

Theorem 4. *Suppose that the condition (C.1') is satisfied. Then for every $p < \min(a/2, 1/4)$*

$$\begin{aligned} \sup_{x \in R} |F_{T_{N,N}^*}^*(x) - F_{T_N}(x)| &= o(N^{-1/2-p}), \\ \sup_{x \in R} |F_{T_{N,N}^*,S}^*(x) - F_{T_{N,S}}(x)| &= o(N^{-1/2-p}), \end{aligned}$$

with probability one, as $N \rightarrow \infty$.

(cf. also Hall & Padmanabhan [16] for a similar result under somewhat stronger conditions)

5. ON THE EMPIRICAL EE, THE M OUT OF N BOOTSTRAP AND EXTRAPOLATION

In this section we will deal with the trimmed mean $T_N = T_{N_{(2)}}$ (cf. (1)) and its bootstrap version $T_{N,M}^* = T_{N,M_{(2)}}^*$ (cf. (2)). Define a quantity $b = \sqrt{N}b_N/\sigma_W = \frac{1}{2\sigma_W} \left(-\frac{\alpha(1-\alpha)}{f(\xi_\alpha)} + \frac{\beta(1-\beta)}{f(\xi_\beta)} \right)$, and denote the functions appearing in the Edgeworth expansions $G_N(x)$ and $H_N(x)$ (cf. (19) and (23)) by $B(x) = -\phi(x) \left[\frac{1}{6} (\lambda_1 + 3\lambda_2) \cdot (x^2 - 1) + b \right]$, and $B_S(x) = \phi(x) \left[\frac{1}{6} (\lambda_1(2x^2 + 1) + 3\lambda_2(x^2 + 1)) - b \right]$ respectively.

Hence, we know that

$$\begin{aligned} F_{T_N}(x) &= \Phi(x) + \frac{1}{\sqrt{N}}B(x) + R_{1,N}, \\ F_{T_N,S}(x) &= \Phi(x) + \frac{1}{\sqrt{N}}B_S(x) + R_{2,N} \end{aligned} \quad (31)$$

where $R_{1,N}$ and $R_{2,N}$ are the remainder terms of EE's. In Gribkova & Helmers [11] we estimate the unknown parameters appearing in $B(x)$ and $B_S(x)$ to obtain empirical EE's (EEE): $\widehat{G}_N(x) = \Phi(x) + \frac{1}{\sqrt{N}}\widehat{B}(x)$ and $\widehat{H}_N(x) = \Phi(x) + \frac{1}{\sqrt{N}}\widehat{B}_S(x)$. In that paper we assume that (C.1) holds with $a = 1$, and apply plug-in estimates for the moments and step-kernel estimates for the values of the density at the points ξ_α, ξ_β , appearing in the Edgeworth correction terms for skewness and bias.

In practical applications, when the density is perhaps close to zero at the points ξ_α, ξ_β , estimating these small values may cause problems, if the sample size N is not very large. In this situation we might prefer the bootstrap procedure, and to obtain second order correctness, provided we perform simulations with $M = N + o(N)$ (cf. Theorem 3). This will of course require a lot of computation. Fortunately, however, we can considerably reduce the computation time by using the M out of N bootstrap (cf. Bickel et. al [3], Bickel & Sakov [4], Sakov [24]) when $N \wedge M \rightarrow \infty$ and $M/N \rightarrow 0$, in conjunction with extrapolation.

Define functions $B_N(x)$ and $B_{S,N}(x)$ to be the same as $B(x)$ and $B_S(x)$, with parameters $\lambda_1, \lambda_2, \sigma_W, f(\xi_\alpha)$ and $f(\xi_\beta)$ replaced by $\lambda_{1,N}, \lambda_{2,N}, \sigma_{W,N}, f(\xi_{\alpha N:N})$ and $f(\xi_{\beta N:N})$. Then, with $M \ll N$ we have the bootstrap version of (31):

$$\begin{aligned} F_{T_{N,M}}^*(x) &= \Phi(x) + \frac{1}{\sqrt{M}}B_N(x) + R_{1,N,M}, \\ F_{T_{N,M},S}^*(x) &= \Phi(x) + \frac{1}{\sqrt{M}}B_{S,N}(x) + R_{2,N,M}, \end{aligned} \quad (32)$$

where $R_{j,N,M}$, $j = 1, 2$, are of the order $o(M^{-1/2-p})$ for every $p < \min(a/2, 1/4)$, a.s. $[P]$. Note that functions $B_N(x)$ and $B_{S,N}(x)$ depend on F, N , but not on M (and here it is crucial that we consider $T_N = T_{N(2)}$ instead of $T_{N(1)}$ (cf. (1)).

Therefore (32) allow us to estimate values of these functions with an error of order $o(M^{-p})$ a.s. $[P]$. The next step is to employ extrapolation for $M = N$, i.e. we substitute these estimates in the expressions for $G_{N,N}(x)$ and $H_{N,N}(x)$ to obtain the estimates

$$\widehat{F}_{T_{N,N}}^*(x) = \Phi(x) + \frac{1}{\sqrt{N}}\widehat{B}_N(x), \quad \widehat{F}_{T_{N,N},S}^*(x) = \Phi(x) + \frac{1}{\sqrt{N}}\widehat{B}_{S,N}(x),$$

with an error of order $o(N^{-1/2}M^{-p})$, implying second order correctness. Thus, we obtain the following result (cf. Gribkova & Helmers [12]) for the M out of N bootstrap, whenever M is of smaller order than N (cf. also Bickel et. al [3], Bickel & Sakov [4] for related results).

Theorem 5. *Suppose that the condition (C.1') is satisfied, and $M/N \rightarrow 0$ as $M \rightarrow \infty$. Then*

$$\sup_{x \in R} |\hat{F}_{T_N^*}^*(x) - F_{T_N}(x)| = o\left(N^{-1/2}M^{-p}\right),$$

$$\sup_{x \in R} |\hat{F}_{T_{N,N}^*,S}^*(x) - F_{T_N,S}(x)| = o\left(N^{-1/2}M^{-p}\right),$$

for every $p < \min(a/2, 1/4)$, a.s. $[P]$, as $M \rightarrow \infty$.

In particular, for $a \geq 1/2$ and with $M = [N^q]$, $0 < q < 1$, we obtain a M out of N bootstrap approximation with error of the order $o(N^{-(1/2+q/4-\epsilon)})$, for every $\epsilon > 0$. Hence, like in Theorem 3, we obtain second order correctness of the bootstrap, but now with reduced computational time. We also note in passing that extrapolation for the trimmed mean given by $T_{N(1)}$ (cf. (1)) only yields a slow rate of order $O(N^{-1/2})$. In other words: improvement over the normal approximation is not possible here for $T_{N(1)}$.

The next step would be to see whether we can possibly improve the M out of N bootstrap by extrapolation using a linear regression model. However, this is outside the scope of this short survey paper. We refer to Bickel et. al [3], Bickel & Sakov [4], Gribkova & Helmers [12] and to an unpublished Berkeley PhD thesis of Sakov [24] on the effect of extrapolation on bootstrap.

REFERENCES

1. R.R. BAHADUR, A note on quantiles in large samples. *Ann. Math. Statist.*, **37**, 577–580, 1966.
2. P.J. BICKEL, Edgeworth expansion in non-parametric statistics, *Ann. Statist.*, **2**, 1–20, 1974.
3. P.J. BICKEL, F. GÖTZE AND W.R. VAN ZWET, Resampling fewer than n observations: gains, losses and remedies for losses, *Statistica Sinica*, **7**, 1–31, 1997.
4. P.J. BICKEL AND A. SAKOV, Extrapolation and the bootstrap, *Sankhya*, **64**, Ser. A, Pt. 3, 640–652, 2002.
5. S. BJERVE, *Error bounds and asymptotic expansions for linear combinations of order statistics*, Unpublished Ph.D. thesis, Berkeley, 1974.
6. S. BJERVE, Error bounds for linear combinations of order statistics. *Ann. Statist.*, **5**, 237–255, 1977.
7. H. CHERNOFF, J.I. GASTWIRTH AND M.V. JOHNS, Asymptotic distribution of linear combinations of functions of order statistics with application to estimation, *Ann. Math. Statist.*, **38**, 52–72, 1967.
8. S. CSÖRGÖ, E. HAEUSLER AND D.M. MASON, The asymptotic distribution of trimmed sums. *Ann. Statist.*, **16**, 672–699, 1988.

9. P. DEHEUVELS, D.M. MASON AND G.R. SHORACK, Some results on the influence of extremes on the bootstrap, *Ann. Inst. Henri Poincaré*, **29**, 83–103, 1993.
10. B. EFRON, Bootstrap methods: another look at jackknife, *Ann. Statist.*, **7**, 1–26, 1979.
11. N.V. GRIBKOVA AND R. HELMERS, The empirical Edgeworth expansion for a Studentized trimmed mean, *Mathematical Methods of Statistics*, **15**, 61–87, 2006.
12. N.V. GRIBKOVA AND R. HELMERS, On the Edgeworth expansion and the M out of N bootstrap accuracy for a Studentized trimmed mean, *Mathematical Methods of Statistics*, **16**, in press, 2007.
13. N.V. GRIBKOVA AND R. HELMERS, On the M fewer than N bootstrap approximation to the trimmed mean, *submitted to Theory Prob. Appl.*, 2007.
14. P.S. GRIFFIN AND W.E. PRUITT, Asymptotic normality and subsequential limits of trimmed sums. *Ann. Statist.*, **16**, 1196–1219, 1988.
15. P. HALL, *The Bootstrap and Edgeworth Expansion*, Springer-Verlag, New York, 1992.
16. P. HALL AND A.R. PADMANABHAN, On the bootstrap and the trimmed mean, *J. of Multivariate Analysis*, **41**, 132–153, 1992.
17. F.R. HAMPEL, E.M. RONCHETTI, P.J. ROUSSEEuw AND W.A. STAHEL, *Robust Statistics. The Approach Based on Influence Functions*, John Wiley & Sons, New York, 1986.
18. R. HELMERS, Edgeworth expansions for trimmed linear combinations of order statistics, *In: Proc. 2nd Prague Symp. in Asymptotic Statist.*, (P. Mandl and M. Hušková Eds), pp. 221–232, North-Holland, Amsterdam, 1979.
19. R. HELMERS, On the Edgeworth expansion and the bootstrap approximation for a Studentized U -statistic, *Ann. Statist.*, **19**, 470–484, 1991.
20. R. HELMERS, B.-Y. JING, G. QIN AND W. ZHOU, Saddlepoint approximations to the trimmed mean, *Bernoulli*, **10**, 465–501, 2004.
21. R. HELMERS AND H. PUTTER, *Bootstrap Resampling: a Survey of Recent Research in the Netherlands*, CWI report BS-R9517, Amsterdam; also in Proceedings of the SEAMS Regional Conference on Mathematical Analysis and Statistics (1995), 20–36, Yogyakarta, Indonesia, July 10–13, 1995.
22. P.J. HUBER, *Robust Statistics*, John Wiley & Sons, New York, 1981.
23. H. PUTTER AND W.R. VAN ZWET, Empirical Edgeworth expansions for symmetric statistics, *Ann. Statist.*, **26**, 1540–1569, 1998.
24. A. SAKOV, *Using the m out of n Bootstrap in Hypothesis Testing*, Ph.D. thesis, University of California, Berkeley, 1998.
25. G.R. SHORACK, Asymptotic normality of linear functions of order statistics, *Ann. Math. Statist.*, **40**, 2041–2050, 1969.
26. S.M. STIGLER, The asymptotic distribution of the trimmed mean, *Ann. Statist.*, **1**, 472–477, 1973.
27. J.W. TUKEY AND D.H. MCCLAUGHLIN, Less vulnerable confidence and significance procedures for location based on a single sample (Trimming/Winsorization 1), *Sankhyā*, Ser. A. **25**, 331–352, 1963.
28. T. DE WET, Berry – Esseen results for the trimmed mean, *South African J. Statist.*, **10**, 77–96, 1976.

29. W.R. VAN ZWET, A Berry – Esseen bound for symmetric statistics, *Z. Wahrsch. Verw. Gebiete*, **66**, 425–440, 1984.
30. W.R. VAN ZWET, Van de Hulst on robust statistics: a historical note, *Statistica Neerlandica*, **39**, 81–95, 1985.

GRIBKOVA N.V.: St.-Petersburg State Transport University, St.-Petersburg, Russia.

E-mail: nvg@rol.ru

HELMERS R.: CWI, Amsterdam, the Netherlands.

E-mail: R.Helmerts@cwi.nl

A MoDeST Evaluation of the Interplay between Energy Consumption and Clock Drift in ZigBee

CHRISTIAN GROSS, HOLGER HERMANN AND REZA PULUNGAN

Abstract. Wireless embedded sensor networks are predicted to provide attractive application possibilities in industry as well as at home. IEEE 802.15.4 and ZigBee are proposed as standards for such networks with a particular focus on pairing reliability with energy efficiency, while sacrificing high data rates.

IEEE 802.15.4 is configurable in many aspects, including the synchronicity of the communication, and the periodicity in which battery-powered sensors need to wake up to communicate. This paper develops a formal behavioral model for the energy implications of these options. The model is modularly specified using the language MODEST, which has an operational semantics mapping on stochastic timed automata. The latter are simulated using a variant of discrete-event simulation implemented in the tool MÖBIUS. We obtain estimated energy consumptions of a number of possible communication scenarios in accordance with the standards, and derive conclusions about the energy-optimal configuration of such networks. As a specific fine point, we investigate the effects of drifting clocks on the energy behavior of various application scenarios.

Key words and Phrases: Sensor networks, formal modelling, distributed coordination, power-aware design, clock drift.

INTRODUCTION

Quantitative analysis of ad hoc and wireless networks has in the past been concentrating on scalability and routing questions [6]. The predominantly applied techniques are based on simulation using enhanced tools such as GloMoSim or NS-2 [24, 5], Omnet [22] or commercially available simulation tools, such as Opnet.

2000 Mathematics Subject Classification:

Another approach is based on instruction-level simulation of the actual microcontroller codes [23, 21].

The credibility of simulation results obtained using the above enhanced modelling tools seems not to be free of doubts. Such studies generally suffer from (1) unclarities of how simulation models are obtained from the modelling language, (2) the sheer number of parameters with non-obvious effects adjustable by the user, some of them having hidden effects on the simulation outcomes, (3) excessive simulation times needed to simulate the ensemble of many protocol stacks and states, and (4) the impossibility to validate simulation results through reproducible real-life experiments.

As a matter of fact, some recent articles have criticized the extremely poor quality and reproducibility of simulation-based experimentations with ad hoc or wireless networks [13, 6]. Indeed, this questions the validity of simulation-based predictions for this area as a whole.

The work presented in this paper is among the few which attempt to attack the above mentioned principal problems. Other loosely similar approaches include [17, 12]. While we still rely on discrete-event simulation as our analytic workhorse, we proceed in a drastically different way. The main difference is that (1) we use a language with a strictly formal semantics, which is equipped with well-established abstraction techniques. Consequently, the underlying stochastic model for simulation is well-defined and the obtained simulation results are trustworthy. (2) We expose all assumptions explicitly, since they are part of the formal system specification. (3) We do not model entire protocol stacks, but work with well-justified abstractions of lower layer effects. We consider worst-case scenarios, if no other information is available. But (4) we do not yet provide real-life experiments to back up our simulation results.

The model is modularly specified using the language MODEST, which has an operational semantics mapping on stochastic timed automata. The latter are simulated using a variant of discrete-event simulation implemented in the tool MÖBIUS. In this paper, that approach is applied to the IEEE 802.15.4 and ZigBee standards. This is a protocol family dedicated to low-bandwidth sensor networks operating on battery. IEEE 802.15.4 is configurable in many aspects, including the synchronicity of the communication, and the periodicity in which battery-powered sensors need to wake up to communicate. The particular configuration chosen has obvious – and non-obvious – implications on the lifetime of battery-powered devices. An obvious rule of thumb is, for instance, that battery-operated devices can survive longer timespans if they need to wake up less often. This paper investigates the non-obvious rules.

We obtain estimated energy consumptions of a number of possible communication scenarios in accordance with the standards, and derive conclusions about the energy-optimal configuration of such networks. In particular, we investigate the effects of time-slotted and unslotted medium access techniques, and their interplay with drifting clocks. Our observations allow us to establish rules of the following kind: (1) Unslotted CSMA/CA is more favorable w.r.t. energy saving

than slotted CSMA/CA. (2) If devices using GTSSs and CSMA/CA coexist, those operating in GTSSs expend considerably less energy. (3) Small clock drifts can have far over-proportional effects on energy consumption, but with only minor adverse effects on battery lifetimes. For sure, power consumption does not only depend on clock drift and synchronization policy. It also depends on many other factors, such as link quality and other environmental conditions. For our studies, these are assumed constant, since we see no way to include them in our studies without loosing focus. Further, we do not include comparison with experimental measurements. The reason is that (1) controlling physical clock drifts is virtually impossible with available hardware, and (2) real-life experiments would take several months or years to show measurably distinct effects on battery lifetimes. The main contribution of this paper is that it pioneers a model-based analysis of the interplay of clock drift and energy cost in sensor networks. The paper is organized as follows: Section provides a brief introduction to ZigBee and IEEE 802.15.4. In Section , the modelling formalism we employ is described together with the tool chain supporting it. Section describes the modelling of ZigBee and IEEE 802.15.4 in MODEST. We also clarify the modelling assumptions we make and discuss in detail a particular model representative. In Section , we describe and discuss the simulation experiments and their results. Section concludes the paper.

PROBLEM DOMAIN

This section provides a general introduction to ZigBee and IEEE 802.15.4, focussed on the characteristics and features that are important for the scope of the paper. For more detailed information, the interested reader is invited to consult the standard documents [2, 1].

ZigBee

ZigBee is a wireless communications standard for low-cost and low-power consumption networks, based on the IEEE 802.15.4 standard (see Section). The standard was developed by ZigBee Alliance, which is a cooperation of a number of industrial enterprises established in 2002. The ZigBee Specification Version 1.0 [2] was ratified in December 2004, and is publicly available since June 2005.

ZigBee-compliant products will typically be embedded in consumer electronics, home and building automation, industrial controls, PC peripherals, medical sensor applications, toys and games [2]. The strict emphasis on low power consumption allows implementation of wireless communications with devices operating for several years without needing battery replacement.

The architecture of ZigBee follows the layering principle. Figure 1 depicts its simplified stack architecture [2]. Both physical and medium access control (MAC) layers are defined in IEEE 802.15.4 standard. The ZigBee standard, on the other hand, provides the definition of the Network layer and the framework for the Application layer.

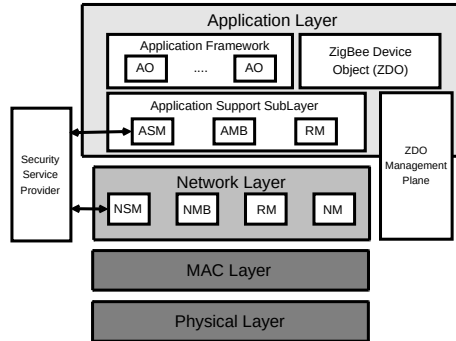


Figure 1: Simplified ZigBee Stack Architecture

Beside data delivery between layers (Network Message Broker–NMB), the Network layer provides mechanisms for joining and leaving networks, together with starting a new network and address assignment in ZigBee coordinator (Network Management–NM); applying security to data frames (Network Security Management–NSM); and routing-related activities, including neighbors discovery, creating and maintaining routing information between devices and routing the data frames (Routing Management–RM).

In the Application layer, Application Support sublayer is responsible for maintaining tables for binding, namely matches between two devices based on their services and needs, and forwarding messages between these devices. The ZigBee Device Object defines the role of the devices within the network, binding activities, establishing secure relationships between devices, discovering devices on the network and determining which services they provide.

ZigBee supports star, tree and mesh network topologies. In these topologies, a device called ZigBee coordinator controls the network. In the star topology, all devices communicate with the coordinator, while in tree and mesh topologies, communication can be mediated by ZigBee routers. Peer-to-peer communication is allowed in mesh topology.

IEEE 802.15.4

IEEE 802.15.4 is an open standard for ultra-low complexity, cost, power consumption and low data rate wireless connectivity among inexpensive devices in wireless personal area networks (WPAN) [1]. The standard specifies the physical and the medium access control (MAC) layers.

Devices participating in WPAN can be distinguished into full-function devices (FFD) and reduced-function devices (RFD). Communication between FFDs and RFDs is possible, but an RFD can only communicate with an FFD. An FFD may become a personal area network (PAN) coordinator, a coordinator, or a device. There are two topologies where a WPAN can operate: star and peer-to-peer. In

the star topology a PAN coordinator controls all communications between devices. In the peer-to-peer topology, any device can communicate with any other device as long as they are in the range of each others.

The physical layer specifies mechanisms for (de)activating the radio transceiver, detecting the channel's energy level, indicating the link quality of received packets, performing Clear Channel Assessment (CCA) for CSMA/CA, selecting channel frequency and data transmission and reception. The standard maintains that compliant devices shall operate on 2.4 GHz, 915 MHz and 868 MHz bands. The 2.4 GHz band, which is available worldwide, consists of 16 channels, each with data rate 250 kbps. The 915 MHz band is available in the North America and has 10 channels, each with data rate 40 kbps. The 868 MHz band has only one channel with data rate 20 kbps and is available in Europe. In all three bands, the radio transmissions use the direct sequence spread spectrum coding with chip rate 2000, 600 and 300 kcps, respectively. The 2.4 GHz band uses the Orthogonal Quadrature Phase-Shift Key modulation, while the others use the Binary Phase-Shift Key. As a result, the symbol rates of the three bands are 62.5, 40 and 20 ksymbols/s, respectively.

Beside managing access to the physical radio channel and providing a reliable link between peer MAC entities, the MAC layer is responsible for generating network beacons (for coordinators), synchronizing to the beacons, supporting PAN (dis)association, supporting device security, and employing the channel access mechanism.

The functional characteristic of low-rate PAN can be distinguished into beacon-enabled and nonbeacon-enabled networks. The simplest manner of operation is nonbeacon-enabled, where the network operates by using the (unslotted) Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA). CSMA/CA is a multiple access protocol similar to CSMA/CD. As opposed to CSMA/CD, which operates by listening to the channel while sending in order to detect collisions, CSMA/CA tries to avoid collisions by listening to the channel for a predetermined amount of time prior to transmissions. If a collision anyhow occurs, it can be determined by the absence of acknowledgements from the receiver. CSMA/CA is more suitable for wireless communication since listening to the medium is energy-expensive, and the spatial setup might not allow to witness collisions anyhow.

In beacon-enabled mode, on the other hand, the coordinator periodically emits beacon signals, which provide a frame of reference for a time-slotted access to the medium. More precisely, a so-called superframe structure, which is defined by the coordinator, is used. Inside of the superframe structure, communication can be carried out with a guaranteed time slot (GTS) mechanism or with slotted CSMA/CA mechanisms. Devices using GTS and those using CSMA/CA may coexist, as we will explain below.

The structure of a superframe is depicted in Figure 2. The coordinator broadcasts network beacons regularly to all devices. These beacons mark the beginnings and the ends of superframes. The beacons can be used for synchronization purpose, to identify the PAN, to describe the structure of the superframe, as well as

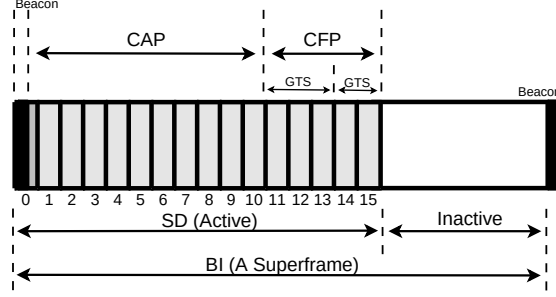


Figure 2: The Superframe Structure

to announce the GTS allocations.

The coordinator can divide a superframe into an active and an inactive portion. The active portion is further divided into 16 equally sized slots. The coordinator may decide to allocate up to 7 of these slots as GTSs. These GTSs form the Contention-Free Period (CFP) and must appear at the end of the active portions. The rest of the slots forms the Contention-Access Period (CAP), in which devices compete using a slotted CSMA/CA mechanism.

The length of a superframe (Beacon Interval-BI) is determined by the coordinator by varying the value of Beacon Order (BO) which influences the length exponentially as follows

$$BI = \text{aBaseSuperframeDuration} \times 2^{BO}, 0 \leq BO \leq 14,$$

where $\text{aBaseSuperframeDuration} = 960$ symbols. The duration of active portion (Superframe Duration-SD) is set by varying the value of Superframe Order (SO) and follows

$$SD = \text{aBaseSuperframeDuration} \times 2^{SO}, 0 \leq SO \leq BO \leq 14.$$

A device wishing to transmit using unslotted CSMA/CA mechanism, begins by setting $NB = 0$ and $BE = \text{macMinBE}$, where NB is the number of performed backoff so far, and BE is the backoff exponent. According to the standard, the value of macMinBE is in the range $\{0 \dots 3\}$ with default 3. The device then randomly selects a number r from the interval $[0, 2^{BE} - 1]$ and waits for r backoff periods before performing a clear channel assessment (CCA). If the channel is idle, the device transmits immediately.

If the channel is not idle then another backoff period must expire before a new CCA is performed. For this, the value of NB is increased by one, and similarly with BE as long as it does not exceed $\text{aMaxBE} = 5$. Thus, the upper bound of the interval, from which the random number is selected, grows exponentially every time the CCA finds the channel busy. However, this does not go on indefinitely. The number of backoff (NB) is limited by macMaxCSMABackoff , a value in the range

$\{0 \dots 5\}$ with default value 4. Once NB reaches this limit, the channel access attempt is considered failed, and the device has to start the transmission anew.

In a beacon-enabled network, slotted CSMA/CA can be employed during the CAP. The backoff period boundaries of each device in the PAN are then aligned with the superframe slot boundaries, and a device can only start transmitting on the boundary of a backoff period. A backoff period, in both slotted and unslotted mode, is the same as the required time to transmit `aUnitBackoffPeriod` = 8 symbols.

Beside the two variables used in the unslotted version, in the slotted version a transmitting device begins by setting $CW = 2$, where the CW is the contention window length. It then tries to locate the backoff period boundary, and proceeds to randomly select some number r from the interval $[0, 2^{BE} - 1]$, and waits for r backoff periods, similar to the unslotted mechanism. The differences are that on finding the channel idle on the first CCA, a further CCA is commenced before the actual transmission is started, and that the CCAs are performed on the boundaries of backoff periods. Every time the CCA confirms that the channel is idle, CW is decreased by one. During these CCAs, a busy channel will trigger backoff.

MODEST AND SUPPORTING TOOLS

The *Modelling and Description Language for Stochastic and Timed Systems* (MODEST) [8] is a specification formalism for stochastic real-time systems. The language is rooted in classical process algebra, *i.e.* the specification of models is compositional. Basic activities are expressed with atomic actions, more complex behavior with constructs for sequential composition, nondeterministic choice, parallel composition with CSP-style synchronization, looping and exception handling. A special construct exists to describe probabilistic choice. Clocks, variables and random variables are used to describe stochastic real-time aspects.¹

All constructs and language concepts have a pleasant syntax, inspired by Promela, LOTOS, FSP, and Java. We will see examples later. MODEST is equipped with a structural operational semantics mapping on stochastic timed automata (STA). The MODEST semantics is described in full detail in [8]. MODEST allows one to describe a very large spectrum of models, including: ordinary labelled transition systems, timed automata, probabilistic automata, stochastic automata [15], Markov decision processes, and various combinations thereof (*cf.* [8]). Remarkably, the language is designed in a way that all these models correspond to syntactic subsets of the language, and can thus be identified while parsing a MODEST specification.

MODEST takes a *single-formalism, multi-solution* approach. This advocates to have a single specification that addresses various aspects of the system under consideration. This is contrary to the more common approach to construct different models to describe different aspects of a system and then analyze these models.

¹MODEST also supports for modelling time variation, especially time non-determinism, but this feature is not used in this case study.

Generally, no guarantee of consistency between these models can be given, be it for lack of a rigorous semantics or because a proper relation between the different model classes is not known. Thus, the validity of results w.r.t. the original system under study is often questionable.

In order to facilitate the analysis of the different models, tool support is indispensable. The MODEST Tool Environment (MOTOR) is a software tool that implements the MODEST semantics and is the central vehicle in the multi-solution analysis of MODEST models. The fundamental idea behind MOTOR is to simplify specifying MODEST models (*e.g.*, by providing a macro-preprocessor), and to translate or adapt the models in a way such that the actual analysis work can be carried out by third-party state-of-the-art tools, such as PRISM [20], UPPAAL [7], or CADP [18].

MOTOR is designed to facilitate easy access to all language features of MODEST, and thus to allow easy extraction of all imaginable model classes from a specification. The design allows straightforward extensibility of the tool. To realize this, MOTOR provides two abstract programming interfaces, one on the level of the abstract syntactic representation of the MODEST specification, and another one which provides backend-developers with access means to the STA defined by the MODEST semantics.

The by now most mature backend of MOTOR is provided by a link to the MÖBIUS performability evaluation environment. MÖBIUS has been developed independently from MODEST and MOTOR at the University of Illinois at Urbana-Champaign [14]. MÖBIUS is designed as an integrated tool environment for the analysis of performability and dependability models. It allows specification of models in different formalisms, based on, for instance, Petri net-like formalisms or Markovian process algebra. The tool provides efficient discrete-event simulation capabilities and numerical solvers, such as Markov chain solvers. From a user perspective, the MOTOR/MÖBIUS tandem enables one to perform simulation of MODEST models, and to gather performability results.

Simulation-based analysis covers the largest language fragment of MODEST: the only concept that cannot be supported by simulation is nondeterminism, in particular of delay durations and non-deterministic choice between actions. We exclude the former by assuming maximal-progress with respect to delays. We do not restrict action nondeterminism, since it is a convenient modelling instrument. However, no mechanisms, like a well-specified-check [16], is implemented yet to ensure the validity of the simulation statistics when action nondeterminism is present.

MOTOR and its connection to MÖBIUS is mature and has been tested in a number of non-trivial case studies. In [19], it has been used for reliability analysis of the upcoming European Train Control System (ETCS) standard. In [9], it has been applied to the analysis of an innovative plug-and-play communication protocol, which has led to a patent application of our industrial partner. In [10], MOTOR has been used for the optimization of production schedules, in combination with timed automata-based schedule synthesis with UPPAAL.

MODELLING

In this section, we describe the modelling of ZigBee and IEEE 802.15.4 in MODEST. First, we provide an overview of the models. At this point, we clarify the parts of the standard being modelled, as well as the parts we abstract from. Subsequently, we describe the framework model, taking the slotted CSMA/CA version as representative. In the end of the section, the modelling of energy consumption and clock precision is explained.

Overview

We model ZigBee/IEEE 802.15.4-compliant personal area networks in a star topology. Each network consists of a single PAN coordinator and a number of stations or devices. We assume that the PAN coordinator has continuous power supply, while the stations do not. A station can be either an FFD or an RFD, attached to a sensor. Periodically, a station communicates with the PAN coordinator, either to transmit its gathered sensor data or to receive instructions, but a station cannot communicate with other stations.

Two separate models, beacon and nonbeacon-enabled PAN models, are developed. The simpler, nonbeacon-enabled model is parameterized by BI, but this is not used to signal beacons. Instead, BI is used to indicate the arrival of messages to each station from its sensor. It is assumed that each station, in both models, always has message to send: almost one slot-length of data every beacon interval.

In the beacon-enabled PAN model, some of the stations use the CAP for CSMA/CA communications, and some use GTS in the CFP, the detailed scenarios considered are described later. The model is parameterized by BI, the beacon interval. We set the superframes to have the same duration as BI, hence there is no inactive period. Every BI-equivalent time units, the PAN coordinator broadcasts a beacon, and all stations must be ready to receive it. Henceforth, the coordinator is ready to receive transmissions from the stations. The stations which are not assigned any GTSs compete with other similar stations to send their messages by using slotted CSMA/CA, while the stations with assigned GTSs wait for their turns.

To save energy, a station goes to sleep mode whenever it has a chance to. For instance, a station sleeps whenever it has no messages to transmit, or while it must wait for its GTS turn or when it is performing backoff. However, a station must always wake up before its turn to transmit or before the beacon is transmitted by the PAN coordinator.

We abstain from modelling the initialization of the PANs. Hence, the selection of the name, the (dis)association of stations into the PAN, and the address assignment of stations by the coordinator are not modelled. Instead, we concentrate on the typical operations of the PAN, when the stations are transmitting messages. The physical layer of IEEE 802.15.4 is not modelled either. For instance, we assume that there is no significant propagation delay and no channel

selection procedure. Nevertheless, some physical layer constants that affect the timing of the communication, for instance the duration of CCAs, are taken into consideration. Furthermore, we model PAN in 2.4 GHz band, which means that all devices transmit at 62.5 ksymbols/s. The medium is assumed uniform, in the sense that all communicating parties have complete knowledge of it.

Models

The model of the PAN in beacon-enabled mode is described in this section, focusing on slotted CSMA/CA, which is more complex and interesting than the GTS case. Complete details are available as a tutorial-style modelling guide [4]. The model consists of two distinct process definitions: `coordinator()` and `station()`, modelling the behaviors of a PAN coordinator and a station, respectively. In all experiments, we set `macMaxCSMABackoff` (the maximum backoff attempts before declaring a channel access failure) to 5 and `macMinBE` (the minimum value of the backoff exponent) to 2.

The System The model of the overall system is depicted in MODEST model 1. The system consists of 11 process instances, one coordinator and ten stations, run in parallel. This is achieved by using the parallel composition construct `par{}`. Pro-

MODEST **model 1** The complete system

```

01 par{
02   :: coordinator()
03   :: relabel {...} by {...} station(1)
04   ...
12   :: relabel {...} by {...} station(10)
13 }
```

cesses inside of a parallel composition construct run concurrently and synchronize on their common actions, if existing. The relabelling operator `relabel {} by {} p()` relabels the actions in the first set by the actions in the second in a particular instance of process `p()`. This allows multiple instantiations of process definition `station()`.

The Coordinator A simple model of the PAN coordinator is shown in MODEST model 2, especially to highlight some of the clock manipulation features of MODEST. The coordinator has two clock variables: `btimer` modelling the time progress between beacons, and clock `c` modelling the transmission time of a beacon. Clock variables increase linearly with time and can only be reset to zero. The coordinator process begins by immediately sending a beacon (action `sendb_start`). All beacons are of length 52 μ s (lines 05 & 09), namely the duration of the smallest possible beacon. Action `sendb_end` signals the end of the beacon's transmission. From then on, the coordinator waits until clock `btimer` is equal to the value of `binterval` (line

```

MODEST model 2 The coordinator
01 process coordinator() {
02   clock btimer, c;
03
04   sendb_start {= bintheair=true =} ;
05   when(c==52) sendb_end {= bintheair=false =} ;
06   do{
07     ::when(btimer==binterval)
08       sendb_start {= c=0, btimer=0, bintheair=true =} ;
09       when(c==52) sendb_end {= bintheair=false =}
10   }
11 }

```

07). At this point, a beacon interval has expired and the coordinator broadcasts a new beacon. The coordinator proceeds thus continuously, broadcasting beacons every time a beacon interval expires.

The process makes use of two global variables, *i.e.* variables accessible to all processes in the model. The variables are **binterval**, of type integer, which corresponds to the time it takes for a BI in the standard, namely the length of the time interval between two beacons; and **bintheair**, which is a boolean variable used to indicate to the whole system that a beacon is being transmitted. Delimiters {= =} wrap a set of variable assignments. Such sets of assignments are executed atomically at the time instant of the action preceding them.

The Stations The model of the station is shown in MODEST model 3 and 4. In the beginning, a station waits until **bintheair** is true, indicating the beginning of the transmission of a new beacon. At the same time, it resets its main clock and sets **ttosend** (the duration of the remaining data to send) equal to constant **sendingtime** = **binterval**/16 – 194 μ s, almost a slot duration in the used BI. Once the beacon finishes, the station aligns its backoff boundary with the superframe slot boundary by waiting until a multiple of the backoff period (20 symbols, hence 320 μ s in 2.4 GHz band) expires since the beginning of the previous beacon (line 20). The station then performs backoff and attempts to transmit. This is repetitively done as long as there is still something to send and enough time to do so (line 17), otherwise (line 15) the station just waits for the next beacon.

The maximum length of messages submittable to the physical layer is 133 bytes, which takes 4256 μ s to transmit in 2.4 GHz band. Thus the remaining data is split accordingly (line 19) into **attosend**, the actual time required to send the partitioned message. In line 23, more lines of code are needed to model the random selection of variable **r** than the code indicates. For readability we shorten the code and put **r=//Uniform(0,2^{BE}-1),//** as a placeholder.

As shown in MODEST model 4, once the duration of the backoff delay is determined, the backoff is performed only if there is still enough time to complete

```

MODEST model 3 The station (Part 1)
01 process station(int id) {
.. ...
10 do{
11  ::when(bintheair) beacon_received
12    {= mainclock=0, ttosend=sendingtime =} ;
13    when(!bintheair) start_waiting;
14    do{
15      ::when(ttosend==0 || !enoughtime) do_nothing
16        {= enoughtime=true =} ; break
17      ::when(ttosend>0 && enoughtime) start_csmaca
18        {= NB=0, CW=2, BE=2,
19          attosend=(ttosend>=4256)?4256:ttosend =} ;
20      when(mainclock%320==0)
21        do{
22          ::choose_random
23          {= c=0, r=//Uniform(0,2^BE-1),//
24            backofftime=r*320 =} ;
.. ...

```

the backoff together with a CCA before the contention-access period ends (line 28). CCA detection time, denoted by `ccatime` in the code, is equal to 8 symbols period, namely $128 \mu\text{s}$ in 2.4 GHz band. If there is not enough time, the station stops and waits until the next beacon comes (line 26). A CCA is carried out immediately after the backoff finishes (line 29 & 31). There are three possible outcomes of the CCA: a busy channel, an idle channel with enough time to send the current portion of the message and an idle channel but not enough time. The outcome of a CCA is determined by the value of global variable `sending`. Variable `sending` indicates the number of stations transmitting at the current moment. A station about to transmit increases this variable and decreases it again once the transmission finishes.

A busy channel triggers another backoff if the number of backoff so far does not exceed the maximum allowed – `maxbackoff` – (line 33 & 35), otherwise the whole CSMA/CA procedure must be restarted to transmit the current portion of the message (line 37). When the channel is idle but there is not enough time to complete the transmission (line 51) the station escapes the CSMA/CA procedure and waits for the next beacon for further attempts.

In a beacon-enabled PAN, two CCAs are required after backoffs before the transmission of the message. Hence, when the channel is idle and there is enough time to complete the transmission (line 39), the transmission is only commenced if $CW = 0$ (line 42), namely when two consecutive CCAs find the channel idle. The transmission, which starts at the next backoff boundary, is announced to all other devices by increasing the global variable `sending`, and it takes the amount of time to send the portion of the message, namely `attosend`.

Once the portion of the message is transmitted, the station decreases the

```

MODEST model 4 The station (Part 2)
.. ...
25 alt{
26 ::when(mainclock>=CAP-backofftime-ccatime)
27   {= enoughtime=false =} ; break
28 ::when(mainclock<CAP-backofftime-ccatime)
29   when(c==backofftime) {= c=0 =} ;
30   do{
31     ::when(c==ccatime)
32       alt{
33         ::when(sending>0)
34           alt{
35             ::when(NB<=maxbackoff) channel_busy
36               {= CW=2, NB=NB+1, BE=(BE<6)?BE+1:6 =} ; break
37             ::when(NB>maxbackoff) {= restart=true =} ; break
38           }
39         ::when(sending==0 && mainclock<CAP-attosend-960)
40           count_down_CW {= CW=CW-1 =} ;
41         alt{
42           ::when(CW==0) wait_for_boundary ;
43           when (c==320) send_message_start
44             {= sending+=1, c=0 =} ;
45           when (c==attosend) send_message_end
46             {= ttosend-=attosend, sending-=1,
47               restart=true =} ; break
48           ::when(CW>0) wait_for_boundary ;
49           when(c==320) {= c=0 =}
50         }
51         ::when(sending==0 && mainclock>=CAP-attosend-960)
52           {= enoughtime=false, restart=true =} ; break
53       }
54   } ; alt{
55     ::when(restart) {= restart=false =} ; break
56     ::when(!restart)
57       } } } } } }

```

global variable `send` and updates the remaining portions of the message to transmit by changing `ttosend` (line 46). Henceforth, the station restarts the CSMA/CA procedure to transmit the remaining message, without waiting for a new beacon.

Energy Consumption

In modelling the energy consumption of a station, we use the technical specification of CC2420 [3]. CC2420 is a 2.4 GHz ZigBee/IEEE 802.15.4-compliant radio frequency (RF) transceiver produced by Chipcon AS. Figure 3 summarizes the specification relevant for this paper. During its operation, the transceiver can be in four modes: shutdown, idle, transmitting or receiving modes. The rate of energy

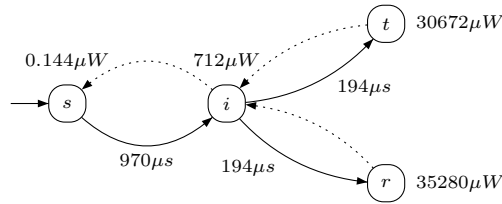


Figure 3: Energy Modes of CC2420

consumption while occupying these modes [11] are 0.144, 712, 30672 and 35280 μW , respectively. A small amount of energy is required during the shutdown mode to power clock and to witness power-ups.

Furthermore, the transitions from mode *s* to *i*, from mode *i* to *t*, and from mode *i* to *r* do not happen instantaneously, but take 970, 194 and 194 μs to complete, respectively. During these transitions [11], the transceiver is considered to be still in the original mode, while consuming energy at the level of the destination mode. This reflects that the transceiver requires time (and power) to turn on its transmitting and receiving devices. All other transitions, namely those with dotted edges in the figure, take no time to complete. In the models, all stations are assumed to use transceivers with such energy consumption.

When the PAN coordinator broadcasts a beacon, all stations must be in receiving mode. That means, some time beforehand, they must have woken up and proceeded to the idle and then to the receiving mode. If a station is assigned GTSSs, it immediately changes to shutdown mode upon the completion of the beacon. However, the station must already be in transmitting mode, when its assigned GTSSs begin. To anticipate this, the station must leave the shutdown mode 1164 (namely 970+194) μs before the GTSSs. After finishing the transmission in the GTSSs, the station moves to idle mode. If there is still ‘enough’ time before the next beacon, it changes further to shutdown mode.

A station without assigned GTSSs must compete with similar stations by using slotted or unslotted CSMA/CA to gain access to the medium. Immediately after the end of a beacon, the station enters the CSMA/CA procedure. Depending on

the duration of the backoff delay, the station may transition to idle or shutdown mode. If the backoff delay is long enough, the station may sleep and wake up $1164 \mu s$ before it must perform a CCA, and be ready in receiving mode. If two consecutive CCAs result in idle channel, the station changes to transmitting mode and sends the message portion. This is performed continuously until the whole message is sent. Afterwards, the station may move to shutdown mode if there is enough time to do so before the next beacon.

To incorporate the energy consumption to the PAN models, each instant of station changes variables `time_in_s_mode`, `time_in_i_mode`, `time_in_t_mode` and `time_in_r_mode` every time it spends some portions of time in shutdown, idle, transmitting, and respectively receiving mode. The framework models are annotated accordingly and the four variables are turned into four global variable arrays, which each instant of station accesses through its `id`. The amount of energy consumed by a station during a beacon interval is the sum of amount of time it spends in each mode weighted by the rate of the energy consumption of the mode.

Referring back to MODEST model 3 and 4, the stations start being in receiving mode in lines 11, 29 and 49. They are starting in transmitting mode in line 44. They start in idle mode in lines 13, 31, and 37. In lines 16, 23, 27, 36 46 and 52, they have to make decisions whether to enter idle or shutdown mode based on the progress of their main clocks so far. If there is enough time to sleep and then follow the wake up procedure, which takes time, they enter shutdown mode. Otherwise, they just enter idle mode.

Clock Precision

Clock precision is important to ensure the correct functioning of ZigBee/IEEE 802.15.4-compliant PANs. In the case of PANs which provide CFPs, it is crucial that stations with assigned GTSs transmit exactly in their allocated period of time. Similarly with PANs employing the slotted CSMA/CA mechanism, the ability to correctly determine the backoff boundaries, which requires a precise clock, is necessary to avoid collisions.

Quartz-based clocks are generally used for electronic components in industry. Such clocks suffer from inaccuracy due to aging and temperature variations. Usually, the manufacturers guarantee a certain upper bound inaccuracy for their clock products. Assuming this upper bound to be constant is actually not realistic. However, we do so in our models. Hence, a clock with a guaranteed accuracy may deviate from the real time within a given time interval based on the accuracy and exhibit at different times a different deviation from the real time.

Clock inaccuracy is usually expressed in ‘parts per million’ (ppm), namely the maximum deviation is some certain time units within a million time units. In the models, we assume that a clock with a guaranteed accuracy of p ppm may show at time t , a clock value within the interval

$$[t - p \cdot 10^{-6} \cdot t, t + p \cdot 10^{-6} \cdot t].$$

A station, however, does not know the exact value of the deviation. Nevertheless, it must be able to precisely observe the deadlines, for instance the arrival of a beacon. Therefore, it counts on the maximum inaccuracy and be ready for the beacon even before the actual time. It is also assumed that the stations synchronize their clocks to the PAN coordinator's, while receiving the beacons.

We model the effect of clock inaccuracy to energy consumption in the following way. Assume that a station must wait for W time units to be ready for some event. The clock of the station has inaccuracy p ppm. The actual waiting time W' for the station is

$$W' = W - W \cdot p \cdot 10^{-6} + W \cdot p' \cdot 10^{-6}, \quad (1)$$

where $W \cdot p \cdot 10^{-6}$ is the maximum deviation of the clock given the clock inaccuracy and $W \cdot p' \cdot 10^{-6}$ is the actual deviation. The actual inaccuracy p' is a value in the interval $[-p, p]$. The models can be parameterized by the actual deviation. In full generality, a clock's inaccuracy may be time-dependent, in which case the above formula involves integration.

SIMULATIONS

In this section, we describe the simulations of the MODEST models presented in the previous section. The simulation was done with discrete-event simulator of MÖBIUS, and we only present an excerpt of several thousand simulation runs we performed. First, the experimental setups are described. The result of the experiments is presented afterwards, followed by its analysis and discussions.

Experiments

In all experiments, the system is a personal area network, which consists of a single PAN coordinator and 10 stations with star topology. The network uses 2.4 GHz band and all durations appearing in the standard which are defined by amount of symbols are adapted accordingly. The networks do not use any inactive periods, thus the beacon interval is always the same as the superframe duration (BI=SD).

The experiments cover $BO = 0, \dots, 10$, which means a beacon interval ranges from 15360 μs to 15728640 μs . Similarly the length of a slot ranges from 960 μs to 983040 μs . In all experiments, a station always has about a slot-length message to transmit during the duration between two beacons. A station transmitting in GTSSs can transmit the whole message continuously. On the other hand, a station using slotted or unslotted CSMA/CA can only transmit 133 bytes at a time, which takes 4256 μs to transmit.

The simulation time of each experiment takes the duration of 10 beacon intervals. For instance, if $BO = 7$, a beacon interval takes 1966080 μs . Therefore, the simulation is run for around 20 seconds (19660800 μs). Simulations are carried out to estimate the mean values of some measures of interest, such as the time a station spends in shutdown mode. The simulation is repeated until the mean

values of all measures of interest converge with relative confidence interval 0.1 and confidence level 95%. MÖBIUS allows users to adjust these settings as desired.

The experiments are divided into two groups: (P) experiments for models with perfect clocks and (D) experiments for models with drifting clocks. There are 3 series of experiments in experiment group P. In series 1, 7 stations use GTSS, while 3 stations use slotted CSMA/CA. This is due to the restriction that there can only be at most 7 slots used for CFP in beacon-enabled networks. In series 2 and 3, 10 stations use slotted and unslotted CSMA/CA, respectively. In experiment group D there are 12 series of experiments, namely the three series in group P repeated for clock inaccuracies of 40, 20, 10 and 5 ppm.

As mentioned earlier, we consider worst-case scenarios whenever appropriate. For clock inaccuracy, the worst scenarios occur when the clock of the PAN coordinator is progressing as slow as possible within its inaccuracy bounds, while the clock of a station is in its fastest possible progress. Referring to Equation (1), the clock of a station is fastest when $p' = -p$. Then the station, which needs to wake up some time before the actual deadlines, actually wakes up even earlier because its clock is too fast. We proved this situation to be worst case by analytical means, which is backed up by simulation. The experiments reported in group D are worst scenario experiments.

All simulations are conducted on a PC with a Pentium 4 3.0 GHz processor with 1 GB RAM running Linux 2.6.17-2.686. The CPU time required ranges from the briefest around 5.2 seconds for experiment group P series 1 with $B0 = 0$ precise clock to the longest around 15429 seconds (around 4.25 hours) for experiment group D series 3 with $B0 = 10$ and clock inaccuracy 40 ppm.

Results

The graphs in Figure 4 summarize the result of experiment group P, where clocks are perfect. The first graph in the figure shows the percentage of the time spent in transmit, receive and idle modes for the 7 stations assigned with GTSS in experiment series 1 (we denote this as series 1G). While the portion of time spent in transmit mode is constant for all $B0$'s, the portions in receive and idle modes diminish as the beacon interval gets larger, and become less than 1% after $B0 = 5$. This is because the actual time spent in receive and idle modes remains constant for all beacon intervals, thus their percentages decrease as the beacon intervals get longer. They are constant because a station only goes to receive mode while receiving beacons and to idle mode prior to the arrivals of beacons and before the transmissions of the messages.

The second graph in the figure depicts the same percentage distribution for the 3 stations using slotted CSMA/CA in experiment series 1 (denoted as series 1C). The portion of time spent in the transmit mode remains almost constant and around 0.25% higher throughout than the previous case. The increase can be explained by the fact that there are more transitions from idle to transmit mode compared to the previous case, because of the restriction on the length of message

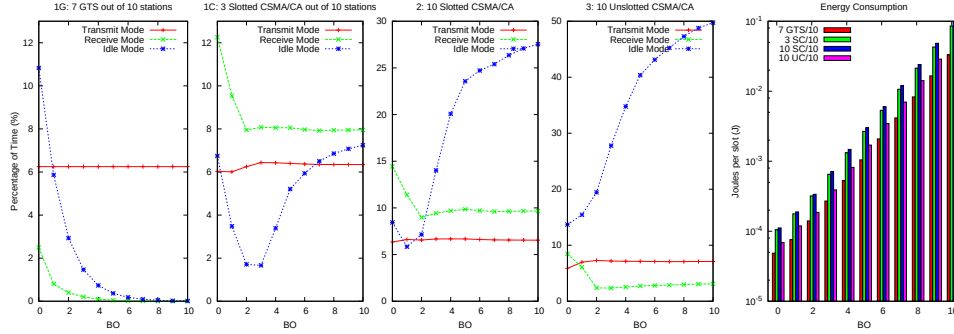


Figure 4: P: Percentage of time per energy mode and energy consumption for perfect clocks

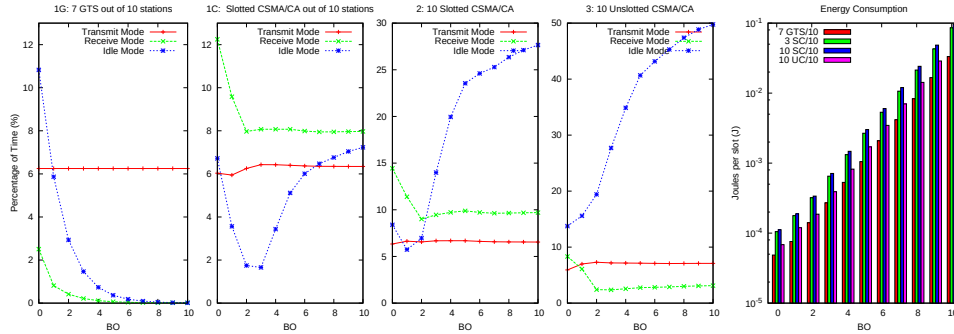


Figure 5: D: Percentage of time per energy mode and energy consumptions for drifting clocks (40 ppm)

portions. When the beacon intervals are too brief, some of the stations may find not enough time to completely transmit its apportioned message due to the overhead of the CSMA/CA procedure, hence the small perturbations in the percentage of transmit mode when $BO < 5$. The percentage of time spent in receive mode is also almost constant at around 8%, which is higher than in series 1G. The percentage increase in receive mode is due to the CCAs, and since the number of CCAs to be performed is proportional to the length of the message, the percentage is steady for most BO values. The time spent in the idle mode, on the other hand, increases with the length of beacon intervals. The longer the beacon interval, the longer a slot-length message. However, since the length of message portion transmittable is restricted, the more often slotted CSMA/CA procedure is repeated to send the whole message. Hence the station spends more percentage of time in idle mode.

The third and fourth graphs in the figure show similar percentage distribution for experiment series 2 and 3, namely 10 stations using slotted and unslotted CSMA/CA, respectively. Compared to series 1C, the percentage of time spent in

receive mode is higher in series 2. This is due to the fact that there are 10 stations for 15 CAP slots in series 2 compared to 3 stations for 8 CAP slots in series 1C. Thus stations perform more CCAs and spend more percentage of time in receive mode. Similar reason applies to the percentage of idle mode. In series 3, however, the percentage of time spent in receive mode is lower than in series 1C and 2 because the stations perform only one CCA in every CSMA/CA cycle. Between themselves, the stations in experiment series 3 spend more time in idle mode compared to those in series 2. This can be explained by the absence of beacons in unslotted CSMA/CA. Once a station finishes transmitting the whole message, it does not have any estimation when the next message will be available and thus must stay in idle mode. The station only enters the shutdown mode when it is performing backoff.

The rightmost graph in Figure 4 depicts the amount of energy consumed by a station during a beacon interval time for each value of B_0 and for each of the formerly described series. Note that the y-axis of the graph is in logarithmic scale. The largest energy consumption is around 0.096946940587 Joules per beacon interval when $B_0 = 10$. We can observe that for all values of B_0 , a station using GTSS consumes the least amount of energy. A station using the unslotted CSMA/CA consumes more energy than that using GTSS but less than the energy consumption of a station using slotted CSMA/CA. A question arises as to why the energy consumption of a station in series 1C is lower than that in series 2, considering the fact that the former has more CAP-slots per station. The reason is in series 1C, a station spends more time in shutdown mode, namely during the CFP period. This can also be observed from the second graph in the figure. It is also worth noting that the differences between the energy consumptions amongst the series are significant, even though they are not apparent due to the logarithmic scale. For instance, a station in series 2 consumes almost thrice the energy of a station in series 1G and almost twice of a station in series 3.

Figure 5 depicts the corresponding graphs for the experiment group D, in which clocks are now drifting with inaccuracy 40 ppm. Recall that we consider the worst case of the coordinator driven by a slowest possible clock, and the stations being fastest. The graphs show similar trends as the graphs in Figure 4, but with some differences. These differences are not apparent, due to the scale of the graphs. There is actually a slight increase of percentage of time spent in receiving and idle modes, decrease of percentage of time spent in shutdown mode and no changes in transmitting mode, which will be emphasized in the next figure.

Figure 6 shows the influence of worst-case clock drifts on the energy consumptions of stations in series 1G, 2 and 3, respectively. Consistently, clock drifts increase the amount of energy consumed for all experiments. The graphs in the figure show the increase in energy expended relative to the amount of energy needed in the precise clock setting. The first graph depicts the deviations for a station in series 1G with clock inaccuracies 40, 20, 10 and 5 ppm for $B_0 = 0, 2, 4, 6, 8, 10$. We observe a linear correlation between the clock inaccuracies and the deviations in the graph. For instance, for most of B_0 values, a station with clock inaccuracy

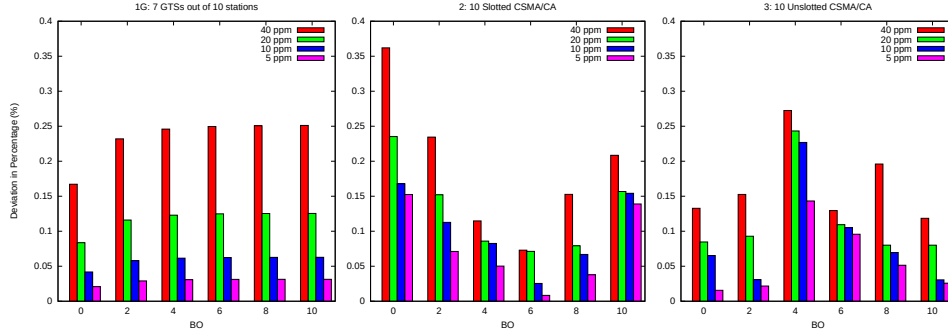


Figure 6: Increase of energy consumptions when clocks drift (40, 20, 10 and 5 ppm) relative to perfect clocks

40 and 20 ppm spends approximately 0.25% and respectively 0.125% more energy than a station with precise clock. The second and third graphs in Figure 6 depicts the deviations for a station in series 2 and 3, respectively. We abstain from explaining the precise patterns exhibited by these plots for increasing BO. However, the effects of the level of clock inaccuracies are obvious. In both graphs, higher clock inaccuracy tends to cause higher energy consumption level.

Overall, a deviation of 0.25% or 0.125% may seem small. If extrapolated, it amounts to about half a day per year decrease in lifetime of a battery. However, a clock drift of 40 ppm is – percentwise – only 0.004%, and we thus conclude that such a clock drift may be amplified by a factor of 62.5 when looking at the excess in energy consumption caused.

CONCLUSIONS

This paper has advocated a rigorous approach to modelling and simulation of energy-aware networked embedded systems. In this approach, MODEST, a language with a strictly formal semantics is used to specify the system under study. Since all assumptions are exposed explicitly in the specification, the underlying stochastic model is well-defined and the obtained simulation results are trustworthy. This is different from many other simulation-based results lately published in the networking literature. The simulation is carried out via the tools MOTOR and MÖBIUS. MOTOR is available as source code from <http://www.purl.org/net/motor> under the GPL license. MÖBIUS is freely available for educational and research purposes from <http://www.mobius.uiuc.edu/>. MOTOR can be installed as an add-on package to the MÖBIUS installation.

We applied the proposed approach to the modelling of crucial medium access mechanisms in IEEE 802.15.4, with a particular emphasis on energy specific configuration relevant to ZigBee operation. When investigating the effects of time-

slotted (*i.e.*, beacon-enabled) and unslotted medium access techniques, we observed that slotted CSMA/CA is clearly inferior to unslotted CSMA/CA w.r.t. energy efficiency. In slotted mode, devices using GTSs use considerably less energy than those using CSMA/CA.

Our studies of drifting clocks reveal some interesting observations. Most of them are consistent with what one might expect – but we did not see them published elsewhere. However interesting, they only show minor quantitative effects. The IEEE standard limits the allowed clock drift to 40 ppm, and this appears to be caused by correctness concerns. (One can show that with 64 ppm the clock boundaries could drift apart to an extent that the protocol could desynchronize and thus malfunction). In the worst case we estimate that a clock drift of 40 ppm induces a shortage in battery life of about half a day per year. While this amplifies the clock drift by a factor of about 60, we think it is not particularly significant, leading us to finally negate the question in the title of this paper.

Acknowledgement. This work is supported by the German Research Council (DFG) as part of the Transregional Collaborative Research Center 'Automatic Verification and Analysis of Complex Systems' (SFB/TR 14 AVACS). See <http://www.avacs.org> for more information.

REFERENCES

- [1] *IEEE 802.15.4: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low-Rate Wireless Personal Area Networks*. IEEE Press, New York, 2003.
- [2] *ZigBee Specification Version 1.0*. ZigBee Alliance, 2004.
http://www.zigbee.org/en/spec_download/download_request.asp.
- [3] *CC2420 Product Information*. Chipcon AS, 2005.
http://www.chipcon.com/index.cfm?kat_id=2&subkat_id=12&dok_id=115.
- [4] *MoDeST Tutorial: development of a complex model in MoDeST*. Dependable Systems and Software, Saarland University, Germany, 2006.
<http://depend.cs.uni-sb.de/modesttutorial/index.html>.
- [5] The network simulator – ns-2 website, 2007.
<http://www.isi.edu/nsnam/ns/>.
- [6] T. R. Andel and A. Yasinac. On the credibility of Manet simulations. *IEEE Computer*, 39(7):48–54, 2006.
- [7] J. Bengtsson, K. G. Larsen, F. Larsson, P. Pettersson, and W. Yi. UPPAAL – a Tool Suite for Automatic Verification of Real-Time Systems. In *Workshop on Verification and Control of Hybrid Systems III*, pages 232–243. Springer, 1995.
- [8] H. C. Bohnenkamp, P. R. D’Argenio, H. Hermanns, and J.-P. Katoen. MoDeST: A compositional modeling formalism for hard and softly timed systems. *IEEE Transactions on Software Engineering*, 32(10):812–830, 2006.
- [9] H. C. Bohnenkamp, J. Gortler, J. Guidi, and J.-P. Katoen. Are you still there? - A lightweight algorithm to monitor node presence in self-configuring networks. In *2005 International Conference on Dependable Systems and Networks, DSN’05*, pages 704–709. IEEE Computer Society, 2005.

- [10] H. C. Bohnenkamp, H. Hermanns, R. Klaren, A. Mader, and Y. S. Usenko. Synthesis and stochastic assessment of schedules for lacquer production. In *1st International Conference on Quantitative Evaluation of Systems, QEST'04*, pages 28–37. IEEE Computer Society, 2004.
- [11] B. Bougard, F. Catthoor, D. C. Daly, A. Chandrakasan, and W. Dehaene. Energy efficiency of the IEEE 802.15.4 standard in dense wireless microsensor networks: Modeling and improvement perspectives. In *8th Conference on Design, Automation and Test in Europe, DATE'05*, pages 196–201. IEEE Computer Society, 2005.
- [12] M. Cadilhac, T. Hérault, R. Lassaigne, S. Peyronnet, and S. Tixeuil. Evaluating complex mac protocols for sensor networks with apmc. *Electron. Notes Theor. Comput. Sci.*, 185:33–46, 2007.
- [13] D. Cavin, Y. Sasson, and A. Schiper. On the accuracy of MANET simulators. In *2nd ACM International Workshop on Principles of Mobile Computing*, pages 38–43. ACM Press, 2002.
- [14] D. Daly, D. D. Deavours, J. M. Doyle, P. G. Webster, and W. H. Sanders. Möbius: An extensible tool for performance and dependability modeling. In *11th International Conference on Computer Performance Evaluation: Modelling Techniques and Tools, TOOLS'00*, pages 332–336. Springer, 2000.
- [15] P. R. D'Argenio and J.-P. Katoen. A theory of stochastic systems part I: Stochastic Automata. *Information and Computation*, 203(1):1–38, 2005.
- [16] D. D. Deavours and W. H. Sanders. An efficient well-specified check. In *8th International Workshop on Petri Nets and Performance Models*, pages 124–133. IEEE Computer Society, 1999.
- [17] M. Fruth. Probabilistic model checking of contention resolution in the IEEE 802.15.4 low-rate wireless personal area network protocol. In *2nd International Symposium on Leveraging Applications of Formal Methods, Verification and Validation, ISOLA'06*, 2006.
- [18] H. Garavel, F. Lang, and R. Mateescu. An overview of CADP 2001. *European Association for Software Science and Technology Newsletter*, 4:13–24, 2001.
- [19] H. Hermanns, D. N. Jansen, and Y. S. Usenko. From StoCharts to MoDeST: a comparative reliability analysis of train radio communications. In *5th International Workshop on Software and Performance, WOSP'05*, pages 13–23. ACM Press, 2005.
- [20] M. Z. Kwiatkowska, G. Norman, and D. Parker. PRISM: Probabilistic symbolic model checker. In *12th International Conference on Computer Performance Evaluation: Modelling Techniques and Tools, TOOLS'02*, pages 200–204. Springer, 2002.

- [21] O. Landsiedel, K. Wehrle, and S. Gotz. Accurate prediction of power consumption in sensor networks. In *2nd IEEE workshop on Embedded Networked Sensors, EmNets'05*, pages 37–44. IEEE Computer Society, 2005.
- [22] G. Pongor. OMNeT: Objective modular network testbed. In *International Workshop on Modeling, Analysis, and Simulation On Computer and Telecommunication Systems*, pages 323–326. Society for Computer Simulation, 1993.
- [23] B. L. Titzer, D. K. Lee, and J. Palsberg. Avrora: scalable sensor network simulation with precise timing. In *4th International Symposium on Information Processing in Sensor Networks, IPSN'05*, page 67. IEEE Press, 2005.
- [24] X. Zeng, R. Bagrodia, and M. Gerla. GloMoSim: A library for parallel simulation of large-scale wireless networks. In *12th Workshop on Parallel and Distributed Simulation*, pages 154–161, 1998.

CHRISTIAN GROSS: comlet Verteilte Systeme GmbH, Amerikastraße 21,
D-66482, Zweibrücken, Germany.
E-mail: christian.gross@comlet.de

HOLGER HERMANN: Department of Computer Science, Saarland University
D-66123, Saarbrücken, Germany
E-mail: hermanns@cs.uni-sb.de

REZA PULUNGAN: Department of Computer Science, Saarland University
D-66123, Saarbrücken, Germany
E-mail: pulungan@cs.uni-sb.de

TESTING FOR INDEPENDENCE IN GAUSSIAN TIME SERIES USING THE CODIFFERENCE FUNCTION: SOME FINITE SAMPLE PROPERTIES

DEDI ROSADI

Abstract. In this paper, we consider a Portmanteau-type test of randomness for given time series data using a test statistic that differs from, but has the same general form as Box-Pierce Q-statistic, which is defined using the codifference function. We obtain that if the underlying distribution of data is normal, under hypothesis H_0 of zero autocorrelation up to order m , the proposed statistic is asymptotically χ_m^2 distributed. Simulation studies are performed to obtain the small sample properties (the size and the power) of the proposed statistic. We compare its performance against the (standard) Box-Pierce Q-Statistics which is defined based on the autocorrelation function, and a similar statistic based on the autocovariation function, introduced recently in Gallagher (2006).

Key words and Phrases: Portmanteau test, the codifference function, the sample autocorrelation function, the autocovariation function

1. INTRODUCTION

In this paper, we consider the problem of testing the hypothesis of zero autocorrelation up to order $m, m < n$ for given Gaussian time series data $X_1, X_2, \dots$, i.e. testing the hypothesis

$$H_0 : \rho(1) = \rho(2) = \dots = \rho(m) = 0, m < n \quad (1)$$

against the serial dependence alternatives,

$$H_1 : \rho(k) \neq 0 \text{ at least for one } k, \text{ with } k = 1, \dots, m$$

2000 *Mathematics Subject Classification:* 62M10

where $\rho(k)$ denotes the autocorrelation function of the process at lag k . If X_i 's are *independent and identically distributed* (iid) and normally distributed or more generally, $E(X_i^2) < \infty$, then this hypothesis can be tested using Box-Pierce Q-statistic (Box and Pierce, 1970)

$$Q_{n,m} = n \sum_{k=1}^m r_k^2 \quad (2)$$

or its finite sample correction, Ljung-Box statistic (Ljung and Box, 1978)

$$\tilde{Q}_{n,m} = n(n+2) \sum_{k=1}^m (n-k)^{-1} r_k^2 \quad (3)$$

with

$$r_k = \frac{\sum_{t=k+1}^n (X_t - \bar{X})(X_{t-k} - \bar{X})}{\sum_{t=1}^n (X_t - \bar{X})^2}, k = 1, \dots, m \quad (4)$$

The statistic $\tilde{Q}_{n,m}$ is asymptotically equivalent to $Q_{n,m}$ but its finite-sample distribution appears better approximated by the χ_m^2 eventhough the approximation remains far from perfect (see Ljung and Box, 1978; Dufour and Roy, 1986).

In this paper we consider to use a Portmanteau statistic which is defined using the codifference function (Rosadi, 2007). Our main interest here is to obtain the finite sample properties of this statistic, by comparing its performance with the statistic (3) and another Portmanteau statistic which is based on the autocovariation function, introduced recently in Gallagher (2006).

The rest of this paper is organized as follows. In the next section, we consider a Portmanteau type test of randomness using a test statistic that differs from, but has the same general form as Box-Pierce Q-statistic, which is defined using the (normalized) codifference function and the autocovariation function. In section three, we perform simulation studies to compare the finite sample properties of the considered statistics. The last section concludes.

2. THE PORTMANTEAU-TYPE STATISTICS

2.1. THE BOX-PIERCE Q-STATISTIC BASED ON THE CODIFFERENCE FUNCTION

The codifference function at lag- k of stationary process $\{X_t\}$ is defined as (Kokoszka and Taqqu, 1994; Yang, Petropulu and Pesquet, 2001)

$$\tau(k) = \tau_k(s, -s) = -\ln E(\exp(is(X_{t+k} - X_t))) + \ln E(\exp(isX_{t+k})) + \ln E(\exp(-isX_t)) \quad (5)$$

where $s \in \mathbb{R}$ and $k \in \mathbb{Z}$. A normalized codifference function can be defined as

$$I(k) = \frac{\tau(k)}{\tau(0)}$$

For symmetric stationary process, $\tau(-k) = \tau(k)$ holds. In Gaussian case, the codifference function is proportional to the covariance function, i.e. $\tau(s, -s; k) = -s^2\gamma(k)$, where $\gamma(\cdot)$ denotes the covariance function of the stationary process $\{X_t\}$. Thus, one directly obtains $I(k) = \rho(k)$ in Gaussian case. Further theoretical properties of the codifference function are discussed in Rosadi and Deistler (2005).

Given a sample $X_1, X_2, \dots, X_n$, Rosadi and Deistler (2005) propose an estimator for the codifference function at lag $k \in \mathbb{Z}$, $s \in \mathbb{R}$, $s \neq 0$ as

$$\hat{\tau}(s, -s; k) = \sqrt{1 - k/n} \times [-\ln \phi(s, -s; k) + \ln \phi(s, 0; k) + \ln \phi(0, -s; k)] \quad (6)$$

where for $u, v \in \mathbb{R}$

$$\phi(u, v; k) = \begin{cases} (n-k)^{-1} \sum_{t=1}^{n-k} \exp i(uX_{t+k} + vX_t) & \text{when } k \geq 0 \\ (n+k)^{-1} \sum_{t=1}^{n+k} \exp i(uX_{t-k} + vX_t) & \text{when } k < 0 \end{cases} \quad (7)$$

Accordingly, $\hat{I}(k) = \frac{\hat{\tau}(k)}{\hat{\tau}(0)}$ can be used as the estimator of the normalized codifference $I(k)$. Some asymptotic properties of the codifference function have been discussed in Rosadi and Deistler (2005).

To test the hypothesis (1), Rosadi (2007) propose to use the statistic

$$Q_{GC} = n \sum_{j=1}^m \left(\text{Re } \tilde{I}(j) \right)^2 \quad (8)$$

where

$$\text{Re } \tilde{I}(j) = \frac{\text{Re } \hat{I}(j)}{\hat{S}} \quad (9)$$

Here, $\text{Re } \hat{I}(j)$ denotes the real part of the sample normalized codifference $\hat{I}(j)$, and $\hat{S}^2$ denotes a consistent estimate of S^2 , the asymptotic variance of $\text{Re } \hat{I}(j)$, $j > 0$ given by

$$S^2 = (\sigma^2 s^2)^{-2} \left[e^{\sigma^2 s^2} \left(\frac{1}{2} e^{\sigma^2 s^2} - 1 \right) + e^{-\sigma^2 s^2} \left(\frac{1}{2} e^{-\sigma^2 s^2} - 1 \right) + 1 \right] \quad (10)$$

In Rosadi (2007), it has been shown that under the null hypothesis that $X_1, X_2, \dots$ is a sequence of iid $N(0, \sigma^2)$ random variables then Q_{GC} is asymptotically χ_m^2 distributed when $n \rightarrow \infty$.

2.2. A PORTMANTEAU-TYPE TEST BASED ON THE AUTOCOVARATION FUNCTION

The covariation function of a strictly stationary ergodic stochastic process with zero mean is given in Gallagher(2006) as

$$\eta(k) = E(X_t \delta_{t-k}), k = 0, \pm 1, \pm 2, \dots$$

where $\delta_t = \text{sign}(X_t)$ and its sample version can be defined (a corrected version of Gallagher, 2006)

$$\hat{\eta}(k) = \frac{1}{n} \sum_{t=l}^r X_t \delta_{t-k}, k = 0, \pm 1, \pm 2, \dots$$

where $l = \max(1, 1+k)$ and $r = \min(n, n+h)$. Gallagher (1998) shown that the $\hat{\eta}(k)$ is a (weak) consistent estimator for $\eta(k)$. Define

$$\lambda(k) = \frac{\eta(k)}{\eta(0)}$$

For Gaussian process, it can be shown that $\lambda(k) = \rho(k)$, with $\rho(k)$ denotes ACF of X_t

To test the hypothesis (1) for i.i.d. symmetric random variables with finite second moment σ^2 , Gallagher (2006) consider the statistic

$$Q_{Acov}(m) = (\hat{\mu}^2 + \hat{\sigma}^2)^{-1} 2n \sum_{j=1}^m \hat{T}^2(j)$$

where $\hat{\mu}$ is any consistent estimator for $E(|X|)$, and $T(k) = (\eta(k) + \eta(-k))/2$. Under null hypothesis, $Q_{Acov}(m)$ is asymptotically χ_m^2 distributed when $n \rightarrow \infty$.

3. SIMULATION RESULTS

Although all portmanteau statistics that we consider in this paper are asymptotically chi-squared distributed, the usage of quantile of chi-square distribution in the finite-sample case is still an open question. In this regard, we will investigate whether the finite-sample distribution of the statistic (8) is well approximated by the chi-square distribution, and furthermore how it compares to the performance of a “standard” statistic, such as Ljung-Box statistic (3) and also the performance of Q_{Acov} . Notice that a similar but more comprehensive simulation studies can be found in Rosadi (2007).

In the first part of simulation study, we investigate the size of the test. For the first part of simulation study, the parameterization in the simulation study is set up similarly as in Dufour and Roy (1986). We generate three “small” samples with sample size $n = 20, 50, 100$ using R/GNU-S version 2.4.1 (R Development Core Team, 2007), where for each sample size we use function `rnorm` to generate $T = 10000$ independent realization of standard Gaussian white noise. We calculate the statistic $\tilde{Q}_{n,m}$, Q_{GC} and Q_{acov} for several values of m . To appreciate the accuracy of χ^2 approximation, the nominal levels of test are chosen as $\theta = 1, 5, 10$ and 20 percent. Here we use the sample variance $\hat{\sigma}^2$ and the sample absolute mean $\frac{1}{n} \sum_{j=1}^n |X_j|$ for $\hat{\mu}$.

The sample normalized codifference function $\text{Re } \hat{I}(\cdot)$ and therefore, the statistic Q_{GC} , depend on s . Here $\text{Re } \hat{I}(\cdot)$ is a function of empirical characteristic function (ECF). For each lag $k = 1, \dots, m$, we consider a discrete estimation procedure for ECF to obtain $\text{Re } \hat{I}(k)$, i.e. we evaluate $\text{Re } \hat{I}(k)$ at some points $\mathbf{s} = s_1, \dots, s_r, r \geq 1$. The estimate of $\text{Re } \hat{I}(k)$ is calculated as the weighted value of the estimates among the choices of grid points, i.e. $\text{Re } \hat{I}(k) = \sum_{j=1}^r w_j \text{Re } \hat{I}(s_j, k)$, $\sum_{j=1}^r w_j = 1$. In this paper, we consider the weights as the simple average of the estimates with $w_j = 1/r$ (we call method $Q_{GCAvg.}$) or the negative exponential weighted average, with $w_j = \exp(-s_j^2) / \sum_{i=1}^r \exp(-s_i^2)$ (we call method $Q_{GCExp.}$).

For the calculation of ecf (and hence also for $\hat{I}(\cdot)$) depends on the choice of the number of points r and more importantly, on their location. For location of the grid point s , for a fixed choice of r , the $s_j, j = 1, \dots, r$ have to be chosen within an interval near zero, since at this interval the values of the real and imaginary parts of ecf (finite sample) will more accurately estimate the true values, and at the other choices of grid points, ecf (similarly for $\hat{I}(\cdot)$) will be highly bias. Using result in Rosadi (2007), we will use $\mathbf{s} = \{0.1, 0.2, \dots, 0.5\}$. Here we also note that for simulation using statistic Q_{acov} , we apply the method using mean corrected version of the data. Results of the experiments are summarized in Table 1.

The findings indicate that in many cases, using the choice $\mathbf{s} = \{0.1, 0.2, \dots, 0.5\}$, the empirical levels of Q_{GC} are closer to the theoretical ones than Ljung-Box statistic $Q_{n,m}$. This result will be more apparent when the lag m is relatively large compared to the sample size n , in which the difference between the empirical and theoretical levels of Ljung-Box statistic $Q_{n,m}$ can be significantly large. Furthermore, in general we can see that the performance of Ljung-Box statistic $Q_{n,m}$ is better than Q_{acov} . On the above choices of $\mathbf{s}$, the performance of both weighting methods is approximately the same.

To examine the power of the test, we consider the alternative hypothesis that the process is autoregressive process of order 1, AR(1). The sample sizes considered are $n = 50$ and 200 . For each sample size, the function `arma.sim` in R is used to generate AR(1) processes $X_t = \phi X_{t-1} + \epsilon_t$, with the parameters $\phi = 0.3$ and $\phi = 0.8$. We generate 10000 independent realization of the process with the normal standard ($N(0,1)$) innovations. The statistic Q_{GC} is evaluated using grid points $\mathbf{s} = \{0.1, 0.2, \dots, 0.5\}$. For the simulation, the level is chosen to be 5% and we consider several values of m . Again, we note that for simulation using statistic Q_{acov} , we apply the method using mean corrected version of the data. The simulation results are summarized in Table 2.

Throughout the experiments, we found that the power of Q_{GC} is approximately the same (slightly lower) as the power of statistic $\tilde{Q}_{n,m}$. The power of the statistic Q_{GC} is better than the power of Q_{acov} . In general, we may conclude that the Pormanteau statistic, which is based on the codifference function, is a powerful statistic for detecting serial dependence in AR(1) form.

4. SUMMARY & CONCLUSION

The above results suggest that for cases when the true underlying distribution is normal, under suitable choice of the design parameter $\mathbf{s}$, the new Portmanteau statistic Q_{GC} is better approximated by χ_m^2 than Ljung-Box statistic $Q_{n,m}$ or Q_{acov} . This result will be more apparent when the lag m is relatively large compared to the sample size n (i.e., closer to $m/n \leq 1/4$). We can conclude that the new Portmanteau statistic Q_{GC} is a useful tool for testing the randomness of time series data, and can be used in addition to the procedure which is based on Ljung-Box statistic $Q_{n,m}$.

Acknowledgement. The financial support from the Ministry of Science and Technology of Austria via ÖAD TSA project and also from the DIKTI Indonesia via Fundamental Research Project number UGM/PFD/35/2005 and LPPM-UGM/583/2007 is gratefully acknowledged.

REFERENCES

1. G. E. P. BOX AND D. A. PIERCE, Distribution of residual correlations in autoregressive-integrated moving average time series models, *JASA*, **65**, 1509 – 1526, 1970.
2. J-M. DUFOUR AND R. ROY, Generalized Portmanteau statistics and tests of randomness, *Communications and Statistics: Theory and Methods*, **15**, 2953–2972, 1986.
3. C.M. GALLAGHER, *Fitting ARMA models to heavy tailed data*, PhD thesis. University of California, Santa Barbara, 1998.
4. C.M. GALLAGHER, Detecting Dependence in Heavy-Tailed Time Series using Portmanteau-type dependence tests, *International Mathematical Forum*, **1** (10), 455–469, 2006.
5. P. S. KOKOSZKA AND M.S. TAQQU, Infinite variance stable ARMA processes, *Journal of Time Series Analysis*, **15**, 203 – 220, 1994.
6. G.M. LJUNG AND G. E. P. BOX, On a measure of lack of fit in time series models, *Biometrika*, **65**, 297 – 303, 1978.
7. R DEVELOPMENT CORE TEAM, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2007. ISBN 3-900051-00-3.
8. D. ROSADI, Testing for Independence in Gaussian Time Series using the Codifference Function. Submitted to, *Journal of Computational Statistics and Data Analysis*, 2007.
9. D. ROSADI AND DEISTLER, M. , *Estimating the codifference function of linear time series models with infinite variance*, Mimeo, 2005. Institute for Mathematical Methods in Economics, Research Unit Econometrics and System Theory, Vienna University of Technology, 2005.
Available on <http://dedirosadi.staff.ugm.ac.id/paper/codifference.pdf>.
10. Q. YANG, A. PETROPULU AND J. C. PESQUET, Estimating long-range dependence in impulsive traffic flows. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 3413 –3416, 2001.

DEDI ROSADI: Department of Mathematics, Gadjah Mada University, Sekip Utara, Yogyakarta, Indonesia.

E-mail: dedirosadi@ugm.ac.id.

Table 1. Empirical significance levels (in percentage)

level %	method	$n = 20$					$n = 50$					$n = 100$							
		m					m					m							
		1	3	5	10		1	3	5	10	15	25	1	3	5	10	15	25	50
0.01	$\hat{Q}_{GCAvg.}$	0.63	0.85	1.41	2.12		0.85	1.13	1.5	2.03	2.81	3.34	0.87	1.01	1.17	1.51	1.99	2.79	4.03
	$\hat{Q}_{GCExp.}$	0.61	0.82	1.4	2.1		0.81	1.13	1.48	2.02	2.8	3.33	0.86	0.98	1.16	1.51	1.99	2.77	3.98
	$\hat{Q}_{n,m}$	0.81	1.07	1.79	2.75		0.9	1.26	1.71	2.2	3.04	3.89	0.87	0.99	1.29	1.59	2.02	2.89	4.42
0.05	$\hat{Q}_{acov}$	0.59	0.59	0.72	0.61		0.76	0.8	1.05	1.24	1.19	0.69	1.07	0.88	0.94	1.24	1.15	1.1	0.35
	$\hat{Q}_{GCAvg.}$	4.45	4.64	4.99	6.45		4.79	4.65	4.96	6.29	7.14	8.21	4.87	5.13	5.18	5.61	6.44	7.8	9.98
	$\hat{Q}_{GCExp.}$	4.43	4.53	4.89	6.41		4.8	4.67	4.96	6.27	7.1	8.15	4.87	5.13	5.15	5.62	6.42	7.78	9.93
0.1	$\hat{Q}_{n,m}$	5.11	5.46	5.9	7.89		5.27	5.29	5.59	6.97	7.84	9.5	5.12	5.45	5.59	5.95	6.69	8.42	10.67
	$\hat{Q}_{acov}$	3.91	3.23	3	2.33		4.43	4.11	4.07	3.54	2.98	1.69	5	4.97	4.79	4.56	4.29	3.68	1.29
	$\hat{Q}_{GCAvg.}$	10.02	8.84	9.36	10.44		9.87	9	9.25	10.63	11.56	12.36	9.87	9.51	9.54	10.08	10.93	12.36	14.24
	$\hat{Q}_{GCExp.}$	9.98	8.72	9.31	10.35		9.84	9	9.21	10.56	11.5	12.33	9.84	9.52	9.51	10.01	10.91	12.28	14.2
	$\hat{Q}_{n,m}$	11.13	10.3	10.73	13.11		10.63	9.6	10.09	11.51	12.6	13.88	9.99	9.73	9.9	10.69	11.46	13.01	15.38
	$\hat{Q}_{acov}$	9.36	6.58	5.78	3.56		9.7	8.72	8.11	6.92	5.52	3	9.76	9.54	8.84	8.28	7.1	5.53	2.05
0.2	$\hat{Q}_{GCAvg.}$	19.78	18.73	18.41	18.91		21.01	19.49	19.15	20.04	20.42	20.99	19.27	19.3	18.82	19.49	19.96	21.43	22.21
	$\hat{Q}_{GCExp.}$	19.74	18.65	18.34	18.82		21.03	19.44	19.12	19.96	20.44	20.88	19.26	19.3	18.71	19.51	19.9	21.4	22.18
	$\hat{Q}_{n,m}$	21.52	21.01	20.99	23.17		21.48	20.36	20.14	21.39	22.21	23.41	19.7	19.7	19.39	20.15	20.84	22.36	23.79
	$\hat{Q}_{acov}$	19.13	15.15	12.18	7.11		20.41	18.24	16.31	13.58	10.77	5.83	19.29	18.7	18.11	16.12	14.47	10.49	3.55

Empirical levels (in %) of the test based on $\hat{Q}_{GC}$, $\hat{Q}_{n,m}$ and $\hat{Q}_{acov}$ at some nominal significance level (in %). Under hypothesis H_0 , the process is assumed $N(0, 1)$ distributed.

Table 2. Empirical power

n	method	$\phi = 0.3$			$\phi = 0.8$		
		m			m		
		5	15	30	5	15	30
50	$Q_{GCAvg.}$	27.06	23.95	24.09	99.4	97.64	96.69
	$Q_{GCExp.}$	27.13	23.86	24.07	99.42	97.64	96.73
	$\tilde{Q}_{n,m}$	29.23	25.52	28.38	99.5	98.02	97.63
	$\tilde{Q}_{acov}$	20.07	13	6.82	98	93.49	86.06
200	$Q_{GCAvg.}$	91.6	78.13	67.87	100	100	100
	$Q_{GCExp.}$	91.61	78.19	67.86	100	100	100
	$\tilde{Q}_{n,m}$	92.11	79.2	69.23	100	100	100
	$\tilde{Q}_{acov}$	82.62	63.92	48.47	100	100	100

Empirical power (in %) of the test based on Q_{GC} (where the weighting method are the simple average and the exponentially weighted average), $\tilde{Q}_{n,m}$ and $\tilde{Q}_{acov}$ at nominal significance level 5%, where the assumed model is i.i.d. $N(0,1)$, but the true model is $AR(1)$, $X_t = \phi X_{t-1} + \epsilon_t$, ϵ_t is $N(0,1)$.

APPLICATIONS OF REPRODUCING KERNELS TO NUMERICAL ANALYSIS THROUGH THE TIKHONOV REGULARIZATION

SABUROU SAITOH

Abstract. This paper is a survey article based on our recent papers on inverse problems using Tikhonov regularization and reproducing kernels. The Tikhonov regularization is a basically important idea and method in numerical analysis, however, the extremal functions of the Tikhonov functional are commonly represented by using the associated singular values and singular functions for the case of compact operators. So, their representations are restrictive and abstract in a sense. Our representations are both analytical and numerical. We introduce a general theory of the Tikhonov regularization using the theory of reproducing kernels including error estimates and convergence rates. We shall present the general theory and its concrete results for the typical inverse problems for heat conduction and for numerical real inversion formulas of the Laplace transform with computer graphs (in the cited references) as the evidences of the power of our inverse formulas.

Key words and Phrases: Reproducing kernel, Tikhonov regularization, numerical analysis, Gaussian convolution, heat equation, real inversion formula of the Laplace transform, inverse problem, spectral theory, noisy data, error estimate, sinc method.

1. REPRODUCING KERNELS

First we recall a basic relation between linear mappings in the framework of Hilbert spaces and reproducing kernels. In particular, we can see here why we meet ill-posed problems, indeed, we can see the idea and method for the avoidance of

2000 Mathematics Subject Classification: Primary 44A15, 35K05, 30C40

the ill-posed problems in the framework of Hilbert spaces. However, this will be a mathematical theory and for the purpose of developing numerical methods, we will need the idea of Tikhonov regularization discussed in Section 5. However we will need essentially the applications of the theory of reproducing kernels to both mathematical and numerical theories for bounded linear operator equations in the framework of Hilbert spaces. See [8,11] for a general elementary theory of inverse problems and ill-posed problems.

We consider any *positive matrix* $K(p, q)$ on a fixed set E ; that is, for an abstract set E the complex-valued function $K(p, q)$ on $E \times E$ satisfies, for any finite points $\{p_j\}$ of E and for any complex numbers $\{C_j\}$,

$$\sum_j \sum_{j'} C_j \overline{C_{j'}} K(p_{j'}, p_j) \geq 0.$$

Then, by the fundamental theorem by Moore–Aronszajn, we have:

Proposition 0.0.1 ([2]) *For any positive matrix $K(p, q)$ on E , there exists a uniquely determined functional Hilbert space (abbreviated RKHS) H_K comprising functions $\{f\}$ on E and admitting the reproducing kernel $K(p, q)$ satisfying and characterized by*

$$K(\cdot, q) \in H_K \text{ for any } q \in E \quad (1)$$

and, for any $q \in E$ and for any $f \in H_K$

$$f(q) = (f(\cdot), K(\cdot, q))_{H_K}. \quad (2)$$

For some general properties of reproducing kernel Hilbert spaces and for various constructions of the RKHS H_K from a positive matrix $K(p, q)$, see the recent book [37] and its Chapter 2, Section 5, respectively.

2. CONNECTIONS WITH LINEAR MAPPINGS

Let us connect linear mappings in the framework of Hilbert spaces with reproducing kernels ([30]).

For an abstract set E and for any Hilbert (possibly finite-dimensional) space $\mathcal{H}$, we shall consider an $\mathcal{H}$ -valued function $\mathbf{h}$ on E

$$\mathbf{h} : E \longrightarrow \mathcal{H} \quad (3)$$

and the linear mapping from $\mathcal{H}$ into a linear space comprising functions on E , given by $\mathbf{f} \longrightarrow f$, where

$$f(p) = (\mathbf{f}, \mathbf{h}(p))_{\mathcal{H}} \text{ for } \mathbf{f} \in \mathcal{H}. \quad (4)$$

This represents, in particular, the Fredholm integral equations of the first kind in the framework of Hilbert spaces.

For this linear mapping (4), we form the positive matrix $K(p, q)$ on E defined by

$$K(p, q) = (\mathbf{h}(q), \mathbf{h}(p))_{\mathcal{H}} \quad \text{on } E \times E, \quad (5)$$

which is, by Proposition 0.0.1, a reproducing kernel.

Then, we have the following fundamental results:

(I) For the RKHS H_K admitting the reproducing kernel $K(p, q)$ defined by (5), the images $\{f(p)\}$ by (4) for $\mathcal{H}$ are characterized as the members of the RKHS H_K .

(II) In general, we have the inequality in (4)

$$\|f\|_{H_K} \leq \|\mathbf{f}\|_{\mathcal{H}}, \quad (6)$$

however, for any $f \in H_K$ there exists a uniquely determined $\mathbf{f}^* \in \mathcal{H}$ satisfying

$$f(p) = (\mathbf{f}^*, \mathbf{h}(p))_{\mathcal{H}} \quad \text{on } E \quad (7)$$

and

$$\|f\|_{H_K} = \|\mathbf{f}^*\|_{\mathcal{H}}. \quad (8)$$

In (6), the isometry holds if and only if $\{\mathbf{h}(p); p \in E\}$ is complete in $\mathcal{H}$.

(III) We can obtain the inversion formula for (4) in the form

$$f \longrightarrow \mathbf{f}^*, \quad (9)$$

by using the RKHS H_K .

However, this inversion formula will depend on, case by case, the realizations of the RKHS H_K .

(IV) Conversely, if we have an isometric mapping $\tilde{L}$ from the RKHS H_K admitting a reproducing kernel $K(p, q)$ on E onto a Hilbert space $\mathcal{H}$, then the mapping is linear and its isometric inversion $\tilde{L}^{-1}$ is represented in the form (4). Here, the Hilbert space $\mathcal{H}$ -valued function $\mathbf{h}$ satisfying (3) and (4) is given by

$$\mathbf{h}(p) = \tilde{L}K(\cdot, p) \quad \text{on } E \quad (10)$$

and, $\{\mathbf{h}(p) : p \in E\}$ is complete in $\mathcal{H}$.

When (4) is isometrical, sometimes we can use the isometric mapping for a realization of the RKHS H_K , conversely — that is, if the inverse L^{-1} of the linear mapping (4) is known, then we have $\|f\|_{H_K} = \|L^{-1}f\|_{\mathcal{H}}$.

3. GENERAL APPLICATIONS

We shall state some general applications of the results (I)~(IV) to several wide subjects and their basic references:

(1) Linear mappings ([32,34,37]).

The fact that the image spaces of linear mappings in the framework of Hilbert spaces are characterized as reproducing kernel Hilbert spaces defined by (5) is the most important one in the general theory of reproducing kernels. This means that the theory of reproducing kernels is fundamental and a general concept in mathematics. To look for the characterization of the image space is a starting point when we consider the linear equation (4). (II) gives a generalization of the Pythagorean theorem (see also [29]) and means that in the general linear mapping (4) there exists essentially an isometric identity between the input and the output. (III) gives a generalized (natural) inverse (solution) of the linear mapping (equation) (4). (IV) gives a general method determining and constructing the linear system from an isometric relation between outputs and inputs by using the reproducing kernel in the output space.

(2) Linear mappings among smooth functions ([42]).

We considered linear mappings in the framework of Hilbert spaces, however, we can consider linear mappings in the framework of Hilbert spaces comprising smooth functions, similarly. Conversely, reproducing kernel Hilbert spaces are considered as the images of some Hilbert spaces by considering some decomposed representations (5) of the reproducing kernels. Such decomposition is, in general, possible. This idea is important in [43] and also in the following items (6) and (7).

(3) Nonharmonic linear mappings ([32]).

If the linear system vectors $\mathbf{h}(p)$ move in a small way (perturbation of the linear systems) in the Hilbert space $\mathcal{H}$, then we can not calculate the related positive matrix (5), however, we can discuss the inversion formula and an isometric identity of the linear mapping. The prototype result is the Paley-Wiener theorem on nonharmonic Fourier series.

(4) Various norm inequalities ([35],[38]).

Relations among positive matrices correspond to those of the associated reproducing kernel Hilbert spaces, by the minimum principle. So, we can derive various norm inequalities among reproducing kernel Hilbert spaces from relations among positive matrices. We were able to derive many beautiful norm inequalities.

(5) Nonlinear mappings ([38],[39]).

In a very general nonlinear mapping of a reproducing kernel Hilbert space, we can look for a natural reproducing kernel Hilbert space containing the image space and furthermore, we can derive a natural norm inequality in the

nonlinear mapping. What is a basic relation between linear mappings and non-linear mappings in the framework of reproducing kernel Hilbert spaces? It seems that the theory of reproducing kernels gives a fundamental and interesting answer for this question.

(6) Linear (singular) integral equations ([15],[43]).

We can apply our general theory to various integral equations containing Volterra type in the framework of Hilbert spaces. We were able to apply our theory to linear singular integral equations, also ([15]).

(7) Linear differential equations with variable coefficients ([3,43,51]).

In linear integro-differential equations with general variable coefficients, we can discuss the existence and construction of the solutions, if the solutions exist. This background method is called a backward transformation method and by reducing the equations to Fredholm integral equations of the first kind and we can discuss the classical solutions, in very general linear equations.

(8) Approximation theory ([1],[37]).

Reproducing kernel Hilbert spaces are very nice function spaces, because the point evaluations are continuous. Therefore, reproducing kernels are a fundamental tool in the related approximation theory. See also Section 4.

(9) Representations of inverse functions ([36],[37]).

For any mapping, we discussed the problem of representing its inverse in terms of the direct mapping and we derived a unified method for this problem. As a simple example, we can represent the Taylor coefficients of the inverse of the Riemann mapping function on the unit disc on the complex plane in terms of the Riemann mapping function. This fact was important in the representation of analytic functions in terms of local data in ([44,46]).

(10) Various operators among Hilbert spaces ([39]).

Among various abstract Hilbert spaces, we can introduce various operators of sum, product, integral and derivative by using the linear mapping or very general nonlinear mappings. The prototype operator is convolution and we discussed it from a wide and general viewpoint with concrete examples ([49]).

(11) Sampling theorems ([37], Chapter 4, Section 2).

The Whittaker-Kotel'nikov-Shannon sampling theorem may be interpreted by (I) and (II) very well, and we can discuss the truncation errors in the sampling theory. J. R. Higgins [13] established a fully general theory for [37]. See also, Section 7, as its essence.

(12) Interpolation problems of Pick-Nevanlinna type ([32],[33]).

General and abstract theory of interpolation problems of Pick-Nevanlinna type may be discussed by using the general theory of reproducing kernels. For a recent book, see ([1]).

- (13) Analytic extension formulas and their applications ([44],[47]).

We were able to obtain various analytic extension formulas and their applications from various isometric identities (II). For their applications to nonlinear partial differential equations, see the survey article by N. Hayashi [12].

- (14) Inversions of a family of bounded linear operators on a Hilbert space into various Hilbert spaces ([50]).

From fully general theorems for the Pythagorean theorem, we derived the concept of inversions of a family of bounded linear operators on a Hilbert space into various Hilbert spaces.

4. BEST APPROXIMATION PROBLEMS

We shall consider the fundamental and well-known best approximation problem

$$\inf_{f \in H_K} \|Lf - \mathbf{d}\|_{\mathcal{H}} \quad (11)$$

for a vector $\mathbf{d}$ in $\mathcal{H}$ where $L : H_K \longrightarrow \mathcal{H}$ is a bounded linear operator. Then, we have, by our theory in Section 2

Proposition 0.0.2 ([6],[37]) *For a vector $\mathbf{d}$ in $\mathcal{H}$, there exists a function $\tilde{f}$ in H_K such that*

$$\inf_{f \in H_K} \|Lf - \mathbf{d}\|_{\mathcal{H}} = \|L\tilde{f} - \mathbf{d}\|_{\mathcal{H}} \quad (12)$$

if and only if, for the RKHS H_k admitting the reproducing kernel defined by

$$k(p, q) = (L^*LK(\cdot, q), L^*LK(\cdot, p))_{H_K}, \quad (13)$$

$$L^*\mathbf{d} \in H_k. \quad (14)$$

Furthermore, if the best approximation $\tilde{f}$ satisfying (12) exists, then there exists a unique extremal function $f_{\mathbf{d}}$ with the minimum norm in H_K , and the function $f_{\mathbf{d}}$ is expressible in the form

$$f_{\mathbf{d}}(p) = (L^*\mathbf{d}, L^*LK(\cdot, p))_{H_k} \quad \text{on } E. \quad (15)$$

In Proposition 0.0.2, note that

$$(L^*\mathbf{d})(p) = (L^*\mathbf{d}, K(\cdot, p))_{H_K} = (\mathbf{d}, LK(\cdot, p))_{\mathcal{H}}; \quad (16)$$

that is, $L^*\mathbf{d}$ is expressible in terms of the known $\mathbf{d}, L, K(p, q)$ and $\mathcal{H}$.

This representation will, in particular, show that for bounded linear operators on a Hilbert space, if the Hilbert space admits a reproducing kernel, then the representations of their adjoint operators are concretely given. This fact will be important for general bounded linear operator theory.

$f_{\mathbf{d}}$ in (15) is the Moore-Penrose generalized inverse solution $L^\dagger \mathbf{d}$ of the equation $Lf = \mathbf{d}$. Therefore, Proposition 0.0.2 gives a necessary and sufficient condition for the existence of the Moore-Penrose generalized inverse. Proposition 0.0.2 is rigid and is not useful in practical applications, because, practical data contain noise or errors and the criteria (14) is not suitable. So, we shall consider the Tikhonov regularization and we shall establish a good relation between the Tikhonov regularization and the theory of reproducing kernels.

5. SPECTRAL THEORY

In order to discuss operator equations for general bounded linear operators L , following [8] we shall fix the well-established theory among spectral theory, the Moore-Penrose generalized inverse and the Tikhonov regularization. See [11] for the corresponding results for compact operators L .

Let $\{E_\lambda\}$ be a spectral family for the self-adjoint operator L^*L . If L^*L is continuously invertible, then

$$(L^*L)^{-1} = \int \frac{1}{\lambda} dE_\lambda.$$

In this case, the Moore-Penrose generalized inverse (15) can be represented by the Gaussian normal equation

$$f_{\mathbf{d}}(p) = \int \frac{1}{\lambda} dE_\lambda L^* \mathbf{d}. \quad (17)$$

If $\mathcal{R}(L)$ is non-closed and $\mathbf{d} \notin \mathcal{D}(L^\dagger)$, i.e. if the equation $Lf = \mathbf{d}$ is ill-posed, then the integral in (17) does not exist. Then, we shall define

$$f_{\mathbf{d},\alpha}(p) = \int \frac{1}{\lambda + \alpha} dE_\lambda L^* \mathbf{d}. \quad (18)$$

By construction, the operator on the right-hand side of (18) acting on $\mathbf{d}$ is continuous, so that, for noisy data $\mathbf{d}^\delta$ with $\|\mathbf{d} - \mathbf{d}^\delta\|_{\mathcal{H}} \leq \delta$, we can bound the error between $f_{\mathbf{d},\alpha}$ and

$$f_{\mathbf{d},\alpha}^\delta(p) = \int \frac{1}{\lambda + \alpha} dE_\lambda L^* \mathbf{d}^\delta \quad (19)$$

as follows:

Proposition 0.0.3 ([8], pages 71-73) *For any $\mathbf{d} \in \mathcal{D}(L^\dagger)$,*

$$\lim_{\alpha \rightarrow 0} (L^*L + \alpha I)^{-1} L^* \mathbf{d} = \lim_{\alpha \rightarrow 0} f_{\mathbf{d},\alpha} = f_{\mathbf{d}}. \quad (20)$$

Furthermore,

$$\|Lf_{\mathbf{d},\alpha} - Lf_{\mathbf{d},\alpha}^{\delta}\|_{\mathcal{H}} \leq \delta \quad (21)$$

and

$$\|f_{\mathbf{d},\alpha} - f_{\mathbf{d},\alpha}^{\delta}\|_{H_K} \leq \frac{\delta}{\sqrt{\alpha}}. \quad (22)$$

Proposition 0.0.4 ([8], pages 117-118) *For any $\mathbf{d}^{\delta} \in \mathcal{D}(L^{\dagger})$ with $\|\mathbf{d} - \mathbf{d}^{\delta}\|_{\mathcal{H}} \leq \delta$, the function $f_{\mathbf{d},\alpha}^{\delta}$ defined by (19) is the unique minimizer of the Tikhonov functional*

$$\inf_{f \in H_K} \{\alpha \|f\|_{H_K}^2 + \|\mathbf{d}^{\delta} - Lf\|_{\mathcal{H}}^2\}. \quad (23)$$

If $\alpha = \alpha(\delta)$ is such that

$$\lim_{\delta \rightarrow 0} \alpha(\delta) = 0$$

and

$$\lim_{\delta \rightarrow 0} \frac{\delta^2}{\alpha(\delta)} = 0,$$

then

$$\lim_{\delta \rightarrow 0} f_{\mathbf{d},\alpha}^{\delta} = f_{\mathbf{d}} = L^{\dagger}(\mathbf{d}). \quad (24)$$

These results are important, because real physical data contain error and noises.

6. REPRESENTATION OF THE EXTREMAL FUNCTIONS IN THE TIKHONOV REGULARIZATION

Our main purpose here is to give an effective representation of the extremal function $f_{\mathbf{d},\alpha}$ or $f_{\mathbf{d},\alpha}^{\delta}$ in the Tikhonov regularization, since the representation by spectral theory is abstract, in many practical problems.

We set

$$K_L(\cdot, p; \alpha) = (L^*L + \alpha I)^{-1} K(\cdot, p).$$

Then, by introducing the inner product, for any fixed positive $\alpha > 0$

$$(f, g)_{H_K(L; \alpha)} = \alpha(f, g)_{H_K} + (Lf, Lg)_{\mathcal{H}}, \quad (25)$$

we shall construct the Hilbert space $H_K(L; \alpha)$ comprising functions of H_K . This space, of course, admits a reproducing kernel. Furthermore, we obtain

Proposition 0.0.5 ([52]) *The extremal function $f_{\mathbf{d},\alpha}(p)$ in the Tikhonov regularization*

$$\inf_{f \in H_K} \{ \alpha \|f\|_{H_K}^2 + \|\mathbf{d} - Lf\|_{\mathcal{H}}^2 \} \quad (26)$$

is represented in terms of the kernel $K_L(p, q; \alpha)$ as follows:

$$f_{\mathbf{d},\alpha}(p) = (\mathbf{d}, LK_L(\cdot, p; \alpha))_{\mathcal{H}} \quad (27)$$

where the kernel $K_L(p, q; \alpha)$ is the reproducing kernel for the Hilbert space $H_K(L; \alpha)$ and it is determined as the unique solution $\tilde{K}(p, q; \alpha)$ of the equation:

$$\tilde{K}(p, q; \alpha) + \frac{1}{\alpha} (L\tilde{K}_q, LK_p)_{\mathcal{H}} = \frac{1}{\alpha} K(p, q) \quad (28)$$

with

$$\tilde{K}_q = \tilde{K}(\cdot, q; \alpha) \in H_K \quad \text{for } q \in E, \quad (29)$$

and

$$K_p = K(\cdot, p) \in H_K \quad \text{for } p \in E.$$

In (27), when $\mathbf{d}$ contains error or noises, we need its error estimate. For this, we can obtain the general result:

Proposition 0.0.6 ([15],[25]). *In (27), we obtain the estimate*

$$|f_{\mathbf{d},\alpha}(p)| \leq \frac{1}{\sqrt{\alpha}} \sqrt{K(p, p)} \|\mathbf{d}\|_{\mathcal{H}}.$$

Propositions 0.0.5 and 0.0.6 will show the importance to consider bounded linear operators on a Hilbert space which admits a reproducing kernel. For many concrete applications of these general theorems, see, for example, [3,15,19-28,45,46,50,55].

7. A TYPICAL EXAMPLE FOR THE INVERSE OF THE HEAT CONDUCTION

We shall give simple approximate real inversion formulas for the Gaussian convolution (the Weierstrass transform)

$$u_F(x, t) = (L_t F)(x) = \frac{1}{(4\pi t)^{n/2}} \int_{\mathbf{R}^n} F(\xi) \exp \left\{ -\frac{|\xi - x|^2}{4t} \right\} d\xi \quad (30)$$

for the functions of $L_2(\mathbf{R}^n)$. This integral transform which represents the solution $u(x, t)$ of the heat equation

$$u_t(x, t) = u_{xx}(x, t) \quad \text{on} \quad \mathbf{R}^n \times \{t > 0\}$$

satisfying the initial condition

$$u(x, 0) = F(x) \quad \text{on} \quad \mathbf{R}^n,$$

is very fundamental and has many applications to mathematical sciences. See the recent article [57] and its references for its many significant applications to medical science and physics.

Over twenty years ago, in the one dimensional case $n = 1$, the author [31] gave a surprise characterization of the image $u_F(x, t)$ of (30) for $L_2(\mathbf{R}) = L_2(\mathbf{R}, dx)$ functions in terms of an analytic function and established a very simple complex inversion formula following the idea (I). The paper created a new method and many applications to general integral transforms in the framework of Hilbert spaces and analytic extension formulas. See, for example [32] and [37], and their many references. However, in particular, its real inversion formulas are very involved, for example, recall that:

For a bounded and continuous function $F(x)$ and for $t = 1$, for the differential operator $D = \frac{d}{dx}$

$$e^{-D^2}[(L_1 F)(x)] = F(x) \quad \text{pointwisely on } \mathbf{R}$$

([16], p. 182). So, one might think that its real inversion formulas will be essentially involved for catching "analyticity" in terms of the data on the real line as in the real inversion formulas of the Laplace transform. See also [17,18] for recent related articles.

Indeed, this inverse problem is very famous as a typical ill-posed problem that is very difficult.

In those papers [23,25], however we were able to obtain simple and practical approximate real inversion formulas by the method in Section 6 using the Sobolev reproducing Hilbert spaces. Furthermore, we illustrated their numerical experiments by using computers and we can realize that we were able to obtain practical real inversion formulas.

In [25], we applied the Paley-Wiener spaces as the reproducing kernel Hilbert spaces in the above theory and we got an improved numerical inversion.

At first we shall fix notations and basic results in the Paley-Wiener spaces following the book by F. Stenger[56] and at the same time we shall show the basic relation of the sampling theory and the theory of reproducing kernels.

We shall consider the integral transform, for $L_2(\mathbf{R}^n, (-\pi/h, +\pi/h)^n)$, ($h > 0$) functions g

$$f(z) = \frac{1}{(2\pi)^n} \int_{\mathbf{R}^n} \chi_h(t) g(t) e^{-iz \cdot t} dt. \quad (31)$$

Here, $z = (z_1, z_2, \dots, z_n)$, $t = (t_1, t_2, \dots, t_n)$, $dt = dt_1 \cdot dt_2 \cdots dt_n$, $z \cdot t = z_1 t_1 + \cdots + z_n t_n$ and

$$\chi_h(t) = \prod_{\nu=1}^n \chi(t_\nu, (-\pi/h, +\pi/h)),$$

the characteristic function χ of $(-\pi/h, +\pi/h)$. In order to identify the image space following the idea (I), we form the reproducing kernel

$$\begin{aligned} K_h(z, \bar{u}) &= \frac{1}{(2\pi)^n} \int_{\mathbf{R}^n} \chi_h(t) e^{-iz \cdot t} \overline{e^{-iu \cdot t}} dt \\ &= \prod_{\nu} \frac{1}{\pi(z_\nu - \bar{u}_\nu)} \sin \frac{\pi}{h} (z_\nu - \bar{u}_\nu). \end{aligned} \quad (32)$$

The image space of (31) is called the Paley-Wiener space $W(\frac{\pi}{h})$ ($:= W_h$) comprised of all analytic functions of exponential type satisfying, for each ν , for some constant C_ν and as $z_\nu \rightarrow \infty$

$$|f(z_1, \dots, z_\nu, z_{\nu+1}, \dots, z_n)| \leq C_\nu \exp\left(\frac{\pi|z_\nu|}{h}\right)$$

and

$$\int_{\mathbf{R}^n} |f(x)|^2 dx < \infty.$$

From the identity, for multi-index $j = (j_1, j_2, \dots, j_n) \in \mathcal{Z}^n$

$$K_h(jh, j'h) = \prod_{\nu=1}^n \frac{1}{h} \delta(j_\nu, j'_\nu)$$

(the Kronecker's δ), for each ν , since $\delta(j_\nu, j'_\nu)$ is the reproducing kernel for the Hilbert space ℓ^2 , from (II) and the Parseval's identity we have the isometric identities in (31)

$$\begin{aligned} \frac{1}{(2\pi)^n} \int_{\mathbf{R}^n} |g(t)|^2 dt &= h^n \sum_j |f(jh)|^2 \\ &= \int_{\mathbf{R}^n} |f(x)|^2 dx. \end{aligned}$$

That is, the reproducing kernel Hilbert space H_{K_h} with $K_h(z, \bar{u})$ is characterized as a space comprising the Paley-Wiener space W_h and with the norms above in the both senses of discrete and continuous versions. Here we used the well-known result that $\{jh\}_j$ is a uniqueness set for the Paley-Wiener space W_h ; that is, $f(jh) = 0$ for all j implies $f \equiv 0$. Then, the reproducing property of $K_h(z, \bar{u})$ states that

$$\begin{aligned} f(x) &= (f(\cdot), K_h(\cdot, x))_{H_{K_h}} = h^n \sum_j f(jh) K_h(jh, x) \\ &= \int_{\mathbf{R}^n} f(\xi) K_h(\xi, x) d\xi, \end{aligned}$$

in particular, on the real space x . This representation is the sampling theorem which represents the whole data $f(x)$ in terms of the discrete data $\{f(jh)\}_j$. For a general theory of sampling and error estimates for some finite points $\{hj\}_j$, see [13] and [37].

Following our general theory, we can obtain the concrete results:

Proposition 0.0.7 ([25]) *For any function $g \in L_2(\mathbf{R}^n)$ and for any $\lambda > 0$, the best approximate function $F_{t,\lambda,h,g}^*$ in the sense*

$$\begin{aligned} & \inf_{F \in H_{K_h}} \left\{ \lambda \|F\|_{H_{K_h}}^2 + \|g - u_F(\cdot, t)\|_{L_2(\mathbf{R}^n)}^2 \right\} \\ &= \lambda \|F_{t,\lambda,h,g}^*\|_{H_{K_h}}^2 + \|g - u_{F_{t,\lambda,h,g}^*}(\cdot, t)\|_{L_2(\mathbf{R}^n)}^2 \end{aligned} \quad (33)$$

exists uniquely and $F_{t,\lambda,h,g}^*$ is represented by

$$F_{t,\lambda,h,g}^*(x) = \int_{\mathbf{R}^n} g(\xi) Q_{t,\lambda,h}(\xi - x) d\xi \quad (34)$$

for

$$Q_{t,\lambda,h}(\xi - x) = \frac{1}{(2\pi)^n} \int_{\mathbf{R}^n} \frac{\chi_h(p) e^{-ip \cdot (\xi - x)} dp}{\lambda e^{|p|^2 t} + e^{-|p|^2 t}}.$$

If, for $F \in H_{K_h}$ we consider the output $u_F(x, t)$ and we take $u_F(\xi, t)$ as g , then we have the result: as $\lambda \rightarrow 0$

$$F_{t,\lambda,h,g}^* \rightarrow F, \quad (35)$$

uniformly.

Here we note the fact that for the Sobolev space case, for $\lambda = 0$ the corresponding representation (34) does not exist ([23],[52]), meanwhile for the Paley-Wiener space $W\left(\frac{\pi}{h}\right)$ case of (34), for $\lambda = 0$ the representation (34) is still valid; that is, in Proposition 0.0.7, the result is valid for even $\lambda = 0$. Hence, we can consider the results for $\lambda = 0$ in the spirit of Tikhonov regularization in which we are interested in a small λ or λ tending to zero. That is, when we use the Paley-Wiener space $W\left(\frac{\pi}{h}\right)$, we need not to consider the Tikhonov regularization. Then,

$$(L_t F_{t,0,h,g}^*)(x) = (g(\cdot), K_h(\cdot, x))_{L_2(\mathbf{R}^n)}$$

as we see from (32). Since the output is the orthogonal projection of g onto the Paley-Wiener space $W\left(\frac{\pi}{h}\right)$, we can estimate the difference of the output of our inverse $F_{t,0,h,g}^*$ and g , clearly as

$$\|L_t F_{t,0,h,g}^* - g\|_{L_2(\mathbf{R}^n)}$$

which is the distance from g onto the Paley-Wiener space $W\left(\frac{\pi}{h}\right)$. Of course, $F_{t,0,h,g}^*$ is the Moore-Penrose generalized inverse of the operator equation, for any $g \in L_2(\mathbf{R}^n)$ and $F \in W\left(\frac{\pi}{h}\right)$,

$$L_t F = g.$$

For the Paley-Wiener space $W\left(\frac{\pi}{h}\right)$, we need not use Tikhonov regularization and we can look for the Moore-Penrose generalized inverse $F_{t,0,h,g}^*$ by using the theory of reproducing kernels ([32], pp. 178-180). However, we had better to calculate the extremal functions $F_{t,\lambda,h,g}^*$ in the Tikhonov regularization and to set $\lambda = 0$, because the structure of the Moore-Penrose generalized inverses is involved.

We consider the heat conduction for the RKHS H_K , however, our inversion formula in the sense (33) will show that for a very general function containing the delta function, our inversion formula is valid, because we are considering the approximate inversion by the functions H_K .

Meanwhile, we are interested in the convergence property of the extremal functions (34) when h tends to zero for a general g .

However, we can see directly that in (34)

$$\lim_{h \rightarrow 0} (L_t F_{t,0,h,g}^*)(x) = g(x)$$

at the points x where $g(x)$ is continuous.

We are interested in numerical experiments for the both cases of the Sobolev space H_S and the Paley-Wiener space $W\left(\frac{\pi}{h}\right)$ for their convergences as $\lambda \rightarrow 0$ and $h \rightarrow 0$. For the Sobolev spaces and $\lambda \rightarrow 0$, see [23].

The real inversion formula (34) will give a practical formula for the Gaussian convolution. In [25], we have shown experimental results by using computers in Figures. In [23], we stated that when we use the Sobolev spaces as the reproducing kernels, *we will see that in order to overcome the high "ill-posedness" in the real inversion and in order to catch "analyticity" of the image of (30) we must work hardly; that is, we must take a very small λ and we must calculate the corresponding integral (34) hardly in the sense of numerical. Computers help us this hard work to calculate the integral for a very small λ .* However, when we use the Paley-Wiener spaces for $\lambda = 0$ and for h tending to 0, we can obtain a very simple and improved inversion formula.

8. NUMERICAL REAL INVERSION FORMULAS OF THE LAPLACE TRANSFORM

We shall give a very natural and numerical real inversion formula of the Laplace transform

$$(\mathcal{L}F)(p) = f(p) = \int_0^\infty e^{-pt} F(t) dt, \quad p > 0 \quad (36)$$

for functions F of some natural function space. This integral transform is, of course, very fundamental in mathematical science. The inversion of the Laplace

transform is, in general, given by a complex form, however, we are interested in and are requested to obtain its real inversion in many practical problems. However, the real inversion will be very involved and one might think that its real inversion will be essentially involved, because we must catch "analyticity" from the real or discrete data. Note that the image functions of the Laplace transform are analytic on some half complex plane. For complexity of the real inversion formula of the Laplace transform, we recall, for example, the following formulas:

$$\lim_{n \rightarrow \infty} \frac{(-1)^n}{n!} \left(\frac{n}{t}\right)^{n+1} f^{(n)}\left(\frac{n}{t}\right) = F(t)$$

and

$$\lim_{n \rightarrow \infty} \prod_{k=1}^n \left(1 + \frac{t}{k} \frac{d}{dt}\right) \left[\frac{n}{t} f\left(\frac{n}{t}\right)\right] = F(t),$$

([7,8]).

Furthermore, see [4,37,48,58] and the recent related articles [17] and [18]. See also the great references [59,60]. The problem may be related to analytic extension problems, see [44] and [48]. Here, we shall introduce new type and very natural real inversion formulas from the viewpoints of best approximations, generalized inverses and the Tikhonov regularization by combining these fundamental ideas and methods by means of the theory of reproducing kernels.

In order to apply the general theory in Section 2 to the real inversion formula of the Laplace transform, we shall recall the "natural situation" based on [27].

We shall introduce the simple reproducing kernel Hilbert space (RKHS) H_K comprised of absolutely continuous functions F on the positive real line $\mathbf{R}^+$ with finite norms

$$\left\{ \int_0^\infty |F'(t)|^2 \frac{1}{t} e^t dt \right\}^{1/2}$$

and satisfying $F(0) = 0$. This Hilbert space admits the reproducing kernel $K(t, t')$

$$\begin{aligned} K(t, t') &= \int_0^{\min(t, t')} \xi e^{-\xi} d\xi \\ &= \begin{cases} -te^{-t} - e^{-t} + 1 & \text{for } t \leq t' \\ -t'e^{-t'} - e^{-t'} + 1 & \text{for } t \geq t' \end{cases} \end{aligned} \quad (37)$$

(see [37], pages 55-56). Then we see that

$$\int_0^\infty |(\mathcal{L}F)(p)p|^2 dp \leq \frac{1}{2} \|F\|_{H_K}^2; \quad (38)$$

that is, the linear operator

$$(\mathcal{L}F)(p)p$$

on H_K into $L_2(\mathbf{R}^+, dp) = L_2(\mathbf{R}^+)$ is bounded. For the reproducing kernel Hilbert spaces H_K satisfying (38), we can find some general spaces ([27]). Therefore, from the general theory in Section 2, we obtain

Proposition 0.0.8 ([27]). *For any $g \in L_2(\mathbf{R}^+)$ and for any $\alpha > 0$, the best approximation $F_{\alpha,g}^*$ in the sense*

$$\begin{aligned} & \inf_{F \in H_K} \left\{ \alpha \int_0^\infty |F'(t)|^2 \frac{1}{t} e^t dt + \|(\mathcal{L}F)(p)p - g\|_{L_2(\mathbf{R}^+)}^2 \right\} \\ &= \alpha \int_0^\infty |F_{\alpha,g}^{*'}(t)|^2 \frac{1}{t} e^t dt + \|(\mathcal{L}F_{\alpha,g}^*)(p)p - g\|_{L_2(\mathbf{R}^+)}^2 \end{aligned} \quad (39)$$

exists uniquely and we obtain the representation

$$F_{\alpha,g}^*(t) = \int_0^\infty g(\xi) (\mathcal{L}K_\alpha(\cdot, t))(\xi) \xi d\xi. \quad (40)$$

Here, $K_\alpha(\cdot, t)$ is determined by the functional equation

$$K_\alpha(t, t') = \frac{1}{\alpha} K(t, t') - \frac{1}{\alpha} ((\mathcal{L}K_{\alpha,t'})(p)p, (\mathcal{L}K_t)(p)p)_{L_2(\mathbf{R}^{n+})} \quad (41)$$

for

$$K_{\alpha,t'} = K_\alpha(\cdot, t')$$

and

$$K_t = K(\cdot, t).$$

We shall look for the approximate inversion $F_{\alpha,g}^*(t)$ by using (40). For this purpose, we take the Laplace transform of (41) in t and change the variables t and t' as in

$$\begin{aligned} & (\mathcal{L}K_\alpha(\cdot, t))(\xi) \\ &= \frac{1}{\alpha} (\mathcal{L}K(\cdot, t'))(\xi) - \frac{1}{\alpha} ((\mathcal{L}K_{\alpha,t'})(p)p, (\mathcal{L}(\mathcal{L}K_t)(p)p))(\xi))_{L_2(\mathbf{R}^{n+})}. \end{aligned} \quad (42)$$

Note that

$$\begin{aligned} & (\mathcal{L}K(\cdot, t'))(p) \\ &= e^{-t'p} e^{-t'} \left[\frac{-t'}{p(p+1)} + \frac{-1}{p(p+1)^2} \right] + \frac{1}{p(p+1)^2} \end{aligned} \quad (43)$$

and

$$\int_0^\infty e^{-qt'} (\mathcal{L}K(\cdot, t'))(p) dt' = \frac{1}{pq(p+q+1)^2}. \quad (44)$$

Therefore, by setting

$$(\mathcal{L}K_\alpha(\cdot, t))(\xi)\xi = H_\alpha(\xi, t),$$

which is needed in (40), we obtain the Fredholm integral equation of the second type

$$\begin{aligned} \alpha H_\alpha(\xi, t) + \int_0^\infty H_\alpha(p, t) \frac{1}{(p + \xi + 1)^2} dp \\ = -\frac{e^{-t\xi}e^{-t}}{\xi + 1} \left(t + \frac{1}{\xi + 1} \right) + \frac{1}{(\xi + 1)^2}. \end{aligned} \quad (45)$$

We solved this integral equation by the three methods:

In [27], by using Matlab with span 0.01 over $[0, 1]$, we were able to obtain quite good numeral solutions for typical examples.

In [28], by using the sinc method, we were able to obtain much more good results for typical examples.

In [9], Professor Fujiwara were able to obtain a definite numerical result for the solution of (45) by his method; infinite precision method (cf. [14]); and indeed, he solved (45) with 600 digits with $\alpha = 10^{-400}$!!

Surprisingly enough, for even the delta function $F(t) = \delta(t - 1)$, he was able to obtain a good numerical result by his computer algorithm.

In [11], Sawano proved that the operator in (38) is, in fact, compact and Professor Fujiwara gave very interesting results for (39) by using his spectral theory and infinite precision methods. He were able to calculate 1500 terms in eigenvalues and eigenfunction expansions, precisely!! However, the numerical results examined are not better than ones in [9], at this moment.

We may think that the real inversion formula and computer program in [9] is practical and effective as a numerical real inversion of the Laplace transform.

REFERENCES

1. J. Agler and J. E. McCarthy, Pick Interpolation and Hilbert Function Spaces, Graduate Studies in Mathematics, Vol. **44** (2002).
2. N. Aronszajn, Theory of reproducing kernels, *Trans. Amer. Math. Soc.*, **68** (1950), 337–404.
3. M. Asaduzzaman, T. Matsuura, and S. Saitoh, Constructions of approximate solutions for linear differential equations by reproducing kernels and inverse problems, *Advances in Analysis, Proceedings of the 4th International ISAAC Congress* (2005), 335–344. (World Scientific).
4. D.-W. Byun and S. Saitoh, A real inversion formula for the Laplace transform *Z. Anal. Anw.*, **12**(1993), 597–603.
5. D.-W. Byun and S. Saitoh, Approximation by the solutions of the heat equation, *J. Approximation Theory*, **78** (1994), 226–238.

6. D.-W. Byun and S. Saitoh, Best approximation in reproducing kernel Hilbert spaces, *Proc. of the 2th International Colloquium on Numerical Analysis*, VSP-Holland, (1994), 55–61.
7. G. Doetsch, *Handbuch der Laplace Transformation*, Vol. 1., Birkhäuser Verlag, Basel, 1950.
8. H. W. Engl, M. Hanke and A. Neubauer, *Regularization of Inverse Problems, Mathematics and Its Applications*, **376**(2000), Kluwer Academic Publishers.
9. H. Fujiwara, T. Matsuura and S. Saitoh, Numerical real inversion formulas of the Laplace transform by using a Fredholm integral equation of the second kind, (in preparation).
10. H. Fujiwara, T. Matsuura, S. Saitoh and Y. Sawano, The real inversion of the Laplace transform by numerical singular value decomposition, (in preparation).
11. C. W. Groetsch, *Inverse Problems in the Mathematical Sciences*, Vieweg & Sohn Verlagsgesellschaft mbH, Braunschweig/Wiesbaden (1993).
12. N. Hayashi, Analytic function spaces and their applications to nonlinear evolution equations, *Analytic Extension Formulas and their Applications*, (2001), Kluwer Academic Publishers, 59–86.
13. J. R. Higgins, A sampling principle associated with Saitoh's fundamental theory of linear transformations, *Analytic Extension Formulas and their Applications*, (2001), Kluwer Academic Publishers, 73–86.
14. H. Imai, T. Takeuchi and M. Kushida, On numerical simulation of partial differential equations in infinite precision, *Advances in Mathematical Sciences and Applications*, **9**(1999), 1007-1016.
15. H. Itou and S. Saitoh, Analytical and numerical solutions of linear singular integral equations, *Int. J. Appl. Math. Stat.*, **12**(2007), 76-89.
16. I.I. Hirschman and D.V. Widder, *The Convolution Transform*, Princeton University Press, Princeton, New Jersey, (1955).
17. V.V. Kryzhniy, Regularized inversion of integral transformations of Mellin convolution type, *Inverse Problems*, **19**(2003), 573-583.
18. V. V. Kryzhniy, Numerical inversion of the Laplace transform: analysis via regularized analytic continuation, *Inverse Problems*, **22** (2006), 579-597.
19. T. Matsuura and S. Saitoh, Analytical and Numerical Solutions of the Inhomogeneous Wave Equation, *Australian J. of Math. Anal. and Appl.*, **1**(2004), Volume 1, Issue 1, Article 7.
20. T. Matsuura, S. Saitoh and D.D. Trong, Numerical solutions of the Poisson equation, *Applicable Analysis*, **83**(2004), 1037-1051.
21. T. Matsuura and S. Saitoh, Analytical and numerical solutions of linear ordinary differential equations with constant coefficients, *Journal Analysis and Applications*, **3**(2005), 1-17.
22. T. Matsuura and S. Saitoh, Numerical inversion formulas in the wave equation, *Journal of Computational Mathematics and Optimization*, **1**(2005), 1-19.
23. T. Matsuura, S. Saitoh and D.D. Trong, Approximate and analytical inversion formulas in heat conduction on multidimensional spaces, *J. of Inverse and Ill-posed Problems*, **13** (2005), 479-493.

24. T. Matsuura and S. Saitoh, Dirichlet's Principle Using Computers, *Applicable Analysis*, **84**(2005), 989-1003.
25. T. Matsuura and S. Saitoh, Analytical and numerical inversion formulas in the Gaussian convolution by using the Paley-Wiener spaces, *Applicable Analysis*, **85**(2006), 901-915.
26. T. Matsuura and S. Saitoh, Analytical and numerical real inversion formulas of the Laplace transform, *The ISAAC Catani Congress Proceedings* (to appear).
27. T. Matsuura, A. Al-Shuaibi, H. Fujiwara and S. Saitoh, Numerical real inversion formulas of the Laplace transform by using a Fredholm integral equation of the second kind, *Journal Analysis and Applications*, *J. of Analysis and Applications*, **5**(2007), 123-136.
28. T. Matsuura, A. Al-Shuaibi, H. Fujiwara, S. Saitoh and M. Sugihara, Numerical real inversion formulas of the Laplace transform by a sinc method, (submitted).
29. Th. M. Rassias and S. Saitoh, The Pythagorean theorem and linear mappings, *PanAmerican Math. J.*, **12** (2002), 1-10.
30. S. Saitoh, Hilbert spaces induced by Hilbert space valued functions, *Proc. Amer. Math. Soc.*, **89** (1983), 74-78.
31. S. Saitoh, The Weierstrass transform and an isometry in the heat equation, *Applicable Analysis*, **16**(1983), 1-6.
32. S. Saitoh, Theory of Reproducing Kernels and its Applications, *Pitman Research Notes in Mathematics Series*, **189** (1988), Longman Scientific & Technical, UK.
33. S. Saitoh, Interpolation problems of Pick-Nevanlinna type, *Pitman Research Notes in Mathematics Series*, **212** (1989), 253-262.
34. S. Saitoh, One approach to some general integral transforms and its applications, *Integral Transforms and Special Functions*, **3** (1995), 49-84.
35. S. Saitoh, Natural norm inequalities in nonlinear transforms, *General Inequalities* **7**(1997), 39-52. Birkhäuser Verlag, Basel, Boston.
36. S. Saitoh, Representations of inverse functions, *Proc. Amer. Math. Soc.*, **125** (1997), 3633-3639.
37. S. Saitoh, Integral Transforms, Reproducing Kernels and their Applications, *Pitman Research Notes in Mathematics Series*, **369** (1997), Addison Wesley Longman, UK.
38. S. Saitoh, Nonlinear transforms and analyticity of functions, *Nonlinear Mathematical Analysis and Applications*, (1998), 223-234. Hadronic Press, Palm Harbor.
39. S. Saitoh, Various operators in Hilbert space induced by transforms, *International J. of Applied Math.*, **1** (1999), 111-126.
40. S. Saitoh, Applications of the general theory of reproducing kernels, *Reproducing Kernels and their Applications*, (1999), Kluwer Academic Publishers, 165-188.
41. S. Saitoh, D. Alpay, J.A. Ball and T. Ohsawa (eds), *Reproducing Kernels and their Applications*, (1999), Kluwer Academic Publishers.
42. S. Saitoh and M. Yamamoto, Integral transforms involving smooth functions, *Reproducing Kernels and their Applications*, (1999), Kluwer Academic Publishers, 149-164.

43. S. Saitoh, Linear integro-differential equations and the theory of reproducing kernels, *Volterra Equations and Applications*, C. Corduneanu and I.W. Sandberg (eds), Gordon and Breach Science Publishers (2000), Amsterdam.
44. S. Saitoh, Analytic extension formulas, integral transforms and reproducing kernels, *Analytic Extension Formulas and their Applications*, (2001), Kluwer Academic Publishers, 207–232.
45. S. Saitoh, Applications of the reproducing kernel theory to inverse problems, *Comm. Korean Math. Soc.*, **16** (2001), 371–383.
46. S. Saitoh, Principle of telethroscope, *Functional-Analytic and Complex Methods, their Interaction and Applications to Partial Differential Equations*, – *Proceedings of the International Graz Workshop*, Graz, Austria, 12-16 February 2001. World Scientific (2001), 101–117.
47. S. Saitoh, N. Hayashi and M. Yamamoto (eds.), *Analytic Extension Formulas and their Applications*, (2001), Kluwer Academic Publishers.
48. S. Saitoh, Vu Kim Tuan and M. Yamamoto, Conditional Stability of a Real Inverse Formula for the Laplace Transform, *Z. Anal. Anw.*, **20**(2001), 193-202.
49. S. Saitoh, Vu Kim Tuan and M. Yamamoto, Reverse convolution inequalities and applications to inverse heat source problems, *J. of Inequalities in Pure and Applied Mathematics*, **3** (2002), Article 80.
50. S. Saitoh, T. Matsuura and M. Asaduzzaman, Operator Equations and Best Approximation Problems in Reproducing Kernel Hilbert Spaces, *Journal of Analysis and Applications*, **1**(2003), 131-142.
51. S. Saitoh, Constructions by Reproducing Kernels of Approximate Solutions for Linear Differential Equations with L_2 Integrable Coefficients, *International J. of Math. Sci.*, **2**(2003), 261-273.
52. S. Saitoh, Approximate Real Inversion Formulas of the Gaussian Convolution, *Applicable Analysis*, **83**(2004), 727-733.
53. S. Saitoh, Best approximation, Tikhonov regularization and reproducing kernels, *Kodai. Math. J.*, **28**(2005), 359-367.
54. S. Saitoh, Tikhonov regularization and the theory of reproducing kernels, *Finite or Infinite Dimensional Complex Analysis and Applications (Proceedings of the 12th ICFIDCAA)*, Kyushu University Press (2005), 291-298.
55. S. Saitoh, Applications of reproducing kernels to best approximations, Tikhonov regularizations and inverse problems, *Advances in Analysis, Proceedings of the 4th International ISAAC Congress (2005)*, (World Scientific), 439-445.
56. F. Stenger, *Numerical Methods Based on Sinc and Analytic Functions*, Springer Series in Computational Mathematics, **20**, 1993.
57. W. Ulmer and W. Kaissl, The inverse problem of a Gaussian convolution and its application to the finite size of measurement chambers/detectors in photon and proton dosimetry, *Phys. Med. Biol.*, **48**(2003), 707-727.
58. D. V. Widder, *The Laplace Transform*, Princeton University Press, Princeton, 1972.
59. <http://library.wolfram.com/inforcenter/MathSource/4738/>

60. *<http://www.columbia.edu/~ww2040/abate.html>*

SAITOH: GRADUATE SCHOOL OF ENGINEERING
GUNMA UNIVERSITY

KIRYU 376-8515, JAPAN

E-MAIL: SSAITOH@MATH.SCI.GUNMA-U.AC.JP; KBDMM360@YAHOO.CO.JP

ATOMIC DECOMPOSITION OF MORREY SPACES

YOSHIHIRO SAWANO

Abstract. The aim of this talk is to describe briefly decompositions in the Besov-Morrey space and the Triebel-Lizorkin-Morrey space. As a corollary we obtain the atomic decomposition for the (generic) Morrey space.

Key words and Phrases: Triebel-Lizorkin space, Morrey space, decomposition of functions

The Besov-Morrey space emerged originally in [3]. H. Kozono and M. Yamazaki investigated time-local solutions of the Navier-Stokes equations. Later it was investigated by A. Mazzucato. She investigated the atomic decomposition and the molecular decomposition [4, 5]. [3, 4, 5] develop a theory of the function space $\mathcal{N}_{pqr}^s$ with $1 \leq q \leq p < \infty$, $1 \leq r \leq \infty$ and $s \in \mathbb{R}$. L. Tang and J. Xu defined the function spaces $\mathcal{N}_{pqr}^s$ and $\mathcal{E}_{pqr}^s$ with $0 < q \leq p < \infty$, $1 \leq r \leq \infty$ and $s \in \mathbb{R}$ in [11]. The present authors developed a theory of decompositions in $\mathcal{N}_{pqr}^s$ and $\mathcal{E}_{pqr}^s$ with $0 < q \leq p < \infty$, $1 \leq r \leq \infty$ and $s \in \mathbb{R}$. This type of decomposition results will be of more importance because it provides us with a convenient way to analyze something. For example, the synthesis result covers a part of theory of wavelet analysis.

Before we go into the definitions of the Besov-Morrey spaces and the Triebel-Lizorkin-Morrey spaces, let us recall the definitions of the Besov spaces and the Triebel-Lizorkin spaces, which are prototypes of the Besov-Morrey spaces and the Triebel-Lizorkin-Morrey spaces respectively. To describe these function spaces, we fix some notations. $B(r)$ means the open ball centered at the origin of radius $r > 0$.

Let $\{f_j\}_{j \in \mathbb{N}_0}$ be a sequence of functions. Then define

$$\|f_j : l_q(L_p)\| := \left(\sum_{j \in \mathbb{N}_0} \|f_j : L_p\|^q \right)^{\frac{1}{q}}, \|f_j : L_p(l_q)\| := \left\| \left(\sum_{j \in \mathbb{N}_0} |f_j|^q \right)^{\frac{1}{q}} : L_p \right\|$$

for $0 < p, q \leq \infty$. Here a natural modification is made if $q = \infty$. Next we fix a sequence of smooth functions $\{\varphi_j\}_{j \in \mathbb{N}_0} \subset \mathcal{S}(\mathbb{R}^n)$ so that

$$\chi_{B(2)} \leq \varphi_0 \leq \chi_{B(4)}, \chi_{B(4) \setminus B(2)} \leq \varphi_1 \leq \chi_{B(8) \setminus B(4)}, \varphi_j = \varphi_1(2^{-j+1} \cdot)$$

for $j \in \mathbb{N}$. Given $f \in \mathcal{S}'(\mathbb{R}^n)$ and $\tau \in \mathcal{S}(\mathbb{R}^n)$, we define $\tau(D)f := \mathcal{F}^{-1}(\tau \cdot \mathcal{F}f)$.

Under these notations, we define the Besov norm and the Triebel-Lizorkin norm. Let $f \in \mathcal{S}'(\mathbb{R}^n)$. We define

$$\begin{aligned} \|f : B_{pq}^s(\mathbb{R}^n)\| &:= \|2^{js} \phi_j(D)f : l_q(L_p)\|, \quad 0 < p \leq \infty, 0 < q \leq \infty, s \in \mathbb{R} \\ \|f : F_{pq}^s(\mathbb{R}^n)\| &:= \|2^{js} \phi_j(D)f : L_p(l_q)\|, \quad 0 < p < \infty, 0 < q \leq \infty, s \in \mathbb{R}. \end{aligned}$$

Different admissible choices of ϕ_0 and ϕ_1 will yield equivalent quasi-norms. To unify the formulation in the sequel, as was defined in [15], for example, we use $A_{pq}^s(\mathbb{R}^n)$ to denote either $B_{pq}^s(\mathbb{R}^n)$ or $F_{pq}^s(\mathbb{R}^n)$ with $0 < p \leq \infty, 0 < q \leq \infty$ and $s \in \mathbb{R}$. If $A_{pq}^s(\mathbb{R}^n)$ means $F_{pq}^s(\mathbb{R}^n)$, then the case when $p = \infty$ and $q < \infty$ is excluded automatically.

Let us describe the function spaces $\mathcal{N}_{pqr}^s$ and $\mathcal{E}_{pqr}^s$ briefly. Suppose that the parameters p, q, r, s satisfy

$$0 < q \leq p < \infty, 0 < r \leq \infty, s \in \mathbb{R}.$$

Define the Morrey norm of a measurable function f by

$$\|f : \mathcal{M}_q^p\| := \sup_{x \in \mathbb{R}^n, r > 0} r^{\frac{n}{p} - \frac{n}{q}} \left(\int_{B(x, r)} |f(y)|^q dy \right)^{\frac{1}{q}}. \quad (1)$$

p seems to reflect the global integrability while q describes the local integrability. Below it will be understood that $\mathcal{M}_q^\infty = L_\infty$.

The Besov-Morrey spaces and the Triebel-Lizorkin-Morrey spaces are obtained by replacing the L_p norm with the Morrey norm $\|\cdot : \mathcal{M}_q^p\|$ given by (1). Given a sequence of measurable functions $\{f_j\}_{j \in \mathbb{N}_0}$, we define

$$\begin{aligned} \|f_j : l_r(\mathcal{M}_q^p)\| &:= \left(\sum_{j \in \mathbb{N}_0} \|f_j : \mathcal{M}_q^p\|^r \right)^{\frac{1}{r}}, \quad 0 < q \leq p \leq \infty, 0 < r \leq \infty \\ \|f_j : \mathcal{M}_q^p(l_r)\| &:= \left\| \left(\sum_{j \in \mathbb{N}_0} |f_j|^r \right)^{\frac{1}{r}} : \mathcal{M}_q^p \right\|, \quad 0 < q \leq p < \infty, 0 < r \leq \infty. \end{aligned}$$

Definition 0.0.1. Let $f \in \mathcal{S}'(\mathbb{R}^n)$. Then define

$$\begin{aligned}\|f : \mathcal{N}_{pqr}^s(\mathbb{R}^n)\| &:= \|2^{js}\phi_j(D)f : l_r(\mathcal{M}_q^p)\|, \quad 0 < q \leq p \leq \infty, \quad 0 < r \leq \infty, \quad s \in \mathbb{R} \\ \|f : \mathcal{E}_{pqr}^s(\mathbb{R}^n)\| &:= \|2^{js}\phi_j(D)f : \mathcal{M}_q^p(l_r)\|, \quad 0 < q \leq p < \infty, \quad 0 < r \leq \infty, \quad s \in \mathbb{R}.\end{aligned}$$

The function space $\mathcal{A}_{pqr}^s(\mathbb{R}^n)$ covers many families of function spaces such as the Hölder-Zygmund space $\mathcal{C}^s(\mathbb{R}^n)$, the Morrey space $\mathcal{M}_q^p(\mathbb{R}^n)$, the Sobolev-Morrey space, the Besov space $B_{pq}^s(\mathbb{R}^n)$ and the Triebel-Lizorkin space $F_{pq}^s(\mathbb{R}^n)$. As for the Sobolev-Morrey space, we refer to [6, 7, 8]. Recall that the Hölder space $\mathcal{C}^s(\mathbb{R}^n)$, $0 < s < 1$ is a set of all continuous functions normed by

$$\|f : \mathcal{C}^s(\mathbb{R}^n)\| := \|f : L_\infty\| + \sup_{\substack{x, y \in \mathbb{R}^n \\ x \neq y}} \frac{|f(x) - f(y)|}{|x - y|^s}.$$

Proposition 0.0.2. Suppose that $0 < q \leq p \leq \infty$, $0 < r \leq \infty$ and $s \in \mathbb{R}$.

1. Let $k \in \mathbb{N}$. Then $f \in \mathcal{A}_{pqr}^s(\mathbb{R}^n)$ if and only if $f \in \mathcal{A}_{pqr}^{s-k}(\mathbb{R}^n)$ and $\partial_j^k f \in \mathcal{A}_{pqr}^{s-k}(\mathbb{R}^n)$ for every $j = 1, 2, \dots, n$. Furthermore, we have the following norm equivalence

$$\|f : \mathcal{A}_{pqr}^s(\mathbb{R}^n)\| \simeq \|f : \mathcal{A}_{pqr}^{s-k}(\mathbb{R}^n)\| + \sum_{j=1}^n \|\partial_j^k f : \mathcal{A}_{pqr}^{s-k}(\mathbb{R}^n)\|$$

for $f \in \mathcal{A}_{pqr}^s(\mathbb{R}^n)$.

2. $\mathcal{A}_{ppr}^s(\mathbb{R}^n) = \mathcal{A}_{pr}^s(\mathbb{R}^n)$.
3. $\mathcal{E}_{pq2}^0(\mathbb{R}^n) = \mathcal{M}_q^p(\mathbb{R}^n)$, if $1 < q \leq p < \infty$.
4. $\mathcal{N}_{\infty\infty\infty}^s(\mathbb{R}^n) = B_{\infty\infty}^s(\mathbb{R}^n) = \mathcal{C}^s(\mathbb{R}^n)$, if $s \in (0, 1)$.

2. ATOMIC DECOMPOSITION

To describe the atomic decomposition, we introduce some notations, which are based on those in [14, 15]. Now we follow [10, Section 4] to formulate the atomic decomposition.

Notation . 1. Let $\nu \in \mathbb{Z}$ and $m \in \mathbb{Z}^n$. Then define $Q_{\nu m} := \prod_{j=1}^n \left[\frac{m_j}{2^\nu}, \frac{m_j + 1}{2^\nu} \right]$.

Call $Q_{\nu m}$ a dyadic cube in $\mathbb{R}^n$ for each $\nu \in \mathbb{Z}$ and $m \in \mathbb{Z}^n$.

2. Let $0 < p < \infty$, $\nu \in \mathbb{N}_0$ and $m \in \mathbb{Z}^n$. $\chi_{\nu m}^{(p)}$ denotes the p -normalized indicator given by

$$\chi_{\nu m}^{(p)} := 2^{\frac{\nu n}{p}} \chi_{Q_{\nu m}}.$$

3. Given a doubly indexed sequence $\lambda = \{\lambda_{\nu m}\}_{\nu \in \mathbb{N}_0, m \in \mathbb{Z}^n}$, define

$$\begin{aligned} \|\lambda : \mathbf{n}_{pqr}(\mathbb{R}^n)\| &:= \left\| \left\{ \sum_{m \in \mathbb{Z}^n} \lambda_{\nu m} \chi_{\nu m}^{(p)} \right\}_{\nu \in \mathbb{N}_0} : l_r(\mathcal{M}_q^p) \right\|, \quad 0 < q \leq p \leq \infty, \quad 0 < r \leq \infty \\ \|\lambda : \mathbf{e}_{pqr}(\mathbb{R}^n)\| &:= \left\| \left\{ \sum_{m \in \mathbb{Z}^n} \lambda_{\nu m} \chi_{\nu m}^{(p)} \right\}_{\nu \in \mathbb{N}_0} : \mathcal{M}_q^p(l_r) \right\|, \quad 0 < q \leq p \leq \infty, \quad 0 < r \leq \infty. \end{aligned}$$

In order to unify the formulations in the sequel, denote by $\mathbf{a}_{pqr}(\mathbb{R}^n)$ either $\mathbf{n}_{pqr}(\mathbb{R}^n)$ or $\mathbf{e}_{pqr}(\mathbb{R}^n)$. We rule out $\mathbf{a}_{pqr}(\mathbb{R}^n)$ with $p = \infty$ and $r < \infty$. Following [14], let us recall the definition of atoms.

Definition 0.0.3. Let $0 < p < \infty$ and $s \in \mathbb{R}$. Fix $K \in \mathbb{N}_0$, $L \in \mathbb{Z} \cap [-1, \infty)$ and $d > 1$.

1. Suppose further that $\nu \in \mathbb{N}_0$ and $m \in \mathbb{Z}^n$. A C^K -function a is said to be an atom centered at $Q_{\nu m}$, if it is supported on $dQ_{\nu m}$ and satisfies the differential inequality and the moment condition given below :

$$\|\partial^\alpha a : L_\infty\| \leq 2^{-\nu(s - \frac{n}{p}) + \nu|\alpha|} \text{ for } \alpha \in \mathbb{N}_0^n \text{ with } |\alpha| \leq K \quad (2)$$

and

$$\int_{\mathbb{R}^n} x^\beta a(x) dx = 0 \text{ for } \beta \in \mathbb{N}_0^n, \nu \in \mathbb{N}_0 \text{ with } |\beta| \leq L, \nu \geq 1. \quad (3)$$

Here condition (3) means no condition, if $L = -1$.

2. Define

$$\begin{aligned} \mathbf{Atom}_0 &:= \{ \{a_{\nu m}\}_{\nu \in \mathbb{N}_0, m \in \mathbb{Z}^n} : \text{each } a_{\nu m} \text{ is an atom centered at } Q_{\nu m} \}, \\ \mathbf{Atom} &:= \{ \{a_{\nu m}\}_{\nu \in \mathbb{N}_0, m \in \mathbb{Z}^n} : \{c a_{\nu m}\}_{\nu \in \mathbb{N}_0, m \in \mathbb{Z}^n} \in \mathbf{Atom}_0 \text{ for some } c > 0 \}. \end{aligned}$$

Definition 0.0.4 (Molecule). A C^K -function m is called an (s, p) -molecule, $s \in \mathbb{R}$, $0 < u \leq p \leq \infty$, $0 < q \leq \infty$, if it satisfies the following oscillation and decay conditions hold for some point $x_0 \in \mathbb{R}^n$ and $\nu \in \mathbb{Z}$, where M is a constant sufficiently large :

1. $\int_{\mathbb{R}^n} x^\alpha m(x) dx = 0$ for $|\alpha| \leq L$.
2. $|\partial^\alpha m(x)| \leq 2^{-\nu(s - n/p) + \nu|\alpha|} \langle 2^\nu(x - x_0) \rangle^{-M - |\alpha|}$ if $|\alpha| \leq K$.

Here and below we call m molecule centered at x_0 and we always assume that

$$M \geq L + \frac{10n}{\min(1, q, u)}.$$

With this definition in mind, we formulate our atomic decomposition theorem.

Theorem 0.0.5 (Atomic decomposition). Suppose that the parameters $K, L \in \mathbb{Z}$ and $p, q, r, s, d \in \mathbb{R}$ satisfy

$$0 < q \leq p \leq \infty, 0 < r \leq \infty, d > 1, K \geq (1 + [s])_+, L \geq \max(-1, [\sigma_q - s])$$

for the $\mathcal{N}$ -scale and

$$0 < q \leq p < \infty, 0 < r \leq \infty, d > 1, K \geq (1 + [s])_+, L \geq \max(-1, [\sigma_{qr} - s])$$

for the $\mathcal{E}$ -scale.

1. Assume that $\{a_{\nu m}\}_{\nu \in \mathbb{N}_0, m \in \mathbb{Z}^n} \in \mathbf{Atom}_0$ and $\lambda = \{\lambda_{\nu m}\}_{\nu \in \mathbb{N}_0, m \in \mathbb{Z}^n} \in \mathbf{a}_{pqr}(\mathbb{R}^n)$. Then the sum

$$f := \sum_{\nu \in \mathbb{N}_0} \sum_{m \in \mathbb{Z}^n} \lambda_{\nu m} a_{\nu m}$$

converges in $\mathcal{S}'(\mathbb{R}^n)$ and belongs to $\mathcal{A}_{pqr}^s(\mathbb{R}^n)$ with the norm estimate

$$\|f : \mathcal{A}_{pqr}^s(\mathbb{R}^n)\| \leq c \|\lambda : \mathbf{a}_{pqr}(\mathbb{R}^n)\|.$$

Here the constant c does not depend on $\{a_{\nu m}\}_{\nu \in \mathbb{N}_0, m \in \mathbb{Z}^n}$ nor λ . The same can be said if we replace atoms with molecules.

2. Conversely any $f \in \mathcal{A}_{pqr}^s(\mathbb{R}^n)$ admits the following decomposition:

$$f := \sum_{\nu \in \mathbb{N}_0} \sum_{m \in \mathbb{Z}^n} \lambda_{\nu m} a_{\nu m}.$$

The sum converges in $\mathcal{S}'(\mathbb{R}^n)$. We can arrange that $\{a_{\nu m}\}_{\nu \in \mathbb{N}_0, m \in \mathbb{Z}^n} \in \mathbf{Atom}_0$ and the coefficient $\lambda = \{\lambda_{\nu m}\}_{\nu \in \mathbb{N}_0, m \in \mathbb{Z}^n} \in \mathbf{a}_{pqr}(\mathbb{R}^n)$ fulfill the norm estimate

$$\|\lambda : \mathbf{a}_{pqr}(\mathbb{R}^n)\| \leq c \|f : \mathcal{A}_{pqr}^s(\mathbb{R}^n)\|.$$

In view of our actual construction, unfortunately the coefficient λ does not depend linearly on f . We refer to [1, 10, 14] for more details.

2. QUARKONIAL DECOMPOSITION

As we have noted, the atomic decomposition fails to enjoy linear dependency for the constants. In [14, 15] H. Triebel proposed the quarkonial decomposition. The counterpart for $\mathcal{A}_{pqr}^s(\mathbb{R}^n)$ was obtained in [10, Section 5]. Following [15, Chapter 1, Section 1] and [10, Section 5], we describe the quarkonial decomposition for $\mathcal{A}_{pqr}^s(\mathbb{R}^n)$.

Notation . 1. Let $\psi \in \mathcal{S}(\mathbb{R}^n)$ be a compactly supported function that generates the partition of the unit :

$$\sum_{m \in \mathbb{Z}^n} \psi(x - m) \equiv 1$$

with $\text{supp}(\psi) \subset B(2^{\tilde{d}})$.

2. For $\beta \in \mathbb{N}_0$ define $\psi^\beta(x) := x^\beta \psi(x)$.
3. For $\beta \in \mathbb{N}_0$, $\nu \in \mathbb{N}_0$, $m \in \mathbb{Z}^n$, define $(\beta qu)_{\nu m}(x) := 2^{-\nu(s-\frac{n}{p})} \psi^\beta(2^\nu x - m)$.
4. Let $\rho > \tilde{d}$, where $\tilde{d}$ is a fixed constant appearing just above. Given a triply indexed sequence $\lambda = \{\lambda_{\nu m}^\beta\}_{\beta \in \mathbb{N}_0^n, \nu \in \mathbb{N}_0, m \in \mathbb{Z}^n}$, we define

$$\|\lambda : \mathbf{n}_{pqr}(\mathbb{R}^n)\|_\rho := \sup_{\beta \in \mathbb{N}_0^n} 2^{|\beta|} \|\lambda^\beta : \mathbf{n}_{pqr}(\mathbb{R}^n)\|, \quad 0 < q \leq p \leq \infty, \quad 0 < r \leq \infty$$

$$\|\lambda : \mathbf{e}_{pqr}(\mathbb{R}^n)\|_\rho := \sup_{\beta \in \mathbb{N}_0^n} 2^{|\beta|} \|\lambda^\beta : \mathbf{e}_{pqr}(\mathbb{R}^n)\|, \quad 0 < q \leq p \leq \infty, \quad 0 < r \leq \infty,$$

where $\lambda^\beta := \{\lambda_{\nu m}^\beta\}_{\nu \in \mathbb{N}_0, m \in \mathbb{Z}^n}$.

5. We denote by $\|\lambda : \mathbf{a}_{pqr}(\mathbb{R}^n)\|_\rho$ to denote either $\|\lambda : \mathbf{n}_{pqr}(\mathbb{R}^n)\|_\rho$ or $\|\lambda : \mathbf{e}_{pqr}(\mathbb{R}^n)\|_\rho$ as usual. $\mathbf{e}_{\infty qr}(\mathbb{R}^n)$ with $r < \infty$ will not appear for later considerations.

Theorem 0.0.6 (Quarkonial decomposition for the general case). *Suppose that an odd integer L and the parameters p, q, r, s, σ, ρ satisfy*

$$0 < q \leq p \leq \infty, \quad 0 < r \leq \infty, \quad s \in \mathbb{R}, \quad \sigma > \max(\sigma_q, s), \quad \rho > \tilde{d}, \quad L \geq \max(-1, [\sigma_q - s])$$

for the $\mathcal{N}$ -scale and

$$0 < q \leq p < \infty, \quad 0 < r \leq \infty, \quad s \in \mathbb{R}, \quad \sigma > \max(\sigma_{qr}, s), \quad \rho > \tilde{d}, \quad L \geq \max(-1, [\sigma_{qr} - s])$$

for the $\mathcal{E}$ -scale. Let $f \in \mathcal{S}'(\mathbb{R}^n)$. Set

$$\begin{aligned} (\beta qu)_{\nu m}(x) &:= 2^{-\nu(\sigma-\frac{n}{p})} \psi^\beta(2^\nu x - m) \\ (\beta qu)_{\nu m}^{(L)}(x) &:= 2^{-\nu(s-\frac{n}{p})} \left((-\Delta)^{\frac{L+1}{2}} \psi^\beta \right) (2^\nu x - m). \end{aligned}$$

Then $f \in \mathcal{A}_{pqr}^s(\mathbb{R}^n)$ if and only if there exist triply indexed sequences

$$\eta = \{\eta_{\nu m}^\beta\}_{\beta \in \mathbb{N}_0^n, \nu \in \mathbb{N}_0, m \in \mathbb{Z}^n} \quad \text{and} \quad \lambda = \{\lambda_{\nu m}^\beta\}_{\beta \in \mathbb{N}_0^n, \nu \in \mathbb{N}_0, m \in \mathbb{Z}^n} \in \mathbf{a}_{pqr}(\mathbb{R}^n)$$

such that f can be expressed as

$$f = \sum_{\beta \in \mathbb{N}_0^n} \sum_{\nu \in \mathbb{N}_0} \sum_{m \in \mathbb{Z}^n} \eta_{\nu m}^\beta (\beta qu)_{\nu m} + \sum_{\beta \in \mathbb{N}_0^n} \sum_{\nu \in \mathbb{N}_0} \sum_{m \in \mathbb{Z}^n} \lambda_{\nu m}^\beta (\beta qu)_{\nu m}^{(L)}$$

with

$$\|\eta : \mathbf{a}_{pqr}(\mathbb{R}^n)\|_\rho + \|\lambda : \mathbf{a}_{pqr}(\mathbb{R}^n)\|_\rho < \infty. \quad (4)$$

If this is the case, then λ can be taken so that it depends linearly and continuously on f and the following norm equivalence holds:

$$\|\eta : \mathbf{a}_{pqr}(\mathbb{R}^n)\|_\rho + \|\lambda : \mathbf{a}_{pqr}(\mathbb{R}^n)\|_\rho \simeq \|f : \mathcal{A}_{pqr}^s(\mathbb{R}^n)\|. \quad (5)$$

If s is sufficiently large, then we can omit the second term and let $\sigma = s$.

We remark that the quarkonial decomposition turns out to be helpful when we are faced the constructions of operators. One of the reasons of this advantage is that the coefficient can be determined canonically. For more details we refer to [15].

Acknowledgement. The author is supported by Research Fellowships of the Japan Society for the Promotion of Science for Young Scientists.

REFERENCES

1. M. Frazier and B. Jawerth, A discrete transform and decompositions of distribution spaces. *Indiana University Math. J.* **34** (1985), no. 4, 777–799.
2. M. Frazier and B. Jawerth, A discrete transform and decompositions of distribution spaces. *J. Funct. Anal.* **93** (1990), no. 1, 34–170.
3. H. Kozono and M. Yamazaki, Semilinear heat equations and the Navier-Stokes equations with distributions in new function spaces as initial data, *Comm. Partial Differential Equations* **19** (1994) no. 5–6, 959–1014.
4. A. Mazzucato, Decomposition of Besov-Morrey spaces. *Harmonic analysis at Mount Holyoke (South Hadley, MA, 2001)*, 279–294, *Contemp. Math.*, **320**, Amer. Math. Soc., Providence, RI, 2003.
5. A. Mazzucato, Besov-Morrey spaces: Function space theory and applications to non-linear PDE, *Trans. Amer. Math. Society*, **355** (2003), no. 4, 1297–1364.
6. A. M. Najafov, Some properties of functions from the intersection of Besov-Morrey type spaces with dominant mixed derivatives. *Proc. A. Razmadze Math. Inst.* **139** (2005), 71–82.
7. A. M. Najafov, Embedding theorems in the Sobolev-Morrey type spaces $S_{p,a,\kappa,r}^l W(G)$ with dominant mixed derivatives, *Siberian Mathematical Journal*, **47** (2006), no 3, 613–625.
8. A. M. Najafov, On some properties of the functions from Sobolev-Morrey type spaces. *Cent. Eur. J. Math.* **3** (2005), no. 3, 496–507
9. Y. Sawano and H. Tanaka, Morrey spaces for non-doubling measures, *Acta Math. Sinica*, **21** (2005), no. 6, 1535–1544.

10. Decompositions of Besov-Morrey spaces and Triebel-Lizorkin-Morrey spaces, to appear in Math. Z.
11. L. Tang and J. Xu, Some properties of Morrey type Besov-Triebel spaces, Math. Nachr **278** (2005), no 7-8, 904–917.
12. H. Triebel, Theory of function spaces, Birkhäuser (1983).
13. H. Triebel, Theory of function spaces II, Birkhäuser (1992).
14. H. Triebel, Fractal and Spectra, Birkhäuser (1997).
15. H. Triebel, The structure of functions, Birkhäuser (2001).

YOSHIHIRO SAWANO: Department of Mathematics and Information Sciences, Tokyo Metropolitan University, Minami-Ohsawa 1-1, Hachioji-shi, Tokyo 192-0397, Japan,
E-mail: yoshihiro@tmu.ac.jp

MATHEMATICAL FINANCE AND PROBABILITY THEORY

J.A.M. VAN DER WEIDE

Abstract. Inspired by a problem from mathematical finance, we study sojourns of Brownian motion and random walk in the positive and the negative half-line. In [1] we get as a by-result in a study of double-sided Parisian options an interesting formula for the probability that Brownian motion stays above some level $l_2 > 0$ during a time interval of length D_2 before being below $l_1 < 0$ during an interval of length D_1 . If we take the levels $l_1, l_2 = 0$ we get a formula which is, except for taking squarehoods, very similar to formulas for the ruin problem. In this paper we explain these similarities from excursion theory. We also discuss the analog problems for symmetric random walk and from these results we get by scaling and taking limits the formulas for Brownian motion back. The complete results for symmetric random walk will be published in a separate paper [22] elsewhere.

Key words and Phrases: Random Walk, Ruin problem, Excursion, Brownian motion, Sojourn.

INTRODUCTION

The last 50 years we have seen a fast growing interest in probability models for financial applications. In his thesis [2] of 1900, Louis Bachelier proposed Brownian motion to model price processes of stocks traded at the Paris Stock Exchange. His work remained unnoticed for some time even though Poincaré, the greatest French intellect of that time, appreciated the thesis. Later in 1973, the work of Black and Scholes [4] and Merton [18] made it possible to develop a price theory for options. Their model is based on geometric Brownian motion. In their papers [11] and [12] of 1979 and 1981, Harrison, Kreps and Pliska explain the relations between martingales and notions as “arbitrage free” in financial economics. The martingale representation theorem can be used to construct “hedging” strategies. Nowadays,

jump diffusions and Lévy processes are studied for modeling price processes and advanced theories as Malliavin Calculus are applied. From this little overview it is clear that probability theory, in particular stochastic processes, play an important role in mathematical finance, but of course numerical methods, control theory, statistics and simulation are important as well. All this makes mathematical finance into an important motivation for research in probability theory. As an example of a result motivated by mathematical finance we present a new formula for Brownian motion that comes up in joint work with Jasper Anderluh [1] about double-sided Parisian options.

Options are contracts traded on exchanges all over the world. The most simple example is a call option on some stock. It gives the holder the right to buy from the writer of the contract at some future time T a share of the stock for a price K . The time T is called the maturity and K is the strike price. Both maturity and strike price are specified in the option contract. The holder will exercise his right to buy a share of the stock only if the stock price $S(T)$ at time T is higher than the strike. In this case his profit will be the difference $S(T) - K$, since he can sell immediately at the market the share of the stock for the price $S(T)$. Otherwise the holder of the option will not exercise his right because the stock can be bought cheaper in the market. In this case the option expires worthless. More accurately, the contract we described is called a European call option and its value is given by $(S_T - K)^+ = \max(S(T) - K, 0)$. A European put option gives the holder the right to sell a share of the stock at time T for the strike price K , so its value is $(K - S(T))^+$. Unless most of the options traded at exchanges are American options which give the holder the right to exercise the option at any time before maturity T but at most once, we will consider here only European style options. There are nowadays many books available in which the basic principles of pricing of options are discussed. A not too technical introduction with a lot of information about financial markets can be found for example in [13].

A Parisian option is a special type of barrier option. A down-and-in call barrier option with strike K , barrier L and time to maturity T pays only if the asset price $S(t)$ hits the barrier before maturity in which case the pay-off is given by $(S(T) - K)^+$, otherwise it expires worthless. There exist also down-and-out, up-and-in and up-and-out barrier options. The options traded at the Jakarta exchange are a special type of barrier options, see [10]. A down-and-in Parisian option is a barrier option that only knocks in if the asset price is below the barrier during a given unbroken time period of length D . A down-and-in double-sided Parisian option is defined by two barriers L_1 and L_2 , $L_1 < L_2$, and two time lengths D_1 and D_2 . The option knocks in if either the asset price is below L_1 during an unbroken time period D_1 or above L_2 during an unbroken time period D_2 . Modeling the asset price as a geometric Brownian motion, the Laplace transform of the price of the option can be calculated, see [1]. An important step in the derivation of this Laplace transform is a formula for the Laplace transform of the first time at which Brownian is a time interval of length D_1 below some level l_1 or during a time interval of length D_2 above some level l_2 . One of the consequences of this formula is

a nice formula for Brownian motion. The probability that Brownian motion stays positive during an unbroken time period of length D_2 before it is negative during an unbroken time period of length D_1 equals $\sqrt{D_1}/(\sqrt{D_1} + \sqrt{D_2})$. This formula will be explained in Section . The structure of this formula resembles the solution of the ruin problem, a problem very well studied in the probabilistic literature. A gambler starts playing with an amount of A Rupiah against an adversary who plays with a capital of B Rupiah. The game is continued till one of the players has lost his capital. Let S_t denote the total profit of the gambler up to and including time t . The gambler wins if the process S_t reaches level B before reaching level $-A$. The probability that the gambler wins is $A/(B + A)$. This problem can also be formulated for Brownian motion as the probability that the Brownian motion hits barrier B before hitting level $-A$. If A and B are integers, the formula for this probability turns out to be exactly the same as for the ruin problem, which can be reformulated as the probability that symmetric Random Walk hits barrier B before hitting level $-A$. In the same way we could reformulate for symmetric Random Walk the former problem about sojourns of Brownian motion in the positive and the negative real numbers. The Random Walk results will be discussed in Section . In Section we study these results from the point of view of excursion theory for Brownian Motion, which explains the similarity in the formulas for the ruin problem and the formula for sojourns.

DOUBLE-SIDED PARISIAN OPTIONS

As a model for a financial market we consider the Black-Scholes model with one stock and a bank account. The price process $\{S(t) : t \geq 0\}$ of the stock is geometric Brownian motion

$$S(t) = S_0 \exp\left\{\left(\mu - \frac{1}{2}\sigma^2\right)t + \sigma W(t)\right\}$$

where μ is the rate of return and $\sigma > 0$ the volatility. The process $\{W(t) : t \geq 0\}$ is a Wiener process or Brownian motion. The bank account is based on continuously compounding with fixed rate $r > 0$, so an initial capital of 1 grows to

$$B(t) = \exp(rt)$$

at time $t \geq 0$. In a barrier out contract the payoff is $\max(S(T) - K, 0)$ like a European call unless the price process hits some given barrier level L in which case there is no payoff. There exist also barrier in contract with payoff is $\max(S(T) - K, 0)$ only if the price process hits the barrier, otherwise there is no payoff. A Parisian option is a special type of barrier option that knocks in or out if the stock price stays a given period of time above or below the barrier after hitting the barrier. Parisian options are at present not traded at exchanges but they turn up in real options, see [9], and convertible bonds, see [15]. There are also applications in credit risk and life insurance, see [19] and [5] respectively. Parisian options were introduced in [6]. A more general type of option is the double-sided Parisian option which was introduced in [1]. We will discuss here the so-called double-sided Parisian knock-in call with maturity T and strike K . For this contract one has to

specify two barriers $L_1 > 0$ and $L_2 > 0$ (we always assume that $L_1 < L_2$) and two durations $D_1 > 0$ and $D_2 > 0$. Define for a stochastic process $\{X(t); t \geq 0\}$ and $D, L > 0$

$$\mathcal{S}_D^{L+}(X) = \inf\{t > D : X(u) > L \text{ for all } u \in (t - D, t]\}$$

and

$$\mathcal{S}_D^{L-}(X) = \inf\{t > D : X(u) < L \text{ for all } u \in (t - D, t]\}.$$

If, before maturity T , the price of the stock has been below barrier L_1 or above L_2 during a time period of length D_1 or D_2 respectively, i.e. $\min(\mathcal{S}_{D_1}^{L_1-}(S), \mathcal{S}_{D_2}^{L_2+}(S)) < T$, then the double-sided Parisian knock-in call pays at maturity T an amount $(S(T) - K)^+$, otherwise the option is worthless. At time $t = 0$, the value V_{DPIC} of this double-sided Parisian knock-in call is given by

$$V_{DPIC} = e^{-rT} E \left[(S(T) - K)^+; \min(\mathcal{S}_{D_1}^{L_1-}(S), \mathcal{S}_{D_2}^{L_2+}(S)) < T \right],$$

where the expectation is taken with respect to the risk-neutral measure under which the price process is a geometric Brownian motion with drift r , so

$$S(T) = S_0 \exp\left\{\left(r - \frac{1}{2}\sigma^2\right)T + \sigma W(T)\right\}.$$

For a complete calculation of V_{DPIC} we rewrite the problem in terms of Brownian motion. Here we recall one of the steps in this calculation. It can be shown that for $l_1 < 0 < l_2$

$$\begin{aligned} E \left[\exp\left\{-\frac{1}{2}\lambda^2 \min(\mathcal{S}_{D_1}^{l_1-}(W), \mathcal{S}_{D_2}^{l_2+}(W))\right\} : \mathcal{S}_{D_2}^{l_2+}(W) < \mathcal{S}_{D_1}^{l_1-}(W) \right] \\ = \frac{e^{\lambda l_1} \Psi(-\lambda_1) - e^{-\lambda l_1} \Psi(\lambda_1)}{e^{\lambda(l_1 - l_2)} \Psi(-\lambda_1) \Psi(-\lambda_2) - e^{\lambda(l_2 - l_1)} \Psi(\lambda_1) \Psi(\lambda_2)}, \end{aligned}$$

where $\lambda_i = \lambda \sqrt{D_i}$ and

$$\Psi(z) = E[e^{zU}; U \geq 0] = 1 + z \exp(z^2/2) \int_{-\infty}^z e^{-\frac{1}{2}x^2} dx.$$

There is an analogous formula for

$$E \left[\exp\left\{-\frac{1}{2}\lambda^2 \min(\mathcal{S}_{D_1}^{l_1-}(W), \mathcal{S}_{D_2}^{l_2+}(W))\right\} : \mathcal{S}_{D_2}^{l_2+}(W) > \mathcal{S}_{D_1}^{l_1-}(W) \right],$$

see [1] for details.

Here we want to consider the special case $l_1 = l_2 = 0$. Let

$$\begin{aligned} \mathcal{S}^- &= \inf\{t > D_1 : W(u) < 0 \text{ for all } u \in (t - D_1, t]\} \\ \mathcal{S}^+ &= \inf\{t > D_2 : W(u) > 0 \text{ for all } u \in (t - D_2, t]\} \end{aligned}$$

and $\mathcal{S} = \min(\mathcal{S}^-, \mathcal{S}^+)$. By letting $l_2 \downarrow 0$ and $l_1 \uparrow 0$, we get that

$$E \left[e^{-\frac{1}{2}\lambda^2 \mathcal{S}} : \mathcal{S}^+ < \mathcal{S}^- \right] = \frac{\Psi(-\lambda_1) - \Psi(\lambda_1)}{\Psi(-\lambda_1) \Psi(-\lambda_2) - \Psi(\lambda_1) \Psi(\lambda_2)}.$$

Letting $\lambda \rightarrow 0$, we get the formula,

$$P(\mathcal{S}^+ < \mathcal{S}^-) = \frac{\sqrt{D_1}}{\sqrt{D_1} + \sqrt{D_2}}.$$

Further

$$E \left[e^{-\frac{1}{2}\lambda^2 \mathcal{S}} \right] = \frac{\Psi(-\lambda_1) + \Psi(-\lambda_2) - [\Psi(\lambda_1) + \Psi(\lambda_2)]}{\Psi(-\lambda_1)\Psi(-\lambda_2) - \Psi(\lambda_1)\Psi(\lambda_2)}.$$

It follows that

$$E(\mathcal{S}) = -\lim_{\lambda \rightarrow 0} \frac{1}{\lambda} \frac{d}{d\lambda} E \left[e^{-\frac{1}{2}\lambda^2 \mathcal{S}} \right] = 2\sqrt{D_1}\sqrt{D_2}.$$

SOJOURNS

In this section we discuss for a simple symmetric random walk the ruin problem and sojourns in the positive and the negative integers. Let $\{S_n : n \geq 0\}$ be simple symmetric random walk on the integers starting at 0. So $S_0 \equiv 0$ and, for $n \geq 1$, the S_n can be considered as the partial sums of a sequence $X_1, X_2, \dots$ of independent random variables all defined on the same probability space $(\Omega, \mathcal{F}, P)$ with probability distributions

$$P(X_i = 1) = P(X_i = -1) = \frac{1}{2}, \quad i = 1, 2, \dots$$

For any integer a , define

$$(1) \quad \tau_a = \min\{n : S_n = a\}.$$

Let $a, b > 0$ be positive integers. The ruin problem amounts to the calculation of the probability of the event $\{\tau_b < \tau_{-a}\}$. It is well-known that

$$(2) \quad P(\tau_b < \tau_{-a}) = 1 - P(\tau_{-a} < \tau_b) = \frac{a}{a+b}.$$

The time at which either the player is bankrupted or the bank is $\min(\tau_b, \tau_{-a})$. Then

$$(3) \quad E(\min(\tau_b, \tau_{-a})) = ab,$$

see for example Feller[8].

Let d_1 and d_2 be given, positive integers. The first time at which the random walk has been > 0 during d_1 steps in a row will be denoted by $\mathcal{S}_{d_1}^{(+)}$, i.e.

$$\mathcal{S}_{d_2}^{(+)} = \min\{n : S_{n-d_2+1} > 0, \dots, S_n > 0\},$$

and, likewise, $\mathcal{S}_{d_1}^{(-)}$ denotes the first time at which the random walk has been < 0 during d_1 steps in a row. We want to calculate the probability that a run of at least d_2 positive numbers occurs before a run of d_1 negative numbers. Let $q_0 = 1/2$ and

$$q_d = P(S_1 > 0, \dots, S_d > 0), \quad d = 1, 2, \dots$$

By symmetry, $P(S_1 < 0, \dots, S_d < 0) = q_d$. It follows from Feller[8], Chapter III.3, that

$$q_{2n} = q_{2n+1} = \frac{1}{2} \binom{2n}{n} 2^{-2n}, \quad n = 0, 1, 2, \dots$$

Then

$$P(\mathcal{S}_{d_2}^{(+)} < \mathcal{S}_{d_1}^{(-)}) = \frac{q_{d_2}}{q_{d_1} + q_{d_2}},$$

and

$$E(\min(\mathcal{S}_{d_2}^{(+)}, \mathcal{S}_{d_1}^{(-)})) = \frac{\alpha + d_1 q_{d_1} + d_2 q_{d_2}}{q_{d_1} + q_{d_2}},$$

where

$$\alpha = 2 \lfloor \frac{d_1}{2} \rfloor q_{d_1} + 2 \lfloor \frac{d_2}{2} \rfloor q_{d_2}.$$

The proofs of these results, as well explicit formulas for sojourns below and above given levels $-l_1$ and l_2 , where l_1 and l_2 are given positive integers, can be found in [22]. The proofs are based on excursion theory. We give the proof of the first formula. Define, for $k = 0, 1, 2, \dots$ the stopping times

$$\tau_0^{k+1} = \min\{n > \tau_0^k \mid S_n = 0\},$$

where $\tau_0^0 \equiv 0$. Then $\{\tau_0^k : k \geq 0\}$ is the sequence of times at which the sample path of the random walk visits 0. The sub-sample path

$$\mathcal{E}_k = (S_{\tau_0^{k-1}+1}, \dots, S_{\tau_0^k-1}, S_{\tau_0^k}) = (0, S_{\tau_0^{k-1}+1}, \dots, S_{\tau_0^k-1}, 0)$$

is called the k th excursion from 0. The use of the term excursion here could be confusing. We do not, as for example in [7], mean the excursion straddling some fixed time point n , which is the excursion containing S_n . Let U be the discrete space of all excursions $\mathbf{u} = (u_0, u_1, \dots, u_{2n})$ for some value $n = 1, 2, \dots$, such that $u_0 = u_{2n} = 0$, $u_1 \neq 0, \dots, u_{2n-1} \neq 0$ and $u_i - u_{i-1} = \pm 1$, $i = 1, \dots, 2n$. Then, by the strong Markov property, the sequence of random variables $\mathcal{E}_k : \Omega \mapsto U$, $k \geq 1$, is iid with probability distribution

$$P(\mathcal{E}_k = (u_0, u_1, \dots, u_{2n})) = \left(\frac{1}{2}\right)^{2n}, \quad (u_0, u_1, \dots, u_{2n}) \in U.$$

Now consider

$$A = \{\mathbf{u} \in U \mid u_1 > 0, \dots, u_{d_2} > 0\}$$

and

$$\nu_A = \min\{n : \mathcal{E}_n \in A\}.$$

So ν_A is the number of the first excursion in which a sojourn of length d_1 in the positive integers occurs. By definition of the q_n , for any k , $p_A = P(\mathcal{E}_k \in A) = q_{d_2}$. In the same way we define ν_B , where

$$B = \{\mathbf{u} \in U \mid u_1 < 0, \dots, u_{d_1} < 0\},$$

and $p_B = P(\mathcal{E}_k \in B) = q_{d_1}$. Now

$$P(\nu_A = 1, \nu_B > 1) = P(\mathcal{E}_1 \in A) = p_A,$$

and, for $k > 1$,

$$\begin{aligned} P(\nu_A = k, \nu_B > k) &= P(\mathcal{E}_i \notin A \cup B, i \leq k-1, \mathcal{E}_k \in A) \\ &= (1 - p_A - p_B)^{k-1} p_A, \end{aligned}$$

since $A \cap B = \emptyset$. Hence

$$P(\nu_A < \nu_B) = \sum_{k=1}^{\infty} P(\nu_A = k, \nu_B > k) = \frac{p_A}{p_A + p_B}.$$

It follows that

$$P\left(\mathcal{S}_{d_2}^{(+)} < \mathcal{S}_{d_1}^{(-)}\right) = P(\nu_A < \nu_B) = \frac{q_{d_2}}{q_{d_1} + q_{d_2}}.$$

This argument is also helpful to understand the similarity between the formulas for the ruin problem and the formulas for the sojourns. Let

$$M(\mathbf{u}) = \max u_i, \quad \mathbf{u} \in U,$$

be the maximum of the excursion. The probability distribution of $M(\mathcal{E}_k)$ is given by $P(M(\mathcal{E}_1) = 0) = 1/2$ and

$$P(M(\mathcal{E}_1) = a) = \frac{1}{2a(a+1)}, \quad a > 0.$$

The probability distribution of the minimum $m(\mathbf{u}) = \min u_i$ of an excursion is given by

$$P(m(\mathcal{E}_k) = 0) = \frac{1}{2} \text{ and } P(m(\mathcal{E}_k) = -b) = \frac{1}{2b(b+1)}, \quad b > 0.$$

Apparently, the maximum of the excursion straddling n has a different probability distribution, see [7]. Now consider

$$C = \{\mathbf{u} \in U \mid M(\mathbf{u}) \geq b\} \quad \text{and} \quad D = \{\mathbf{u} \in U \mid m(\mathbf{u}) \leq -a\}$$

with $p_C = P(\mathcal{E}_k \in C) = 1/b$ and $p_D = 1/a$. Hence

$$P(\tau_b < \tau_{-a}) = P(\nu_C < \nu_D) = \frac{p_C}{p_C + p_D} = \frac{a}{a+b}.$$

To compare the results for symmetric random walk with the results in Section , we use the classical approximation of Brownian motion by scaled random walk $\frac{1}{\sqrt{n}}S(\lfloor nt \rfloor)$, see for example [3]. So

$$P(\mathcal{S}^+ < \mathcal{S}^-) = \lim_{n \rightarrow \infty} P\left(\mathcal{S}_{nD_2}^{(+)} < \mathcal{S}_{nD_1}^{(-)}\right) = \lim_{n \rightarrow \infty} \frac{q_{nD_2}}{q_{nD_1} + q_{nD_2}},$$

where $\mathcal{S}^+$ and $\mathcal{S}^-$ are defined in Section . By Stirling's formula,

$$n! \sim \sqrt{2\pi n} n^{n+\frac{1}{2}} e^{-n}, \quad n \rightarrow \infty,$$

hence

$$q_n \sim \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{n}} \quad \text{as } n \rightarrow \infty,$$

and it follows that

$$\lim_{n \rightarrow \infty} \frac{q_{nD_2}}{q_{nD_1} + q_{nD_2}} = \frac{\sqrt{1/\sqrt{D_2}}}{\sqrt{1/\sqrt{D_1}} + \sqrt{1/\sqrt{D_2}}} = \frac{\sqrt{D_1}}{\sqrt{D_1} + \sqrt{D_2}}.$$

In the same way with $\mathcal{S}$ as defined in Section , we get

$$\begin{aligned} E(\mathcal{S}) &= \lim_{n \rightarrow \infty} E(\min(\frac{1}{n}\mathcal{S}_{nD_2}^{(+)}, \frac{1}{n}\mathcal{S}_{nD_1}^{(-)})) \\ &= \lim_{n \rightarrow \infty} \frac{2(D_1q_{nD_1} + D_2q_{nD_2})}{q_{nD_1} + q_{nD_2}} = 2\sqrt{D_1}\sqrt{D_2}. \end{aligned}$$

So asymptotically, the results for symmetric random walk are the same as for Brownian motion.

EXCURSION THEORY FOR BROWNIAN MOTION

The results of Section suggest that we should study excursions of Brownian motion. A problem is that any interval $[0, \delta)$, $\delta > 0$, contains almost surely an infinite number of zeros of the Brownian path, so there is no first excursion, like there is no smallest rational number $> \sqrt{2}$. Lévy's studies [17] and [16] of the sample paths of Brownian motion can be considered as the starting point of excursion theory which led finally to Itô's famous paper [14] which contains a description of the excursions as the point of a Poisson point process. We will follow Scheffer [20], see also [21]. Let $\{W(t) : t \geq 0\}$ be standard Brownian motion defined on some probability space $(\Omega, \mathfrak{F}, P)$ with local time

$$L_t = \lim_{\varepsilon \downarrow 0} \frac{1}{2\varepsilon} \int_0^t 1_{(-\varepsilon, \varepsilon)}(W(s)) \, ds.$$

The sample paths $t \mapsto W_t(\omega)$, $\omega \in \Omega$, are continuous functions and an excursion from 0 is defined as a maximal, connected subinterval of $[0, \infty) \setminus \{t \mid W_t(\omega) = 0\}$. So, $e(\omega) = [a, b]$ is an excursion interval iff $W_a(\omega) = W_b(\omega) = 0$ and $W_t(\omega) \neq 0$ for all $t \in (a, b)$. The sign $\mathcal{S}(e(\omega)) = +$ (resp. $-$) if $W_t(\omega)$ is positive (resp. negative) for all $t \in (a, b)$. The local time $\mathfrak{L}_{e(\omega)}$ of the excursion interval $e(\omega) = [a, b]$ is defined as $L_a(\omega)$ and the duration is defined as $\zeta_{e(\omega)} = b - a$. To any $\omega \in \Omega$ we assign the simple point measure N_ω on $[0, \infty) \times \mathbb{R} \setminus \{0\} \times \{-, +\}$ with support

$$\{(\mathfrak{L}_{e(\omega)}, \zeta_{e(\omega)}, \mathcal{S}_{e(\omega)}) \mid e(\omega) \text{ is an excursion interval of } t \mapsto W_t(\omega)\}.$$

The distribution of the point process $\omega \mapsto N_\omega$ is a Poisson process on $U = [0, \infty) \times \mathbb{R} \setminus \{0\} \times \{-, +\}$ with intensity measure

$$\nu(dt \, du, +) = \nu(dt \, du, -) = \frac{1}{2\sqrt{2\pi}u^3} \, dt \, du.$$

In particular

$$t \in [0, \infty) \mapsto N_\omega([0, t) \times [D_1, \infty) \times \{-\})$$

is a Poisson process on the positive real line with intensity

$$\int_{D_1}^{\infty} \frac{1}{2\sqrt{2\pi}u^3} du = \frac{1}{\sqrt{2\pi}D_1}.$$

This process counts, in local time, the number of negative excursions with durations exceeding D_1 . In the same way, the number of positive excursions with durations exceeding D_2 is a Poisson process with intensity $1/\sqrt{2\pi}D_2$. In addition, these Poisson processes are independent. So, measured in local time, the time till the first positive excursion with duration $\geq D_2$ and the time till the first negative excursion with duration $\geq D_1$ are independent, exponentially distributed random variables with parameters $1/\sqrt{2\pi}D_2$ and $1/\sqrt{2\pi}D_1$ respectively. So the probability that a positive excursion with duration $\geq D_2$ occurs before a negative excursion with duration $\geq D_1$ is equal to

$$\frac{1/\sqrt{2\pi}D_2}{1/\sqrt{2\pi}D_1 + 1/\sqrt{2\pi}D_2} = \frac{\sqrt{D_1}}{\sqrt{D_1} + \sqrt{D_2}}.$$

To calculate the expected value of $\mathcal{S} = \min(\mathcal{S}_-, \mathcal{S}_+)$, we have to sum the durations of all excursions before $\mathcal{S}$:

$$\sum_{(\mathfrak{L}_{e(\omega)}, \zeta_{e(\omega)}, -)} \zeta_{e(\omega)} \mathbf{1}_{\{\mathfrak{L}_{e(\omega)} < \mathfrak{L}_{\mathcal{S}(\omega)}\}}$$

is the sum of the durations of all negative excursions before the first negative excursion with duration $\geq D_1$ or the first positive excursion with duration $\geq D_2$, which is equivalent to $\mathfrak{L}_{e(\omega)} < \mathfrak{L}_{\mathcal{S}(\omega)}$. We prefer to write this sum as an integral

$$\int_{[0, \infty) \times \mathbb{R} \setminus \{0\}} u \mathbf{1}_{\{N_\omega([0, t] \times A) = 0\}} N_\omega(dt du, -)$$

where

$$A = [D_1, \infty) \times \{-\} \cup [D_2, \infty) \times \{+\}.$$

Hence $\mathcal{S}(\omega)$ can be written as

$$\begin{aligned} \int u \mathbf{1}_{\{N_\omega([0, t] \times A) = 0\}} N_\omega(dt du, -) + \int u \mathbf{1}_{\{N_\omega([0, t] \times A) = 0\}} N_\omega(dt du, +) \\ + D_1 \mathbf{1}_{\{\mathcal{S} = \mathcal{S}_-\}} + D_2 \mathbf{1}_{\{\mathcal{S} = \mathcal{S}_+\}}, \end{aligned}$$

where both integrals are over $[0, \infty) \times \mathbb{R} \setminus \{0\}$. To calculate $E(\mathcal{S})$, we use the Palm formula for Poisson processes. This formula is well known and it describes how we can interchange the order of an integral over the sample space Ω with respect to P followed by an integral over the phase space of the point process with respect to

the point measure N_ω . We get

$$\begin{aligned}
& \int_{\Omega} \int u \mathbf{1}_{\{N_\omega([0,t] \times A)=0\}} N_\omega(dt du, +) P(d\omega) \\
&= \int_{\Omega} \int_{\Omega} u \mathbf{1}_{\{(N_\omega + \delta_{(t,u,+)})([0,t] \times A)=0\}} P(d\omega) \nu(dt du, +) \\
&= \int_{t=0}^{\infty} \int_{u=0}^{D_2} u \exp\left(-t \left[\frac{1}{\sqrt{2\pi D_1}} + \frac{1}{\sqrt{2\pi D_2}}\right]\right) \frac{1}{2\sqrt{2\pi}u^3} dt du \\
&= \frac{D_2 \sqrt{D_1}}{\sqrt{D_1} + \sqrt{D_2}}.
\end{aligned}$$

In the same way we get

$$\int_{\Omega} \int u \mathbf{1}_{\{N_\omega([0,t] \times A)=0\}} N_\omega(dt du, -) P(d\omega) = \frac{D_1 \sqrt{D_2}}{\sqrt{D_1} + \sqrt{D_2}},$$

and it follows that

$$\begin{aligned}
E(S) &= \frac{D_2 \sqrt{D_1} + D_1 \sqrt{D_2}}{\sqrt{D_1} + \sqrt{D_2}} + D_1 \frac{\sqrt{D_2}}{\sqrt{D_1} + \sqrt{D_2}} + D_2 \frac{\sqrt{D_1}}{\sqrt{D_1} + \sqrt{D_2}} \\
&= 2\sqrt{D_1} \sqrt{D_2}.
\end{aligned}$$

So, the expected value of the time till the first positive excursion with duration $\geq D_2$ or the first negative excursion with duration D_1 , whichever occurs first, is equal to $\sqrt{D_1} \sqrt{D_2}$. The factor 2 in the formula for $E(S)$ can be explained as coming from the contribution of the time spent in the last excursion to complete the duration D_1 or D_2 .

Acknowledgement.

*

References

- [1] J.H.M. Anderluh and J.A.M. van der Weide. Double-sided Parisian Option Pricing. 2007 (submitted).
- [2] L. Bachelier. Théorie de la spéculation. *Ann. Ecole Norm. Sup.*, 17:21–87, 1900.
- [3] P. Billingsley. *Convergence of Probability Measures*. Wiley, 1968.
- [4] F. Black and M. Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3):637–659, 1973.
- [5] A. Chen and M. Suchaneki. Default risk, bankruptcy procedures and the market value of life insurance liabilities. *Insurance, Mathematics and Economics*, 40:231–255, 2007.
- [6] M. Chesney, M. Jeanblanc-Picqué and M. Yor. Brownian excursions and parisian barrier options. *Adv. Appl. Prob.*, 29:165–184, 1997.
- [7] E. Csáki and S.G. Mohanty. Excursion and meander in random walk. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, 9(1):57–70, 1981.
- [8] W. Feller. *An Introduction to Probability Theory and its Applications; Volume 1*. New York: John Wiley & Sons, 1950.
- [9] L. Gauthier. Excursions height- and length-related stopping times and application to finance. *Adv. Appl. Prob.*, 34:846–868, 2002.

- [10] Gunardi, J.H.M. Anderluh, J.A.M. van der Weide, Subanar and Sri Hayatmi. Indonesian options. *Working paper Gadjah Mada University, Yogyakarta*, 2007.
- [11] J. Harrison and J. Kreps. Martingales and arbitrage in multiperiod securities markets. *Journal of Economic Theory*, 11:418–443, 1979.
- [12] J. Harrison and S. Pliska. Martingales and stochastic integrals in the theory of continuous trading. *Stochastic Processes and Applications*, 11:215–260, 1981.
- [13] John C. Hull. *Options, Futures, and other Derivatives, 6th Edition*. Prentice Hall, 2005.
- [14] K. Itô. Poisson point processes attached to markov processes. In *Proceedings of the sixth Berkeley Symposium on Mathematical Statistics and Probability*, pages 255–329. University of California Press, 1970.
- [15] K.W. Lau and Y.K. Kwok. Optimal calling policies in convertible bonds. *Proceedings of International Conference on Computational Intelligence for Financial Engineering*, pages 109–114, 2003.
- [16] P. Lévy. Sur certain processus stochastiques homogènes. *Comp. Math.*, 7:283–339, 1939.
- [17] P. Lévy. *Processus stochastiques et mouvement Brownien*. Gauthier-Villars, Paris, 1948.
- [18] R.C. Merton. Theory of rational option pricing. *Bell Journal of Economics and Management Science*, 4(1):141–183, 1973.
- [19] F. Moraux. Valuing corporate liabilities when the default threshold is not an absorbing barrier. *Working paper Université de Rennes*, 2007.
- [20] C.L. Scheffer. The rank of the present excursion. *Stoch. Proc. Appl.*, 55(1):101–118, 1995.
- [21] J.A.M. van der Weide. *Stochastic Processes and Point Processes of Excursions*. Stichting Mathematisch Centrum, Amsterdam, 1994.
- [22] J.A.M. van der Weide, A.T. Hensbergen and J.H.M. Anderluh. Sojourns of simple random walk. *Working paper Delft Institute of Applied Mathematics*, 2007.

J.A.M. VAN DER WEIDE: Department of Applied Mathematics, Delft University of Technology, P.O. Box 5301, NL-2600 GA Delft, The Netherlands.

E-mail: seams2007@ugm.ac.id.

THE DISCOVERY OF TEMPORAL PATTERNS AND RULES FROM SEQUENTIAL DATA

EDI WINARKO

Abstract. A huge amount of data is collected everyday and a real universal challenge is to find actionable knowledge from such large amount of data. Data mining is an emerging research direction to meet this challenge. Data mining techniques can be deployed to search large databases to discover useful information that might otherwise remain unknown. Many data mining problems involve temporal aspects, with examples ranging from transaction databases in health care and insurance, stock exchange and customer goods in market sectors, sensor data collected from sensor networks, to scientific databases in geophysics and astronomy. Temporal data mining deals with the problem of mining patterns or rules from this temporal data. This paper discusses the discovery of temporal rules from sequential data, that is, ordered lists of events (nominal symbols from a particular alphabet) where each event has an associated time of occurrence. An ordered list of events occurring at an instant is called an *event sequence*, while an ordered list of events occurring over a time interval is called an *interval sequence*. A large number of studies have been concentrated on analyzing event sequence data. In contrast, the analysis of interval sequence data has received relatively little attention, so that there are still many issues that need to be addressed. The first part of this paper reviews previous studies on the discovery of temporal patterns from event sequence data. The second part discusses our solutions to three main issues related to the pattern discovery from interval sequence data.

Temporal mining, temporal rules, sequential data:

1. INTRODUCTION

As our world is now in its information era, a huge amount of data is collected everyday. A real universal challenge is to find actionable knowledge from such large

2000 Mathematics Subject Classification: 68P15

amount of data. Data mining is an emerging research direction to meet this challenge. Data mining techniques can be deployed to scour large databases to discover useful information that might otherwise remain unknown. Data mining is actually an integral part of *Knowledge Discovery in Database* (KDD) process, which is the overall process of converting raw data into useful information [11]. This process consists of many steps: 1) preparing the data, 2) setting search parameters, 3) searching for and extracting patterns, 4) evaluating and interpreting the results, and 5) refining the search for further iterations. Data mining and knowledge discovery have become important issues for research over the past decade. This has been caused not only by the growth in the size of datasets but also in the availability of otherwise unavailable datasets over the Internet and the increased value that organizations now place on the knowledge that can be gained from data analysis.

Data mining techniques have been applied for analyzing various data. In this paper, we focus our discussion on mining patterns from temporal data. Examples of temporal data include alarms/events and performance measurements generated by distributed computer systems and by telecommunication networks, the web server logs, on line transaction logs, financial data, and sensor data collected from sensor networks.

This paper is organized as follows. Section 2 describes briefly the field of temporal data mining. Section 3 reviews previous work on the discovery of patterns from event sequence data, while Section 4 describes three main issues related to the pattern discovery from interval sequence data as well as our proposed solutions in dealing with these issues. Our conclusion is given in Section 5.

2. TEMPORAL DATA MINING

Temporal data mining is an extension of data mining that deals with the problem of mining patterns from temporal data, which can be either *symbolic sequences* or *numerical time series*. It has the capability to look for interesting correlations or rules in large sets of temporal data, which might be overlooked when the temporal component is ignored or treated as a simple numeric attribute [49]. Currently, temporal data mining is a fast expanding field and covers a wide spectrum of paradigms for knowledge modeling and discovery. Based on its main tasks, temporal data mining can be grouped into five broad categories: prediction, classification, clustering, search and retrieval, and pattern discovery [26].

Prediction

Prediction is the task of explicitly modeling variable dependencies to predict a subset of the variables from others. The task of time series prediction is to forecast future values of the time series based on its past samples. In order to perform the prediction, one needs to build a predictive model from the data. Koskela *et al.* (1996) have studied neural networks for nonlinear modeling of time series data. The prediction problem for symbolic sequences has been addressed in AI research by Dietterich and Michalski (1985).

Classification

Classification is the task of assigning class labels to the data according to a model learned from the training data where the classes are known. Classification is one of the most common tasks in supervised learning, but it has not received much attention in temporal data mining [5]. In sequence classification, each sequence presented to the system is assumed to belong to one of predefined classes and the goal is to automatically determine the corresponding category for a given input sequence. Examples of sequence classification applications include signature verification [41], gesture recognition [56], and hand-written word recognition [25].

Clustering

Clustering¹ is the process of finding intrinsic groups, called clusters, in the data. Clustering of time series (or sequences) is concerned with grouping a collection of time series (or sequences) based on their similarity. Time series clustering has been shown effective in providing useful information in various domains [30]. For example, in financial data, clustering can be used to group stocks that exhibit similar trends in price movements. Another example could be clustering of fMRI time series for identifying regions with similar patterns of activation [12]. Clustering of sequences is relatively less explored but is becoming increasingly important in data mining applications such as web usage mining and bioinformatics [26]. A survey on clustering time series has been presented by Liao (2005).

Searching and Retrieval

Searching and retrieval are concerned with efficiently locating subsequences or sub-series (often referred to as queries) in large databases of sequences or time series. In data mining, query based searches are more concerned with the problem of efficiently locating approximate matching than exact matching, known as content-based retrieval. An example of a time series retrieval application is to find out all the days of the year in which a particular stock had similar movements to those of today. Another example is finding products with similar demand cycles. An example of a sequence retrieval is finding gene expression patterns that are similar to the expression pattern of a given gene. In order to address the time series retrieval problem, different notions of similarity between time series and indexing techniques have been proposed. There is considerably less work in the area of sequence retrieval, and the problem is more general and difficult. For more detail about time series and sequence retrieval can be found in Das and Gunopulos (2003).

¹Clustering is sometimes called unsupervised classification.

Pattern Discovery

Unlike in search and retrieval applications, in the pattern discovery there is no specific query in hand with which to search the database. The objective is simply to discover all patterns of interest. While the other tasks described earlier have their origins in other disciplines like statistics, machine learning or pattern recognition, the pattern discovery task has its origin in data mining itself. A pattern is a local structure in the data. There are many ways of defining what constitutes a pattern. There is no universal notion for interestingness of a pattern either. However, one concept that is normally used in data mining is that of frequent patterns, that is, patterns that occurs many times in the data. Much of data mining literature is concerned with formulating useful pattern structures and developing efficient algorithms for discovering frequent patterns.

Methods for finding frequent patterns are important because they can be used for discovering useful rules, which in turn can be used to infer some interesting regularities in the data. A rule usually consists of a pair of a left-hand side proposition (the antecedent) and a right-hand side proposition (the consequent). The rule states that when the antecedent is true, then the consequent will be true as well.

3. PATTERN DISCOVERY FROM EVENT SEQUENCE DATA

A large number of previous studies on discovering temporal patterns from sequential data have focused on data that are stamped with, and interpreted as, time points. Several variations of the classical association rule mining [7] have been proposed under the name of temporal association rules mining, whose task is to discover temporal association rules from a set of timestamped transactions. Abstractly, such data can be viewed as a sequence of event sets; and the temporal association rule represents a temporal relationship between a set of events.

Agrawal and Srikant (1995) introduced the discovery of sequential patterns from a set of data sequences. While in temporal association rule mining each transaction is normally treated individually with no record of the associated object, in sequential pattern mining each transaction is associated with a particular object at a given time. Each data sequence consists of an ordered list of transactions belonging to a particular object (e.g. customer or client who purchased the transactions). Even though the framework is described using an example of mining a database of customer transaction sequences, the concept of sequential patterns is quite general and can be used in many other areas, such as analysing web access sequences and protein sequences. In general, a sequential pattern represents a correlation among the events in the sequence.

Another framework for discovering temporal patterns in sequential data is the mining of frequent episodes [36]. In this framework, the data are given in a single long sequence of events, and the task is to discover frequent episodes in the sequence, which are later used to generate episode rules. A frequent episode is essentially a frequent sequence, but instead of being frequent across many data

sequences, an episode is frequent within one sequence. Later, Han *et al.* (1998) introduced the discovery of partially periodic patterns from a long sequence of events.

This section reviews previous studies on the discovery of temporal patterns from point-based sequential data. In particular, the review focuses on the discovery of temporal association rules, sequential patterns, episodes and periodic patterns. The section is structured as follows. First, we provide a survey of different types of temporal association rule and methods to discover the rules (Subsection 3.1). Subsection 3.2 reviews algorithms for solving the basic sequential pattern mining problem. Various extensions to the sequential pattern mining problem are also discussed. Subsection 3.3 describes two basic algorithms for mining frequent episodes, highlights several extensions to episode mining model, and briefly discusses periodic pattern mining.

3.1 Temporal Association Rules

Several types of temporal association rules have been proposed as extensions of the classical association rules [1]. Ozden *et al.* (1998) introduced cyclic association rules and presented two algorithms to discover such rules. A cyclic association rule is an association rule that occurs periodically over time. However, periodicity has limited power in describing real-life variations. For instance, periodicity cannot describe real-life concepts such as *the first business day of every month* in which the distances between two such consecutive business days are not always the same. Ramaswamy *et al.* (1998), therefore, consider the discovery of association rules that hold during the time intervals described by a calendar algebraic expression. The calendar algebra adopted is considered more powerful than periodicity in describing time intervals of interest. Although the work of Ramaswamy *et al.* (1998) is more flexible than that of Ozden *et al.* (1998), it requires the user to define the calendar algebraic expression. To provide more flexibility to the user, Li *et al.* (2001) propose a calendar schema as a framework for discovering temporal association rules.

Ale and Rossi (2000) consider the discovery of association rules hold during the lifetime (lifespan) of items involved in the rules. An item's lifespan is a period between the first and the last time the item appears in transactions in the dataset. The lifespan is intrinsic to the data so that the users do not need to define it. Similarly, Lee *et al.* (2001) consider the problem of mining general temporal association rules in publication databases. A publication database is a set of transactions where each transaction contains an individual exhibition period. Zimbrão *et al.* (2002) combine the work of Li *et al.* (2001) and Ale and Rossi (2000) and propose a new approach to discover calendar-based association rules with an item's lifespan restriction.

Chen and Petrounias (2000) present the discovery of the longest interval and the longest periodicity of association rules. Rainsford and Roddick (1999) propose to add temporal features to association rules by associating a conjunction of binary temporal predicates that specify the relationships between the timestamps

of transactions. Subsequently, Rainsford and Roddick (2000) provide visualisation methods for viewing these temporal predicate association rules.

Previous studies on temporal association rules mining have focused on developing temporal association rule models. The proposed models can be classified based on three different aspects: measures of interestingness used, temporal feature associated with the rules, and algorithms used to discover the rules. In classical association rule mining, two commonly used measures of interestingness are *support* and *confidence* [1]. Both reflect the usefulness and certainty of discovered rules [15]. However, for some of the temporal association rule models discussed here, these two measures are considered insufficient and additional measures are proposed, for example, *temporal support*, *frequency*, and *temporal confidence*. Temporal support is used to filter the items with high support but short life. The combination of support and temporal support is used to determine if an itemset is frequent [4]. Frequency measures the proportion of the intervals during which the rules hold with respect to a set of given time intervals [8]. Temporal confidence determines the strength of the temporal relationships between temporal items in the rule [46].

Most of the algorithms for discovering temporal association rules are *Apriori*-based algorithms². The only algorithm that is not based on the Apriori is the Progressive Partition Miner (PPM) algorithm [27].

3.2 Sequential Patterns

Sequential pattern mining, which discovers frequent sequences in a sequence database, is an important data mining problem with broad applications, such as the analysis of customer purchase patterns, web access patterns, scientific experiments, disease treatments, natural disasters, and DNA sequences [45]. In sequential pattern mining a database D is a set of transactions $(CID, TID, ITEMS)$, where CID is a customer-id, TID is a transaction-id (based on the transaction time), and $ITEMS$ is a list of items in the transaction. For a given customer-id, there are no transactions with the same transaction time, which means that there are no transactions with the same transaction-id. All the transactions with the same CID can be regarded as a sequence of itemsets ordered by increasing TID.

Since it was first introduced by Agrawal and Srikant (1995), it is one of the active research areas in temporal data mining. As a result, a number of algorithms and techniques have been proposed for mining sequential patterns. Earlier algorithms for sequential pattern mining are Apriori-like algorithms, based on the Apriori property proposed for mining frequent itemsets [2]. A series of Apriori-like algorithms have been proposed, among others, AprioriAll [3], GSP [50], and SPADE [57]. As alternatives to the Apriori-based approach, other methods have also been proposed, such as FreeSpan [16] and PrefixSpan [45]. These new methods use a pattern-growth approach, based on the FP-growth algorithm for mining frequent itemsets [18]. Various extensions of sequential pattern mining have been developed, such as mining constraint-based sequential patterns, mining closed sequential pat-

²Note that the Apriori algorithm refers to the algorithm to discover association rules [2]

terns, incremental and interactive mining of sequential patterns. Most extensions are solved by extending the basic algorithms for mining sequential patterns.

3.3 Episodes

The problem of discovering episodes from a long sequence of events was introduced by Mannila *et al.* (1995). An episode is a collection of events that occur relatively close to each other in a certain (partial) order, whose total span of time is constraint by a window. There are two basic approaches for discovering episodes. The first approach is based on the occurrences of the patterns in a sliding window along the sequence and used by the WINEPI algorithm [36, 34]. The second one relies on the notion of minimal occurrences of the patterns and used by the MINEPI algorithm [35, 34]. Further studies have been undertaken into variations that attempt to extend the basic episode discovery framework, either by improving the algorithms or by enriching the generated patterns. Furthermore, other studies have also been conducted that focus on the discovery of periodic patterns. A periodic pattern is an ordered list of events that repeats itself in the sequence of events.

4. PATTERN DISCOVERY FROM INTERVAL SEQUENCE DATA

Previous section has presented numerous studies in the discovery of temporal patterns from point-based sequential data, where events in the sequence are instantaneous. According to Böhlen *et al.* (1998), in many applications events are not instantaneous; they instead occur over a time interval. For instance, consider a database application in which a data item is locked and then unlocked sometime later. Instead of treating the lock and unlock operations as two discrete events, it can be advantageous to interpret them together as a single interval event that better captures the significance of placing, holding and releasing the lock. When there are several such events, a series of interval events is formed [53]. Because events are extended in time, different events may overlap in time and interact. Temporal data mining methods can be used to extract patterns from these interacting interval events.

Several applications exist where the discovery of temporal relations among interval events can provide insight about the operation of the system in question. Consider a sign language database that contains useful linguistic information on a variety of grammatical and syntactic structures, as well as manual and non verbal fields. Detecting relations between the above structures and fields could be useful to the linguist and may help them discover new type of relations they have not known before. Another application is in network monitoring. Multiple types of events occurring over certain time periods can be stored in a log, and the goal is to detect general temporal relations of these events that with high probability would describe regular patterns in the network [44]. Other application areas include analysis of weather data [19], muscle activity [40], audio data [39], and video data [38]. Despite these facts, research on mining temporal patterns from interval-based sequential data has received little attention.

As a result, there are still many issues that need to be addressed in the area of analysing interval sequence data. Three main issues that will be addressed in this section are as follows.

1. The first important issue is that of what constitutes an interesting pattern in data. As an example, the notions of sequential patterns or frequent episodes represent the currently popular structures for patterns in event sequence data. Therefore, the problem of defining structure for interesting patterns in interval sequence data would be a problem that deserves attention.
2. In all data mining applications, the primary constraint is the large volume of data. Hence there is always a need for efficient algorithms. Therefore, designing efficient algorithms for the discovery of patterns in interval sequence data is a problem that would continue to attract attention.
3. Another important issue is that of the analysis of discovered temporal patterns so that one can find interesting patterns from a large number of patterns generated by data mining algorithms. Making sense of such a large number of patterns presents a significant challenge. Therefore, there is a need for tools to assist the user finding interesting patterns from a set of discovered patterns.

4.1 Dealing with the First Issue: Defining Patterns

Only a few previous studies consider the discovery of temporal patterns from a long sequence of intervals, which might originate from multivariate time series. Villafane *et al.* (2000) propose a technique to discover containment relationships from interval time series. A containment lattice is constructed from the intervals, and then the containment relationships are discovered with the *Growing Snake Traversal* method. Hoppner (2001) mines temporal rules expressed with Allen's interval logic and a sliding window to restrict the pattern length. The patterns are mined with an Apriori-like algorithm. In other work, temporal patterns are extracted from a symbolic interval sequence obtained from multivariate time series and formulated using the UTG [13, 40, 39]. In Mörchen (2006) the patterns are expressed using the TSKR.

The discovery of temporal patterns from interval sequence databases proposed by Kam and Fu (2000) and Papapetrou *et al.* (2005). Kam and Fu (2001) have proposed *A1 temporal patterns* to represent relationships between intervals found in a set of interval sequences. Papapetrou *et al.* (2005) studied the discovery of frequent *arrangements* from interval sequence databases.

In order to address the problem of defining pattern structure, we have formulated the problem of discovering richer temporal association rules from interval sequence databases. The richer temporal association rules are formulated using the pattern formulation proposed by Höppner (2001). However, our work is different from the work of Höppner, because the richer temporal association rules are mined from a set of interval sequences, instead of a long sequence of intervals.

4.2 Dealing with the Second Issue: Designing Algorithms

Discovering temporal rules from interval-based sequential data is certainly more complex and requires a different approach from mining patterns from point-based sequential data, such as mining sequential patterns or episodes. An interval has duration and therefore the generated patterns have different semantics than simply *before* and *after*.

In order to deal with this issue, we propose a new algorithm called ARMADA for discovering richer temporal association rules [55]. ARMADA first generates temporal patterns by extending the MEMISP algorithm [31] for mining sequential patterns. After the temporal patterns have been found, they are then can be used to generate temporal rules called *richer temporal association rules*. This work is different from the work of Höppner (2001), which discovers temporal rules from a long sequence of intervals rather than from a set of interval sequences.

4.3 Dealing with the Third Issue: Selecting Patterns

Many rule discovery algorithms in data mining generate a large number of rules, often greatly exceeds the size of the underlying database, and only a small fraction of that large volume of rules is of any interest to the user [20]. It is generally understood that interpreting the discovered rules to gain a good understanding of the domain is an important phase of the knowledge discovery process. However, when there are a large number of generated rules, identifying and analysing interesting rules become difficult. Providing the user with a list of rules ranked by their confidence and support might not be a good way of organizing the set of rules as this method would overwhelm the user. Besides, not all rules with high confidence and support are interesting [23]. Therefore, to be useful, a data mining system must manage the large amount of generated rules by offering flexible tools for further selecting rules.

Several approaches for post-processing discovered association rules have been reported. One approach is to group ‘similar’ rules among the discovered rules [51, 28, 32, 33]. This approach works well for a moderately large number of rules. However, for a larger number of rules it will produce too many clusters. A more flexible approach is by identifying the rules that are of special importance to the user through templates or *data mining queries*. This approach can complement the rule grouping approach. Templates have been used to specify interesting and uninteresting classes of rules (association and episode rules) [23]. Several query languages have been proposed in the literature that provide the user the ability to query the data, for example, Mine-Rule [37], DMQL (Han *et al.* 1996), and OLE DB [42]. However, these languages are not designed for querying the discovered rules, but for generating rules from the data. A more powerful data mining query language called MSQL is proposed, which can be used not only for rule generation, but also for querying the discovered rules (using SelectRules operator) [21]. Tuzhilin & Liu (2002) have proposed Rule-QL for querying multiple sets of association rules and developed efficient algorithms for processing the queries.

We especially deal with the post-processing problem of richer temporal association rules by developing a retrieval system that can be utilized for facilitating the selection of interesting rules during the post-processing of a large set of discovered richer temporal association rules [54]. We propose a query language TAR-QL for specifying the criteria of rules to be retrieved. The query language specification is based on the subset of Rule-QL, extended by adding several additional operators and functions to deal with richer temporal association rules. Furthermore, we develop low-level methods for evaluating queries involving rule format conditions. In order to improve the performance of the methods, the indexing technique based on signature files is proposed.

5. CONCLUSION

In this paper we have discussed the discovery of temporal patterns and rules from sequential data. Numerous previous studies have concentrated on the discovery of patterns in event sequence data, such as the discovery of temporal association rules, sequential patterns, episodes, and periodic patterns. In contrast, despite its promising applications, the pattern discovery from interval sequence data has received little attention, and many issues have not been fully explored. In this paper, three main issues related to the pattern discovery from interval sequence data and our proposed solutions for these issues have been described.

REFERENCES

- [1] Rakesh Agrawal, Thomasz Imielinski, and Arun Swami. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD'93)*, pages 207–216, May 1993.
- [2] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB'94)*, pages 487–499, September 1994.
- [3] Rakesh Agrawal and Ramakrishnan Srikant. Mining sequential patterns. In Philip S. Yu and Arbee S. P. Chen, editors, *Proceedings of the 11th International Conference on Data Engineering (ICDE'95)*, pages 3–14, Taipei, Taiwan, 1995. IEEE Computer Society Press.
- [4] Juan M. Ale and Gustavo H. Rossi. An approach to discovering temporal association rules. In *Proceedings of the 2000 ACM Symposium on Applied Computing*, pages 294–300, 2000.
- [5] Claudia M. Antunes and Arlindo L. Oliveira. Temporal data mining: An overview. In *Proceedings of the KDD'01 Workshop on Temporal Data Mining*, pages 1–13, San Francisco, USA, 2001.
- [6] Michael H. Böhlen, Renato Busatto, and Christian S. Jensen. Point-versus interval-based temporal data models. In *Proceedings of the 14th International Conference on Data Engineering (ICDE'98)*, pages 192–200, Orlando, Florida, USA, 1998. IEEE Computer Society Press.
- [7] Aaron Ceglar and John F. Roddick. Association mining. *ACM Computing Surveys*, 38(2), 2006.
- [8] Xiaodong Chen and Ilias Petrounias. An integrated query and mining system for temporal association rules. In Yahiko Kambayashi, Mukesh K. Mohania, and A. Min Tjoa, editors, *Proceedings of the 2nd International Conference on Data Warehousing and Knowledge Discovery (DaWaK'00)*, pages 327–336, London, UK, 2000.

- [9] Gautam Das and Dimitrios Gunopulos. Time series similarity and indexing. *Invited Chapter in Handbook on Data Mining*, 2003.
- [10] T. G. Dietterich and R. S. Michalski. Discovering patterns in sequences of events. *Artificial Intelligent*, 25:187–232, 1985.
- [11] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. The KDD process for extracting useful knowledge from volumes of data. *Communication of the ACM*, 39(11):27–34, 1996.
- [12] C. Goutte, P. Toft, and E. Rostrup. On clustering fMRI time series. *Neuroimage*, 9(3):298–310, 1999.
- [13] G Guimarães and A. Ultsch. A method for temporal knowledge conversion. In *Proceedings of the 3rd International Conference in Intelligent Data Analysis (IDA '99)*, pages 369–380. Springer, 1999.
- [14] J. Han, W. Gong, and Y. Yin. Mining segment-wise periodic patterns in time-related databases. In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*, pages 214–218. AAAI Press, 1998.
- [15] J. Han and M. Kamber. *Data Mining : Concepts and Techniques*. Academic Press, San Diego, 2001.
- [16] J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M.-C. Hsu. FreeSpan: Frequent pattern-projected sequential pattern mining. In *Proceedings of the 2000 International Conference on Knowledge Discovery and Data Mining (KDD'00)*, pages 20–23, Boston, MA, USA, August 2000.
- [17] Jiawei Han, Jenny Y. Chiang, Sonny Chee, Jianping Chen, Qing Chen, Shan Cheng, Wan Gong, Micheline Kamber, Krzysztof Koperski, Gang Liu, Yijun Lu, Nebojsa Stefanovic, Lara Winstone, Betty B. Xia, Osmar R. Zaiane, Shuhua Zhang, and Hua Zhu. DBMiner: a system for data mining in relational databases and data warehouses. In *Proceedings of the International Conference on Knowledge Discovery in Databases and Data Mining (KDD'96)*, 1996.
- [18] Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. In Weidong Chen, Jeffrey Naughton, and Philip A. Bernstein, editors, *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD'00)*, pages 1–12, May 2000.
- [19] Frank Höppner. Learning temporal rules from state sequence. In *Proceedings of IJCAI'01 Workshop on Learning from Temporal and Spatial Data*, pages 25–31, Seattle, USA, 2001.
- [20] Tomasz Imielinski and Aashu Virmani. Association rules... and what's next? towards second generation data mining systems. In Witold Litwin, Tadeusz Morzy, and Gottfried Vossen, editors, *Proceedings of the 2nd East European Symposium on Advances in Databases and Information Systems (ADBIS'98)*, LNCS, pages 6–25, Poznan, Poland, 1998. Springer.

- [21] Tomasz Imielinski and Aashu Virmani. MSQL: A query language for database mining. *Data Mining and Knowledge Discovery*, 3(4):373–408, 1999.
- [22] Po-Shan Kam and Ada Wai-Chee Fu. Discovering temporal patterns for interval-based events. In Yahiko Kambayashi, Mukesh K. Mohania, and A. Min Tjoa, editors, *Proceedings of the 2nd International Conference on Data Warehousing and Knowledge Discovery (DaWaK'00)*, volume 1874 of *LNCS*, pages 317–326, London, UK, 2000. Springer.
- [23] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A.I. Verkamo. Finding interesting rules from large sets of discovered association rules. In N.R. Adam, B.K. Bhargava, and Y. Yesha, editors, *Proceedings of the 3rd International Conference on Information and Knowledge Management*, pages 401–407, Gaithersburg, Maryland, 1994. ACM Press.
- [24] T. Koskela, M. Lehtokangas, J. Saarinen, and K. Kaski. Time series prediction with multilayer perceptron, FIR, and Elman neural networks. In *Proceedings of World Congress on Neural Network*, pages 491–496, 1996.
- [25] A. Kundu, Y. He, and P. Bahl. Word recognition and word hypothesis generation for handwritten script: A Hidden Markov Model based approach. In *Proceedings of 1988 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'88)*, pages 457–462, 1988.
- [26] S. Laxman and P. S. Sastry. A survey of temporal data mining. *Sādhanā*, 31:173–198, 2006.
- [27] Chang-Hung Lee, Cheng-Ru Lin, and Ming-Syan Chen. On mining general temporal association rules in a publication database. In *Proceedings of the 1st IEEE International Conference on Data Mining*, 2001.
- [28] Brian Lent, Arun N. Swami, and Jennifer Widom. Clustering association rules. In W. A. Gray and Per-Åke Larson, editors, *Proceedings of the 13th International Conference on Data Engineering (ICDE'97)*, pages 220–231. IEEE Computer Society, 1997.
- [29] Yingjiu Li, Peng Ning, Xiaoyang Sean Wang, and Sushil Jajodia. Discovering calendar-based temporal association rules. In *Proceedings of the 8th International Symposium on Temporal Representation and Reasoning*, pages 111–118, 2001.
- [30] T. Warren Liao. Clustering of time series data - a survey. *Pattern Recognition*, 38(11):1857–1874, 2005.
- [31] Ming-Yen Lin and Suh-Yin Lee. Fast discovery of sequential patterns by memory indexing. In *Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery (DaWaK'02)*, pages 150–160, Aix-en-Provence, France, September 2002.

- [32] Bing Liu, Wynne Hsu, and Yiming Ma. Pruning and summarizing the discovered associations. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'99)*, pages 125–134, San Diego, CA, USA, 1999.
- [33] Bing Liu, Mingqing Hu, and Wynne Hsu. Multi-level organization and summarization of the discovered rules. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'00)*, New York, NY, USA, 2000. ACM Press.
- [34] H. Mannila, H. Toivonen, and A. Verkamo. Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery*, 1(3):259 – 289, 1997.
- [35] Heikki Mannila and Hannu Toivonen. Discovering generalised episodes using minimal occurrences. In Evangelos Simoudis, Jiawei Han, and Usama Fayyad, editors, *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD'96)*, pages 146–151, Portland, Oregon, 1996. AAAI Press.
- [36] Heikki Mannila, Hannu Toivonen, and A. Inkeri Verkamo. Discovering frequent episodes in sequences. In Usama M. Fayyad and Ramasamy Uthrusamy, editors, *Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining (KDD'95)*, pages 210–215, Montreal, Canada, 1995. AAAI Press.
- [37] Rosa Meo, Giuseppe Psaila, and Stefano Ceri. A new SQL-like operator for mining association rules. In T. M. Vijayaraman, Alejandro P. Buchmann, C. Mohan, and Nandlal L. Sarda, editors, *Proceedings of the 22th International Conference on Very Large Data Bases (VLDB'96)*, pages 122–133, 1996.
- [38] Fabian Mörchén. A better tool than Allen's relations for expressing temporal knowledge in interval data. In *Proceedings of the Workshop on Theory and Practice of Temporal Data Mining (TPTDM'06)*, pages 25–34, Philadelphia, PA, USA, 2006.
- [39] Fabian Mörchén and Alfred Ultsch. Mining hierarchical temporal patterns in multivariate time series. In S Biundo, T. W. Frühwirth, and G Palm, editors, *Proceedings of the 27th Annual German Conference in Artificial Intelligence*, pages 127–140. Springer, 2004.
- [40] Fabian Mörchén, Alfred Ultsch, and Olaf Hoos. Discovering interpretable muscle activation patterns with the temporal data mining method. In Jean-François Boulicaut, Floriana Esposito, Fosca Giannotti, and Dino Pedreschi, editors, *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'04)*, volume 3202 of *LNCS*, pages 512–514, Pisa, Italy, 2004. Springer.
- [41] V. S. Nalwa. Automatic on-line signature verification. *Proceedings of the IEEE*, 85:215–239, 1997.

- [42] Amir Netz, Surajit Chaudhuri, Usama M. Fayyad, and Jeff Bernhardt. Integrating data mining with SQL databases: OLE DB for data mining. In *Proceedings of the 17th International Conference on Data Engineering (ICDE'01)*, pages 379–387, 2001.
- [43] Banu Ozden, Sridhar Ramaswamy, and Abraham Silberschatz. Cyclic association rules. In *Proceedings of the 14th International Conference on Data Engineering (ICDE'98)*, pages 412–421, Orlando, Florida, USA, 1998. IEEE Computer Society Press.
- [44] Panagiotis Papapetrou, George Kollios, Stan Sclaroff, and Dimitrios Gunopulos. Discovering frequent arrangements of temporal intervals. In *Proceedings of the 5th IEEE International Conference on Data Mining (ICDM'05)*, pages 354–361, Houston, Texas, USA, 2005. IEEE Computer Society Press.
- [45] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M-C. Hsu. PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth. In *Proceedings of the 17th International Conference on Data Engineering (ICDE'01)*, pages 215–224, Heidelberg, Germany, April 2001. IEEE Computer Society Press.
- [46] Chris P. Rainsford and John F. Roddick. Adding temporal semantics to association rules. In Jan M. Zytkow and Jan Rauch, editors, *Proceedings of the 3rd European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'99)*, pages 504–509, 1999.
- [47] Chris P. Rainsford and John F. Roddick. Visualisation of temporal interval association rules. In *Proceedings of 2nd International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'00)*, volume 1983 of *LNCs*, pages 91–96, Shatin, N.T., Hong Kong, 2000. Springer.
- [48] Sridhar Ramaswamy, Sameer Mahajan, and Abraham Silberschatz. On the discovery of interesting patterns in association rules. In *Proceedings of the 24th International Conference on Very Large Data Bases*, pages 368–379, 1998.
- [49] John F. Roddick and Myra Spiliopoulou. A survey of temporal knowledge discovery paradigms and methods. *IEEE Transactions on Knowledge and Data Engineering*, 14(4):750–767, 2002.
- [50] Ramakrishnan Srikant and Rakesh Agrawal. Mining sequential patterns: Generalizations and performance improvements. In *Proceedings of the 5th International Conference on Extending Database Technology*, pages 3–17, Avignon, France, 1996.
- [51] H. Toivonen, M. Klemettinen, P. Ronkainen, K. Hatonen, and H. Mannila. Pruning and grouping of discovered association rules. In *ECML-95 Workshop on Statistics, Machine Learning, and Knowledge Discovery in Databases*, pages 47–52, Heraklion, Greece, 1995.

- [52] Alexander Tuzhilin and Bing Liu. Querying multiple sets of discovered rules. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'02)*, pages 52–60, New York, NY, USA, 2002. ACM Press.
- [53] Roy Villafane, Kien A. Hua, Duc Tran, and Basab Maulik. Knowledge discovery from series of interval events. *Journal of Intelligent Information Systems*, 15(1):71–89, 2000.
- [54] Edi Winarko. The discovery and retrieval of temporal rules in interval sequence data, 2007. PhD Dissertation, School of Informatics and Engineering, Flinders University, Australia.
- [55] Edi Winarko and John F. Roddick. ARMADA - an algorithm for discovering richer relative temporal association rules from interval-based data. *Data and Knowledge Engineering*, 63:76–90, 2007.
- [56] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using Hidden Markov Model. In *Proceedings of 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'92)*, pages 379–385, Champaign, IL, 1992.
- [57] Mohammed Javeed Zaki. Efficient enumeration of frequent sequences. In *Proceedings of the 7th International Conference on Information and Knowledge Management (CIKM'98)*, pages 68–75, Washington DC, USA, 1998.
- [58] Geraldo Zimbrão, Jano Moreira de Souza, Victor Teixeira de Almeida, and Wanderson Araujo da Silva. An algorithm to discover calendar-based temporal association rules with item's lifespan restriction. In *Proceedings of the 2nd Workshop on Temporal Data Mining*, 2002.

EDI WINARKO: Department of Mathematics, Gadjah Mada University, Yogyakarta, Indonesia.

E-mail: edwin@ugm.ac.id

Contributed Papers

Part I: Algebra and Combinatorics

Part II: Analysis

Part III: Applied Mathematics

Part IV: Computer Science

Part V: Statistics

PART I

Algebra and Combinatorics

SUFFICIENT AND NECESSARY CONDITIONS OF A BOUNDED KRULL ORDER

SANTI IRAWATI, HIDETOSHI MARUBAYASHI

Abstract. Let R be an order with an automorphism σ in a simple Artinian ring Q . It is proved in [1, (2.2)] that $S = R[x, \sigma]$ is a Krull order if and only if R is a σ -Krull. Considering this idea, it will be investigated the characterization of $S = R[x, \sigma]$ to be bounded Krull order in the case that R is Noetherian prime Goldie ring.

Key words and Phrases : Krull, Noetherian prime Goldie ring.

1. INTRODUCTION

Let C be a multiplicatively closed subset of a ring R . We say that R satisfies the right Ore condition with respect to C or that C is called a right Ore set of R if, for any $a \in R$ and $c \in C$, there exist $b \in R$ and $d \in C$ such that $ad = cb$. If $C \subseteq C_R(0)$, then it is called a regular right Ore set of R . Similarly, we can define a (regular) left Ore set of R . If C is a (regular) right and left Ore set of R , then it is simply called (regular) Ore set of R .

Let C be a regular right Ore set of a ring R . An overring T of R is called the right quotient ring of R with respect to C if (i) $c \in U(T)$, the set of units of T , for any $c \in C$ and (ii) for any $q \in T$, there exist $a \in R$ and $c \in C$ such that $q = ac^{-1}$. We denote the ring T by R_C . We note that for a multiplicative subset C of R with $C \subseteq C_R(0)$, the right quotient ring R_C of R with respect to C exists if and only if C is a regular right Ore set of R ([9, Chap. 2]).

A subring R of a ring Q is called a right order in Q if Q is the right quotient ring of R with respect to $C_R(0)$, and sometimes we denote the ring Q by $Q(R)$. In particular, R is a right order in Q if and only if $C_R(0)$ is a right Ore set of R . Similarly, we can define a left order in Q and a ring which is both a right and left order in Q is called an order in Q .

An order R in a simple Artinian ring Q is called Krull if there are a family $\{R_i\}_{i \in I}$ and $S(R)$ of overrings of R satisfying the following ([8])

- (1) $R = \bigcap_{i \in I} R_i \cap S(R)$, where R_i and $S(R)$ are essential overrings of R and $S(R)$ is an Asano overring of R ;
- (2) Each R_i is a Noetherian, local, Asano order in Q , and $S(R)$ is a Noetherian simple ring;

- (3) If c is any regular element of R , then $cR_i \neq R_i$ for only finitely many i in I and $R_k c \neq R_k$ for only finitely k in I .

In the case $S(R) = Q$, then R is said to be *bounded Krull order*.

A right R -module M is said to have *finite Goldie dimension* if it contains no infinite direct sum of non-zero R -submodules. For any subset X of R , we set $r_R(X) = \{a \in R \mid Xa = 0\}$ and call it the *right annihilator of X* . Sometimes we denote $r_R(X)$ by $r(X)$. The *left annihilator of X* is defined similarly. A ring R is called a *right Goldie ring* if R satisfies the ascending chain condition (acc) on right annihilators and has finite Goldie dimension as a right R -module. A *left Goldie ring* is defined similarly and R is called a *Goldie ring* if R is a right and left Goldie ring.

Let R be an order with an automorphism σ in a simple Artinian ring Q . [1] proved that $S = R[x, \sigma] = \{a_0 + a_1x + a_2x^2 + \dots + a_nx^n \mid a_i \in R\}$, a skew polynomial ring with $\sigma \in \text{Aut}(R)$ satisfies $xb = \sigma(b)x$ for any $b \in R$, is a Krull order if and only if R is σ -Krull. From this idea, it will be investigated the sufficient and necessary conditions of $S = R[x, \sigma]$ to be a bounded Krull order in the case R is Noetherian prime Goldie ring.

Throughout this paper R will denote a Noetherian prime Goldie ring with identity element, and Q will denote the simple Artinian ring of R .

Let R be an order in a ring Q . A right R -submodule I of Q is called a *right R -ideal of Q* if $I \cap U(Q) \neq \emptyset$ and there exists $c \in U(Q)$ such that $cI \subseteq R$. Similarly, we can define a *left R -ideal of Q* . A right and left R -ideal is called an R -ideal.

A one-sided R -ideal I is said to be σ -stable if $\sigma(I) \subseteq I$, and I is said to be σ -invariant if $\sigma(I) = I$.

For any subsets X and Y of Q , we set $(X : Y)_l = \{q \in Q \mid qY \subseteq X\}$ and $(X : Y)_r = \{q \in Q \mid Yq \subseteq X\}$. For an R -ideal I , we set $I_v = (R : (R : I)_l)_r$ and ${}_v I = (R : (R : I)_r)_l$. I is called a v -ideal if $I_v = I = {}_v I$.

Let P be a prime ideal of a ring S . If $C(P) = \{c \in S \mid c \text{ is regular mod } P\}$ is a regular Ore set of S then the quotient ring $S_{C(P)}$ of S with respect to $C(P)$ is denoted by S_P and is called the *localization of S at P* .

Lemma 1. 1. *Let P be prime v -ideal of $S = R[x, \sigma]$ and $P' = PT$, with $T = Q[x, \sigma]$. Then*

- (i) $C_R(0) \subseteq C(P) \subseteq C(P')$, where $C(P) = \{f(x) \in S \mid f(x) \text{ is regular mod } P\}$ and $C(P') = \{f'(x) \in T \mid f(x) \text{ is regular mod } P'\}$.
- (ii) P' is a maximal ideal of T .
- (iii) $S_P = T_{P'}$, where $S_P = \{f(x) c(x)^{-1} \mid f(x) \in S, c(x) \in C(P)\}$ and $T_{P'} = \{f(x) c(x)^{-1} \mid f(x) \in T, c(x) \in C(P')\}$.

Proof. (i). Let $c \in C_R(0)$ and $f(x) \in S$ such that $c f(x) \in P$. We claim $f(x) \in P$ to show that $c \in C(P)$. Since c is a unit of T , then $f(x) \in f(x) S \subseteq f(x) T = f(x) c T \subseteq PT = P'$.

Therefore, $f(x) \in S \cap P' = P$. It follows that $C_R(0) \subseteq C(P)$. Now, let $f(x) \in C(P)$ and $g'(x) \in T$ such that $f(x)g'(x) \in P'$. We claim $g'(x) \in P'$ to show $C(P) \subseteq C(P')$. For $g'(x) \in T$, there exists a regular element c of R such that $g'(x)c \in S$. We get $f(x)g'(x)c \in S \cap P' = P$. It means $g'(x)c \in P$, since $f(x) \in C(P)$. Hence $g'(x) \in Pc^{-1} \subseteq PT = P'$. It follows that $C(P) \subseteq C(P')$.

(ii). P is a prime ideal of S implies $P' = PT$ is a prime ideal of T . Indeed, let I and J be ideals of T such that $IJ \subseteq P'$. If $A = I \cap S$ and $B = J \cap S$, $I = (I \cap S)T = AT$ and $J = (J \cap S)T = BT$. We get $AB = AB \cap S \subseteq IJ \cap S \subseteq P' \cap S = P$ implies either $A \subseteq P$ or $B \subseteq P$ since P is prime. Hence either $I = AT \subseteq PT = P'$ or $J = BT \subseteq PT = P'$. Moreover, since T is Noetherian and principal ideal ring, then P' is maximal. Furthermore, by [5, (2.4)], P' is localizable.

(iii). First, we show that P is left localizable that is for any $f(x) \in C(P)$ and $g(x) \in S$, there exist $k(x) \in C(P)$ and $l(x) \in S$ such that $l(x)f(x) = k(x)g(x)$. Since $C(P) \subseteq C(P')$ and P' is localizable, there exist $k'(x) \in C(P')$ and $l'(x) \in T$ such that $l'(x)f(x) = k'(x)g(x)$. For $k'(x) \in C(P') \subseteq T$ and $l'(x) \in T$, there exist $r \in C_R(0)$ and $k(x), l(x) \in S$ such that $l'(x) = r l(x)$ and $k'(x) = r k(x)$. So, $r l(x)f(x) = l'(x)f(x) = k'(x)g(x) = r k(x)g(x)$ implies $l(x)f(x) = k(x)g(x)$ because $r \in C_R(0) \subseteq C(P)$. Next, we claim that $k(x) \in C(P)$. Let $s(x) \in S$ such that $k(x)s(x) \in P$. We get $k'(x)s(x) = r k(x)s(x) \in P = S \cap P' \subseteq P'$, which implies $s(x) \in P'$ because $k'(x) \in C(P')$. Hence, $s(x) \in P' \cap S = P$. It follows that $k(x) \in C(P)$ and hence P is left localizable. By similar way, we can prove that P is right localizable and hence P is localizable. Now, we will prove $S_P = T_{P'}$. By its definition, $S_P \subseteq T_{P'}$. Conversely, let $g(x)d(x)^{-1} \in T_{P'}$ where $g(x) \in T$ and $d(x) \in C(P')$. For $d(x) \in C(P') \subseteq T = Q[x, \sigma]$, there exists a regular element $r \in C_R(0)$ such that $g(x)r, d(x)r \in S = R[x, \sigma]$. Furthermore, $d(x)r \in C(P)$ since for any $k(x) \in S$ such that $d(x)r k(x) \in P$ we get $d(x)r k(x) \in P = S \cap P' \subseteq P'$. The condition $d(x) \in C(P')$ implies $r k(x) \in P'$ and hence $k(x) \in P'$ because $r \in C(P')$ by (i). Therefore, $g(x)d(x)^{-1} = g(x)r[d(x)r]^{-1} \in S_P$ since $d(x)r \in C(P)$. We proved $S_P = T_{P'}$.

Let R be a ring and let M be a right R -module. An R -submodule L of M is called *essential* if $L \cap N \neq 0$ for any non-zero R -submodule N of M . By Zorn's lemma, we note that for any R -submodule L of M , there exists an R -submodule L' of M such that $L \cap L' = 0$ and $L \oplus L'$ is essential in M ([9, (2.2.2(v))]). If a right ideal I of R is an essential R -submodule of R , then I is called an *essential right ideal*. A right R -module U is said to be *uniform* if, for any non-zero R -submodule U_1 and U_2 of U , $U_1 \cap U_2 \neq 0$, that is, any non-zero R -submodule of U is an essential R -submodule of U .

A ring R is called *right bounded* if for any essential right ideal I of R , there exists a non-zero ideal A of R such that $A \subseteq I$. A *left bounded* is defined similarly. By a *bounded ring* we mean it is right and left bounded.

Lemma 1.2. *If $S = R[x, \sigma]$ is bounded, then $T = Q[x, \sigma]$ is bounded.*

Proof. Let J be any essential right ideal of $T = Q[x, \sigma]$. Then $I = J \cap S$ is a non-zero essential right ideal of S . Since S is bounded, there exists a non-zero (two-sided) ideal A of S such that $A \subseteq I$. Since T is principal ideal ring, then non-zero two-sided ideal AT of T is contained in $IT = J$. It means T is right bounded. By similar way, we can prove that T is left bounded and hence T is bounded.

By [2, (6.3.4)], it is shown that $T = Q[x, \sigma]$ is bounded if and only if (i) the inner order of σ , $o_Q(\sigma)$, is finite and (ii) $M_n(Q)$, the $n \times n$ matrix ring over Q , is algebraic over $Z(Q)$, the center of Q , for any natural number n . While by [1,

(2.2)], it is proved that $S = R[x, \sigma]$ is a Krull order if and only if R is σ -Krull.

A ring R is called *right σ -bounded* if for any essential right ideal I of R , there exists a non-zero σ -stable (hence σ -invariant since R is Noetherian) ideal A of R such that $A \subseteq I$. A *left σ -bounded* is defined similarly. By a σ -bounded we mean that it is right and left σ -bounded.

2. MAIN RESULT

Theorem 2.1. $S = R[x, \sigma]$ is bounded Krull order if and only if (i) $o_Q(\sigma) < \infty$, (ii) $M_n(Q)$ is algebraic over $Z(Q)$, and (iii) R is σ -Krull and σ -bounded.

Proof. If $S = R[x, \sigma]$ is bounded Krull order, then it is proved that (i), (ii) and R is σ -Krull are satisfied by Lemma 1.2, [2, (6.3.4)] and [1, (2.2)]. Now we will only prove that R is σ -bounded. Let c be regular element of R and P be prime v -ideal of S . If $PT = T$, then $P \cap R = (0)$ and hence $c \in C(P)$ by Lemma 1.1 (i) and $cS_P = S_P$. Since S is Krull order, there exist prime v -ideals $P_1, P_2, \dots, P_n$ of S such that $S = \bigcap S_{P_i}$ and $f(x)S_{P_i} \subseteq S_{P_i}$, for any regular element $f(x)$ of S ([4], (1.8.b)). Since $c \in C(P)$, then $cS_{P_i} \subseteq S_{P_i}$. Thus $A_i = P_i \cap R$ are non-zero σ -ideals and there exists a non-zero σ -stable ideal $A = \bigcap A_i \subseteq (\bigcap cS_{P_i}) \cap R = cS \cap R = cR$. Hence R is right σ -bounded. By similar way, it can be proved that R is left σ -bounded and hence R is σ -bounded.

Conversely, assume that (i) – (iii) are held. Again, by [1, (2.2)], S is Krull order, that is, there exist essential overrings S_i of R which are Noetherian, local, Asano orders in Q and an Asano overring S^0 of R which is Noetherian and simple such that $S = \bigcap S_i \cap S^0$. By [6, (1.1)], $S_i = \bigcap (S_i)_{P_i}$, where P_i are prime ideals of S_i . Beside that, by [5, 3.1], $S^0 = \bigcap S_P \cap S'$ where P ranges over maximal ideals of S and $S' = \bigcup \{B^{-1} : B \text{ is non-zero (two-sided) ideal of } S\}$, $B^{-1} = \{q \in Q(S) : BqB \subseteq B\}$. To prove S is bounded, we shall show that $S' = Q(S)$. By its definition, $S' \subseteq Q(S)$. Now, let $f(x)g(x)^{-1} \in Q(S)$ where $f(x), g(x) \in S$ and $g(x)$ is regular element of S (and hence of T). From (i) and (ii), we get $T = Q[x, \sigma]$ is bounded by [2, (6.3.4)]. Hence, there exists a non-zero (two-sided) ideal D' of T such that $D' \subseteq g(x)T$ or $g(x)^{-1}D' \subseteq T$. So, we have $f(x)g(x)^{-1}D' \subseteq f(x)T \subseteq T$. If $D = D' \cap S$, then D is an ideal of S and $DT = D'$. So, we have $f(x)g(x)^{-1}D \subseteq f(x)g(x)^{-1}D' \subseteq T$. Since S is Noetherian, then every ideal of S is finitely generated by [9, (1.7.iii)]. So, $D = d_1(x)S + d_2(x)S + \dots + d_m(x)S$ for some integer m and $d_i(x) \in D$. From $f(x)g(x)^{-1}D \subseteq T$, there exists a regular element $c \in C_R(0)$ such that $c f(x)g(x)^{-1}d_i(x) \in S$ or $c f(x)g(x)^{-1}D \subseteq S$. Since R is σ -bounded and $c \in C_R(0)$, there exists a non-zero (two-sided) σ -stable ideal A of R such that $A \subseteq Rc$. Hence $B = A[x, \sigma]$ is a non-zero ideal of $S = R[x, \sigma]$ such that $B \subseteq Sc$. So, $B f(x)g(x)^{-1}D \subseteq S c f(x)g(x)^{-1}D \subseteq S$ or $f(x)g(x)^{-1}D \subseteq B^{-1}S = B^{-1}$. Hence, $f(x)g(x)^{-1}DB \subseteq (f(x)g(x)^{-1}DB)_v \subseteq (B^{-1}B)_v = S$. Since DB is a non-zero (two-sided) ideal of S , then $f(x)g(x)^{-1} \in S'$. It means $Q(S) \subseteq S'$. We proved $S' = Q(S)$ which means $S = R[x, \sigma]$ is bounded.

Open problems.

The skew polynomial ring (or Ore extension) $R[x, \sigma] =$

$\{a_0 + a_1x + a_2x^2 + \dots + a_nx^n \mid a_i \in R\}$ with $\sigma \in \text{Aut}(R)$ satisfies $xb = \sigma(b)x$ for any $b \in R$ is a special case (that is $\delta = 0$) of $R[x; \sigma, \delta] = \{a_0 + a_1x + a_2x^2 + \dots + a_nx^n \mid a_i \in R\}$ where $\sigma \in \text{Aut}(R)$ and δ is a (left) σ -derivation of R satisfy $xb = \sigma(b)x + \delta(b)$ for any $b \in R$. [2] proved that if $S = R[x; \sigma, \delta]$ is bounded then $T = Q[x; \sigma, \delta]$ is also bounded. This remains a question to get the sufficient and necessary conditions of $S = R[x; \sigma, \delta]$ to be bounded Krull.

REFERENCES

1. G.Q. Abbasi, Kobayashi S., Marubayashi H., Ueda A., Non-commutative unique factorization rings. *Comm. Algebra* **19**(1), 167 – 198., 1991.
2. T. Cauchon, *Les T-anneaux et les anneaux a identities polynomials noetheriens*, These, Orsay, 1977.
3. A.W. Chatters and C.R.Hajarnavis. *Rings with chain condition*, USA: Pitman Advanced Publishing Program, 1980.
4. M.Chamarie, Anneaux de Krull non commutatifs, *J. Algebra* **72**, 210 – 222., 1981.
5. Hajarnavis and Lenagan, Localization in Asano orders, *J. Algebra* **21**, 441 – 449., 1972.
6. Marubayashi H., Non-commutative Krull rings, *Osaka J. Math* **12**, 703 – 714., 1975.
7. Marubayashi H., On bounded Krull prime rings, *Osaka J. Math* **13**, 491 – 501., 1976.
8. Marubayashi H., Polynomial rings over Krull orders in simple Artinian rings, *Hokkaido Math. Journal* vol. **9**, 63 – 78., 1980.
9. J.C. McConnell and J.C. Robson, *Non-commutative Noetherian rings*, John Wiley & Sons, New York, 1987.

Santi Irawati: Mathematics Department, State University of Malang (UM), Malang, Jawa Timur, Indonesia.
E-mail: santira99@yahoo.com

Hidetoshi Marubayashi: Mathematics Department, Tokushima Bunri University, Kagawa, Japan.
E-mail: marubaya@naruto-u.ac.jp

FUZZY VERSION OF THE FUNDAMENTAL THEOREMS OF HOMOMORPHISM GROUPS

KARYATI

Abstract. Using the classical homomorphism group, we will investigate the properties of this homomorphism group in fuzzy setting relate to the fuzzy kernel of homomorphism group, especially about fuzzy version of the fundamental theorems of homomorphism groups.

If $f : G \rightarrow G'$ is a group homomorphism of a group G onto G' , then the following are held: $\mu < G \Rightarrow f(\mu) < G'$, $\mu \triangleleft G \Rightarrow f(\mu) \triangleleft G'$, $\eta < G' \Rightarrow f^{-1}(\eta) < G$ and $\eta \triangleleft G' \Rightarrow f^{-1}(\eta) \triangleleft G$. If $f : G \rightarrow G'$ is an onto group homomorphism with kernel K , then the mapping $\mu \rightarrow f(\mu)$ induces the one-one correspondence $A \rightarrow f(A)$ between the family F of subgroup of G containing the kernel K and the family F' of subgroup of G' . If $f : G \rightarrow G'$ is an onto group homomorphism with kernel K , then the mapping $\mu \rightarrow f(\mu)$ is one correspondence between the family τ_t of all fuzzy subgroups which have the support t at e , the identity of G and containing K_t and the family τ'_t of all those fuzzy subgroup of G' which have support t at e' the identity of G' . If $f : G \rightarrow G'$ is an onto group homomorphism with kernel K , μ is a fuzzy subgroup of G , the map $\mu|_K : G/K \rightarrow [0,1]$ given by $\mu|_K(xK) = \sup_{k \in K} \{\mu(xk)\}$, for $x \in G$ is a fuzzy subgroup of G/K , and is called the fuzzy quotient group of μ relative to K .

Key Words: group homomorphism, fuzzy subgroup, fuzzy kernel, fuzzy quotient group

1. INTRODUCTION

The fundamental concept of a fuzzy set was first introduced by Zadeh in 1965, Aktaş (2004), Shabir (2004), has been applied by many authors to extend of the basic notion of structure algebra. Group is one of the structure algebra that is very interesting to be extended. Based on the fuzzy set theory, the group theory was generalized to fuzzy subgroup theory. Using the classical homomorphism group, we will investigate the properties of this homomorphism group in fuzzy setting relate to the fuzzy kernel of homomorphism group, especially about fuzzy version of the fundamental theorems of homomorphism groups.

Let S be an ordinary set, μ is called a fuzzy set of S if μ is any mapping from S to the interval $[0,1]$. Let G be a group. A fuzzy subset μ of a group G is called a fuzzy subgroup of G if these conditions are held:

(i) $\mu(xy) \geq \min(\mu(x), \mu(y))$ for all $x, y \in G$, and (ii) $\mu(x) = \mu(x^{-1})$ for all $x \in G$. A fuzzy subgroup μ of G is called a fuzzy normal if $\mu(xy) = \mu(yx)$ for all $x, y \in G$. The level subset μ_t and strong level subset $\mu_t^>$ of μ are defined as follows: (i) $\mu_t = \{x \in S \mid \mu(x) \geq t\}$. (ii) $\mu_t^> = \{x \in S \mid \mu(x) > t\}$. We can prove that μ_t and $\mu_t^>$ are ordinary subgroups for each $t \in [0, \mu(e))$.

We give many important characterizations of a fuzzy subgroup which are going to play a vital role in obtaining the result of this paper. If μ is a fuzzy set in a group G , then μ is a fuzzy subgroup in a group G if and only if each non empty level subset of μ is a subgroup G and μ is a fuzzy subgroup in a group G if and only if each level subset μ_t , $t \in \text{Im } \mu$ is a subgroup G . A fuzzy subgroup μ of G is a fuzzy normal if and only if each non empty level subset μ_t is a normal subgroup of G . If $f: G \rightarrow G'$ is a group mapping, μ is a fuzzy subgroup of G , then for $t \in (0, 1]$, $f(\mu)_t = \bigcap_{t > \varepsilon > 0} f(\mu_{t-\varepsilon})$. If $f: G \rightarrow G'$ is a

group homomorphism, μ is a fuzzy subgroup of G , then the homomorphic image $f(\mu)$ is a fuzzy subgroup of G' . If $f: G \rightarrow G'$ is a surjective group homomorphism, μ is a fuzzy normal subgroup of G , then the homomorphic image $f(\mu)$ is a fuzzy normal subgroup of G' . Two fuzzy subgroups μ and η of groups G and G' respectively, are isomorphic to each other, denoted by $\mu \approx \eta$, if there exist an isomorphism from G onto G' or vice versa such that $\eta = f(\mu)$ or $\mu = f(\eta)$.

Let $f: G \rightarrow G'$ be a group homomorphism with kernel K . A fuzzy subgroup μ of G is said to contain the kernel K if each non empty level subset of μ contains K , for the next we denote the containment $K \subseteq \mu$. If A is a subgroup of G and χ_A is its characteristic function then $K \subseteq \chi_A$ if and only if $A \subseteq K$.

If $f: G \rightarrow G'$ is a group homomorphism and η is a fuzzy subgroup of G' , then for $t \in [0, 1]$, $f^{-1}(\eta)_t = f^{-1}(\eta_t)$. We shall denote a fuzzy subgroup μ of G by $\mu < G$ and a fuzzy normal subgroup by $\mu \triangleleft G$. If $f: G \rightarrow G'$ is a group homomorphism of a group G onto G' , then the following are held: $\mu < G \Rightarrow f(\mu) < G'$, $\mu \triangleleft G \Rightarrow f(\mu) \triangleleft G'$, $\eta < G' \Rightarrow f^{-1}(\eta) < G$ and $\eta \triangleleft G' \Rightarrow f^{-1}(\eta) \triangleleft G$.

2. MAIN RESULTS

Theorem 2.1. (Ajmal N) Let $f: G \rightarrow G'$ be a homomorphism from a group G onto a group G' . Then, the following are true: The mapping $\mu \rightarrow f(\mu)$ is a one-to-one correspondence between the family τ of fuzzy subgroup of G containing the kernel K and the family τ' of the fuzzy subgroup of G' .

Furthermore, fuzzy normal subgroup of G correspond to fuzzy normal subgroup of G' .

Proof:

To prove this theorem, we must prove that:

- a. the mapping $\mu \rightarrow f(\mu)$ is onto

Let $\lambda \in \tau' \Rightarrow f^{-1}(\lambda)$ is a subgroup fuzzy of G

Next, we proof that $f^{-1}(\lambda)$ containing the kernel K :

Let $f^{-1}(\lambda)_t$ be a non empty level subset of $f^{-1}(\lambda)$ then λ_t is a non empty level subset of λ . We know that λ_t is a subgroup of G' , so $f(K) \subseteq \lambda_t$. It implies:

$$K \subseteq f^{-1}f(K) \subseteq f^{-1}(\lambda_t) = f^{-1}(\lambda)_t$$

Thus $K \subseteq f^{-1}(\lambda)$ and $ff^{-1}(\lambda) = \lambda$, so $f^{-1}(\lambda) \in \tau'$. So, the mapping $\mu \rightarrow f(\mu)$ is onto.

- b. the mapping $\mu \rightarrow f(\mu)$ is one-to-one

Let $\mu \in \tau$, such that $K \subseteq \mu$. Thus $K \subseteq \mu_{t_0}$, $t_0 = \mu(e)$ and this expression is held:

$$f^{-1}f(\mu)(x) = f(\mu)(f(x)) = \sup_{x_i \in f^{-1}f(x)} \{\mu(x_i)\}$$

Let $x \in \mu_{t_0}$, such that for $x_i \in f^{-1}f(x)$, $x_i = k_i x$ for $k_i \in K \subseteq \mu_{t_0}$.

We have μ_{t_0} is a subgroup of G since μ is a fuzzy subgroup of G .

Therefore $k_i x \in \mu_{t_0}$, $\mu(x_i) = \mu(x) = t_0$. But, for $x \in G \setminus \mu_{t_0}$ we have

$\mu(x) < \mu(k_i) = t_0$, $\forall k_i \in K$. For $x_i \in f^{-1}f(x)$, the following expression is held:

$$\mu(x_i) = \mu(k_i x) \geq \min\{\mu(k_i), \mu(x)\} = \mu(x)$$

$$\mu(x) = \mu(k_i^{-1} k_i x) \geq \min\{\mu(k_i), \mu(k_i x)\} = \mu(k_i x)$$

So, $\mu(x_i) = \mu(x)$, $\forall x_i \in f^{-1}f(x)$ and we have:

$$f^{-1}f(\mu)(x) = \sup_{x_i \in f^{-1}f(x)} \{\mu(x_i)\} = \mu(x), \forall x \in G$$

■

For the next lemma, χ_A is an ordinary characteristic function of a set A .

Lemma 2.1. (Ajmal N). Let be an onto group homomorphism, let A and B be ordinary subgroup of G and G' , respectively. Then :

- (i) $f(\chi_A) = \chi_{f(A)}$
(ii) $f^{-1}(\chi_B) = \chi_{f^{-1}(B)}$

Proof:

It is the consequence of the Theorem 1, we omit it

Theorem 2.2. Let $f : G \rightarrow G'$ be an onto group homomorphism with kernel K . Then, the mapping $\mu \rightarrow f(\mu)$ induces the one-one correspondence $A \rightarrow f(A)$ between the family F of subgroup of G containing the kernel K and the family F' of subgroup of G' .

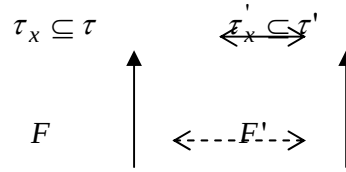
Proof:

Let :

τ_x be the family of characteristics function of the member of F

τ'_x be the corresponding family of F'

See the following diagram:



The one-one correspondence between F and τ_x is obvious, τ_x contains the kernel K , therefore $\tau_x \subseteq \tau$. We obtain a one-one correspondence between τ_x and τ'_x and between τ'_x and F' . Finally, We obtain a one-one correspondence between F and F' .

The following definition of a fuzzy kernel of f in G .

Definition 2.1. Let $f : G \rightarrow G'$ be a group homomorphism with kernel K and

$t \in (0,1]$ be any fixed real number. Then, the fuzzy subset K_t is defined by:

$$K_t(x) = \begin{cases} t & \text{if } x \in K \\ 0 & \text{if } x \in G \setminus K \end{cases}$$

is called a fuzzy kernel of f in G .

A fuzzy subgroup of G is said to contain the fuzzy kernel K_t if $K_t \leq \mu$, where ' $\leq$ ' is the relation of fuzzy set inclusion. Each fuzzy kernel K_t is a fuzzy normal subgroup of G .

Theorem 2.3. (Ajmal N). Let $f : G \rightarrow G'$ be an onto group homomorphism with fuzzy kernel. Then, the mapping $\mu \rightarrow f(\mu)$ is a one-one correspondence

between the family τ_t of all fuzzy subgroup which have the support t at e (the identity of G) and containing the fuzzy kernel K_t and the family τ'_t of all those fuzzy subgroup of G' which have support t at e' (the identity of G')

If G is a group and K is kernel of the homomorphism $f : G \rightarrow G'$, with G' is a group, then $G/K = \{xK | x \in G\}$ is a group. So, we can construct a fuzzy set from G/K to $[0,1]$. If the fuzzy subset is a fuzzy subgroup, then it is called a fuzzy quotient group relative to K . The following theorem gives an example of a fuzzy quotient group.

Theorem 2.4. Let $f : G \rightarrow G'$ be an onto group homomorphism with kernel K , and let μ be a fuzzy subgroup of G . Then the map $\mu|_K : G/K \rightarrow [0,1]$ given by:

$$\mu|_K(xK) = \sup_{k \in K} \{\mu(xk)\}, x \in G$$

is a fuzzy subgroup of G/K , and is called the fuzzy quotient group of μ relative to K .

Proof:

- a. We proof that $\mu|_K(xK) = \sup_{k \in K} \{\mu(xk)\}, x \in G$ is well defined.

Let $xK = yK$, for some $x, y \in G$. We will proof by contradiction method:

$$\text{If } \mu|_K(xK) > \mu|_K(yK), \text{ then } \sup_{k \in K} \{\mu(xk)\} > \sup_{k \in K} \{\mu(yk)\}.$$

Therefore, there is $k_0 \in K$ such that $\mu(xk_0) > \mu(yk), \forall k \in K$.

So, we have $\forall k \in K, yk \notin \mu_t$, where $\mu(xk_0) = t$, hence $\mu_t \cap yK = \emptyset$. On the other hand, $xk_0 \in \mu_t$ and $xk_0 \in xK = yK$, hence $xk_0 \in \mu_t \cap yK$. In the other word, $\mu_t \cap yK \neq \emptyset$. So there is a contradiction and we conclude that

$$\mu|_K(xK) = \mu|_K(yK)$$

- b. We proof that $\mu|_K$ is a fuzzy subgroup of G/K , i.e.:

$$\text{b.1. } \mu|_K(xKyK) \geq \min\{\mu|_K(xK), \mu|_K(yK)\}$$

$$\begin{aligned}
\mu|_K(xKyK) &= \mu|_K(xyK) = \sup_{k \in K} \{\mu(xyk)\} \\
&= \sup_{k_1, k_2 \in K} \{\mu(xk_1yk_2)\} \\
&\geq \sup_{k_1, k_2 \in K} (\min\{\mu(xk_1), \mu(yk_2)\}) \\
&\geq \min \left\{ \sup_{k_1 \in K} \mu(xk_1), \sup_{k_2 \in K} \mu(yk_2) \right\} \\
&= \min \left\{ \mu|_K(xK), \mu|_K(yK) \right\}
\end{aligned}$$

$$\begin{aligned}
\text{b.2. } \mu|_K(xK) &= \mu|_K(-xK) \\
\mu|_K(xK) &= \sup_{k \in K} \mu(xk) = \sup_{k \in K} \mu(-xk) = \mu|_K(-xK) \blacksquare
\end{aligned}$$

Let μ be a fuzzy subgroup and K be a normal subgroup of a group G . Then the fuzzy quotient subgroup $\mu|_K$ is the homomorphic image of μ under the canonical homomorphism from G into G/K . A fuzzy set μ in a group G is a fuzzy subgroup of G if and only if each non empty strong level subset $\mu_t^>$ is a subgroup of G . Non empty strong level subset $\mu_t^>$ of a fuzzy subgroup μ are called strong level subgroup of μ in G . Let $f: G \rightarrow G'$ be a group homomorphism, and let μ be a fuzzy subgroup of G . Then for each $t \in [0,1)$, $f(\mu)_t^> = f(\mu_t^>)$

Theorem 2.5. Let $f: G \rightarrow G'$ be an onto group homomorphism with kernel K , and let μ be a fuzzy subgroup of G . Then for each $t \in [0,1)$,

$$(\mu|_K)_t^> = \mu_t^> K/K$$

Proof

$$\begin{aligned}
\text{Let } xK \in (\mu|_K)_t^> &\Rightarrow \sup_{k \in K} \{\mu(xk)\} > t \Rightarrow \sup_{k \in K} \{\mu(xk_0)\} > t, \text{ for some } k_0 \in K \\
&\Rightarrow xk_0 \in \mu_t^>, \text{ hence } xk_0 \in \mu_t^> K \\
&\Rightarrow xk_0 K = xK \in \mu_t^> K/K \\
&\Rightarrow (\mu|_K)_t^> \subseteq \mu_t^> K/K
\end{aligned}$$

$$\begin{aligned}
\text{Let } xK \in \mu_t^> K/K &\Rightarrow xK = x_0 K, \text{ for some } x_0 \in \mu_t^> \\
&\Rightarrow \mu|_K(xK) = \mu|_K(x_0 K) = \sup_{k \in K} \{\mu(x_0 k)\} > t
\end{aligned}$$

$$\begin{aligned} &\Rightarrow xK \in \left(\mu|_K \right)_t^> \\ &\Rightarrow \left(\mu|_K \right)_t^> \supseteq \mu_t^>K/K \blacksquare \end{aligned}$$

Theorem 2.6. Let $f : G \rightarrow G'$ be an onto group homomorphism with kernel K , and let μ be a fuzzy subgroup of G . Then for each $t \in [0,1)$ the strong level subgroup $\left(\mu|_K \right)_t^>$ and $f(\mu)_t^>$ are isomorphic, i.e.: $\left(\mu|_K \right)_t^> \cong f(\mu)_t^>$

$$\begin{array}{ccc} \mu_t^> & \xrightarrow{\quad} & f(\mu)_t^> \\ \downarrow & \searrow & \\ \left(\mu|_K \right)_t^> & & \end{array}$$

Proof:
Consider this diagram:

$$\begin{array}{ccc} \mu_t^> & \xrightarrow{\quad} & f(\mu)_t^> \\ \downarrow & \searrow & \\ \mu_t^>/(\mu_t^> \cap K) & & \end{array}$$

Here, f is a homomorphism from $\mu_t^>$ onto $f(\mu)_t^>$ with kernel $\mu_t^> \cap K$, by the fundamental theorem of homomorphism :

$$\mu_t^>/(\mu_t^> \cap K) \cong f(\mu)_t^>$$

So, we have $\left(\mu|_K \right)_t^> = \mu_t^>K/K$, and based on the second isomorphism theorem ,

we have $\mu_t^>/(\mu_t^> \cap K) \cong \mu_t^>K/K$. Therefore we get $\left(\mu|_K \right)_t^> \cong \mu_t^>/(\mu_t^> \cap K)$.

On the other hand, $f(\mu)_t^> = f(\mu_t^>)$, so we have $\mu_t^>/(\mu_t^> \cap K) \cong f(\mu_t^>)$.

Finally, we have $\left(\mu|_K \right)_t^> \cong f(\mu)_t^> \blacksquare$

Finally, if $f : G \rightarrow G'$ is an onto group homomorphism with kernel K , then the fuzzy quotient group $\mu|_K$ is isomorphic with the homomorphic image $f(\mu)$ of μ , i.e. $\mu|_K \approx f(\mu)$.

3.CONCLUDING REMARK

If $f : G \rightarrow G'$ is a group homomorphism of a group G onto G' , then the following are held: $\mu < G \Rightarrow f(\mu) < G'$, $\mu \triangleleft G \Rightarrow f(\mu) \triangleleft G'$, $\eta < G' \Rightarrow f^{-1}(\eta) < G$ and $\eta \triangleleft G' \Rightarrow f^{-1}(\eta) \triangleleft G$. If $f : G \rightarrow G'$ is an onto group homomorphism and A, B are subgroups of G, G' respectively, then $f(\chi_A) = \chi_{f(A)}$ and $f^{-1}(\chi_B) = \chi_{f^{-1}(B)}$, which χ_A, χ_B are characteristic functions of A, B respectively. If $f : G \rightarrow G'$ is an onto group homomorphism with kernel K , then the mapping $\mu \rightarrow f(\mu)$ induces the one-one correspondence $A \rightarrow f(A)$ between the family F of subgroup of G containing the kernel K and the family F' of subgroup of G' .

We define a notion of fuzzy kernel of a group homomorphism, i.e.: If $f : G \rightarrow G'$ is a group homomorphism with kernel K and $t \in (0,1]$, then the fuzzy subset K_t is defined by:

$$K_t(x) = \begin{cases} t & \text{if } x \in K \\ 0 & \text{if } x \in G \setminus K \end{cases}$$

A fuzzy subgroup μ of G is said to contain the fuzzy kernel K_t if $K_t \leq \mu$, where $\leq$ is the relation of fuzzy set inclusion. We get : If $f : G \rightarrow G'$ is an onto group homomorphism with kernel K , then the mapping $\mu \rightarrow f(\mu)$ is one correspondence between the family τ_t of all fuzzy subgroups which have the support t at e , the identity of G and containing K_t and the family τ'_t of all those fuzzy subgroup of G' which have support t at e' the identity of G' . If $f : G \rightarrow G'$ is an onto group homomorphism with kernel K , μ is a fuzzy subgroup of G , then the map $\mu|_K : G/K \rightarrow [0,1]$ given by

$$\mu|_K(xK) = \sup_{k \in K} \{\mu(xk)\}, \text{ for } x \in G \text{ is a fuzzy subgroup of } G/K, \text{ and is called the fuzzy}$$

quotient group of μ relative to K

REFERENCES

- Ajmal, Naseem. 1994. Homomorphism of Fuzzy groups, Correspondence Theorem and Fuzzy Quotient Groups. *Fuzzy Sets and Systems* 61, p:329-339. North-Holland
- Aktaş, Hacı. 2004. On Fuzzy Relation and Fuzzy Quotient Groups. *International Journal of Computational Cognition* Vol 2, No 2, p: 71-79
- Asaad, Mohamed. 1991. Group and Fuzzy Subgroup. *Fuzzy Sets and Systems* 39, p:323-328. North-Holland
- Klir, G.J, U.S. Clair, , B. Yuan, 1997. *Fuzzy Set Theory: Foundation and Applications*. Prentice-Hall, Inc. USA
- J.N. Mordeson, Malik, D.S. 1998. *Fuzzy Commutative Algebra*. World Scientifics Publishing Co. Pte. Ltd. Singapore

KARYATI

M. Shabir, 2005. Fully Fuzzy Prime Semigroups. *International Journal of Mathematics and Mathematical Sciences*:1 p:163-168

Karyati: Mathematics Education Department, Faculty of Mathematics and Natural Sciences, Yogyakarta State University, Indonesia.
E-mail: yatiuny@yahoo..Com

A TUTORIAL AND SURVEY ON CODES OVER RINGS

MAHESHANAND AND SIRI KRISHAN WASAN

Abstract. There has been a lot of research on codes over rings since the revelation that certain nonlinear binary codes with very good parameters are actually the binary images under a certain map, of some linear codes over $\mathbb{Z}_4$, the ring of integers modulo 4. This gives an indication that codes with better parameters can be obtained over rings. Further, codes over rings may be useful in certain applications where the finite field arithmetic is not well suited. We survey here some classical and some recent constructions of algebraic codes over rings.

Keywords: Linear codes, group rings, Galois rings, gray map, Hensel's lift.

1. INTRODUCTION

The theory of algebraic codes has been mainly studied over an alphabet when the alphabet is a finite field. However, coding theorists for a long time have felt that algebraic codes could be defined over a less structured alphabet like a finite ring. As a result, several constructions of codes over rings have been obtained. Constructing codes over rings is also important, because in certain applications, like computer to computer communication, codes over finite rings may be more appropriate. Another important reason for investigating codes over finite rings is that, codes with better parameters can be obtained. It has already been shown that some well-known nonlinear binary codes with very good parameters are in fact the binary images under a certain map of some linear codes over $\mathbb{Z}_4$, the ring of integers modulo 4 [4]. Further, codes over rings are continuously finding newer applications.

In this tutorial and survey we consider some important constructions of algebraic codes over rings. Looking at the amount of research done in this area in recent years, this survey is a modest survey. We have considered two approaches for constructing codes over rings. The first approach is the group ring (group algebra) approach in which codes are considered as ideals in the group rings. In the second one, construction of codes over some Galois rings is considered. The latter approach has been very popular in recent years. Many other constructions of codes over rings also have been studied in recent years. Noted among them are codes over more general Galois rings [15] codes over finite chain ring [10] and codes over quasi-Frobenius rings [17].

2. PRELIMINARIES

In this section we discuss some preliminaries of algebraic codes over rings.

Let R be a finite commutative ring with identity. A linear code $\mathcal{C}$ of length n over R is a submodule of the R -module R^n . Such a submodule is not necessarily a free module; a module with a basis. We call a matrix to be a generator matrix of a linear code $\mathcal{C}$ over R if the set of the rows of the matrix generate $\mathcal{C}$ and no proper subset of this set generates $\mathcal{C}$. We define the Hamming distance and Hamming weight in R^n , similarly as in the finite field case. In the study of codes over rings another weight function, called the Lee weight, is sometimes more useful. The Lee weight of an n -tuple $\mathbf{x} = (a_1, a_2, \dots, a_n)$, where a_i are chosen from the set $S = \{0, 1, 2, \dots, m-1\}$ is defined by

$$w_L(\mathbf{x}) = \sum |a_i| \quad \text{where, } |a_i| = \min \{a_i, m - a_i\}$$

The Lee distance between two words $\mathbf{x}$ and $\mathbf{y}$ of S^n is defined by $d_L(\mathbf{x}, \mathbf{y}) = w_L(\mathbf{x} - \mathbf{y})$.

The natural inner product $\mathbf{x} \cdot \mathbf{y}$, of two elements $\mathbf{x}, \mathbf{y}$ of R^n is defined as $\mathbf{x} \cdot \mathbf{y} = (x_1, x_2, \dots, x_n) \cdot (y_1, y_2, \dots, y_n) = x_1 y_1 + x_2 y_2 + \dots + x_n y_n$. Two elements $\mathbf{x}, \mathbf{y}$ of R^n , are said to be orthogonal if $\mathbf{x} \cdot \mathbf{y} = \mathbf{0}$. The dual $\mathcal{C}^\perp$ of a code $\mathcal{C}$ is defined by $\mathcal{C}^\perp = \{\mathbf{x} \in R^n \mid \mathbf{x} \cdot \mathbf{c} = \mathbf{0}, \mathbf{c} \in \mathcal{C}\}$. Further if $\mathcal{C} = \mathcal{C}^\perp$ then $\mathcal{C}$ is called a selfdual code.

A linear code $\mathcal{C}$ over R is called a cyclic code if whenever a codeword $(c_0, c_1, \dots, c_{n-1})$ is in $\mathcal{C}$, its cyclic shift $(c_{n-1}, c_0, \dots, c_{n-2})$, is also in $\mathcal{C}$. As in the case of finite fields, a codeword $(c_0, c_1, \dots, c_{n-1})$ in a cyclic code $\mathcal{C}$ over R is often represented by the polynomial $c(x) = c_0 + c_1 x + \dots + c_{n-1} x^{n-1}$ in $R[x]$. It is straightforward to show, that cyclic codes of length n over R are precisely the ideals of the residue class ring $\mathcal{R}_n = R[x]/(x^n - 1)$.

There is another representation of cyclic codes. It is well known that cyclic codes over finite fields can be represented as ideals in the group algebras [8]. Similarly cyclic codes over R can be represented as ideals in the group rings [1, 13, 16]. Let $C_n = \{1, x, x^2, \dots, x^{n-1}\}$, $x^n = 1$, be a cyclic group of order n . Then the group ring RC_n of C_n over R consists of all formal sums $\sum c_i x^i$, $c_i \in R$. Addition in RC_n is coordinatewise and multiplication is modulo $x^n - 1$. Then clearly RC_n coincides with $\mathcal{R}_n$. Therefore; a cyclic code of length n over R is an ideal in the group ring RC_n . This representation of cyclic codes can give some useful information about their structures. To generalize further, if G is finite Abelian group of order n , then a code of length n over R can be defined as an ideal in the group ring RG . Such a code is called an Abelian code of length n over G . Here we formally define the group ring RG of a finite Abelian group G over R .

Definition: The group ring RG of a finite Abelian group G over a ring R is the set of formal sums $\sum f_i g_i$, $f_i \in R$, $g_i \in G$, with addition, scalar multiplication and ordinary multiplication defined by

$$\begin{aligned} \sum r_i g_i + \sum r'_i g_i &= \sum (r_i + r'_i) g_i, \quad r \sum r_i g_i = \sum r r_i g_i, \text{ and} \\ \sum r_i g_i \sum f_j g_j &= \sum_i \sum_j r_i f_j g_i g_j = \sum_k r_k g_k, \text{ where } r_k = \sum_{g_i g_j = g_k} r_i f_j \end{aligned}$$

3. Codes Over $\mathbb{Z}_m$ (The Group Ring Approach)

In this section, we describe codes over $\mathbb{Z}_m$, the ring of integers modulo m , following the group ring approach.

The study of algebraic codes over rings mainly started in early seventies with two important papers by Blake [1, 2]. In [1] Blake gave a construction of cyclic codes over $\mathbb{Z}_m$, where m is a product of distinct primes. In his second paper [2], he described some constructions of Hamming, Reed Solomon and BCH codes over $\mathbb{Z}_m$, where m is an arbitrary positive integer. These codes were natural analogs of such codes over finite fields.

We consider Blake's construction of cyclic codes over $\mathbb{Z}_m$ when m is a product of distinct primes p_i , i.e. $m = \prod_{i=1}^s p_i$, p_i distinct primes, and $(m, n) = 1$. A cyclic code of length n over $\mathbb{Z}_m$ is an ideal of $\mathbb{Z}_m C_n$. Since $m = \prod_{i=1}^s p_i$, therefore by the Chinese Remainder Theorem

$$\mathbb{Z}_m C_n \cong \prod_{i=1}^s \mathbb{Z}_{p_i} C_n$$

The corresponding isomorphism being $a \rightarrow (a(\bmod p_1), a(\bmod p_2), \dots, a(\bmod p_s))$, which is a special case of a more general result given in equation (1). Now each ideal of $\mathbb{Z}_m C_n$ is a direct product of ideals of $\mathbb{Z}_{p_i} C_n$. Since $\mathbb{Z}_{p_i}$ is a finite field with p_i elements and an ideal of $\mathbb{Z}_{p_i} C_n$ is a cyclic code of length n over $\mathbb{Z}_{p_i}$, therefore, every cyclic code of length n over $\mathbb{Z}_m$ is a direct product of cyclic codes over the fields $\mathbb{Z}_{p_i}$, $1 \leq i \leq s$. Now if C_i is an $[n, k_i]$ cyclic code over the field $\mathbb{Z}_{p_i}$, $1 \leq i \leq s$, then $\prod_{i=1}^s C_i$ is a cyclic code of length n over $\mathbb{Z}_m$ with $\prod_{i=1}^s p_i^{k_i}$ codewords. The minimum distance of the code is $\min_i d_i$, where d_i is the minimum distance of the code C_i . Thus when m a product of distinct primes, constructing cyclic codes over $\mathbb{Z}_m$ is straightforward.

Now we discuss Blake's construction [2] of some linear and cyclic codes over $\mathbb{Z}_m$, where m is an arbitrary positive integer. Let $m = \prod_{i=1}^s p_i^{e_i}$ be the canonical factorization of m . Then by the Chinese Remainder Theorem

$$\mathbb{Z}_m C_n = \prod_{i=1}^s \mathbb{Z}_{p_i^{e_i}} C_n \quad (1)$$

The corresponding isomorphism being $a \rightarrow (a(\bmod p_1^{e_1}), a(\bmod p_2^{e_2}), \dots, a(\bmod p_s^{e_s}))$

Every ideal (submodule) of $\mathbb{Z}_m C_n$ is a direct product of ideals (submodules) of $\mathbb{Z}_{p_i^{e_i}} C_n$, $1 \leq i \leq s$. As a cyclic (linear) code of length n is an ideal (submodule) of $\mathbb{Z}_m C_n$. From above identity it is sufficient to consider the

case of $m = p^r$. Note that a submodule of $\mathbb{Z}_m C_n$ is a linear code over $\mathbb{Z}_m$ as $\mathbb{Z}_m C_n$ is also a module over $\mathbb{Z}_m$.

3.1 Hamming Codes Over $\mathbb{Z}_{p^r}$ [2]

We first consider the construction of Hamming codes over $\mathbb{Z}_{p^r}$. Let $q = p^r$. It is well known from ring theory that $\mathbb{Z}_q^*$, the group of units of $\mathbb{Z}_q$, is a cyclic group if either p is an odd prime or $p = 2$ and $r = 1$ or 2 . If we consider the zero element also as a zero divisor then every element of $\mathbb{Z}_q$ is either a unit or a zero divisor. Now, there are $\phi(q) = \phi(p^r) = p^{r-1}(p-1)$ units in $\mathbb{Z}_q$, where ϕ is the Euler's phi function. Therefore there are $q - \phi(q) = p^r - p^{r-1}(p-1) = p^{r-1}$ zero divisors in $\mathbb{Z}_q$.

Now let Z be the set of all elements of $\mathbb{Z}_q^k$ whose components are all zero divisors. There are $p^{(r-1)k}$ such elements as there are p^{r-1} choices of a zero divisor for each component. Define two elements $\mathbf{u}, \mathbf{v} \in \mathbb{Z}_q^k \setminus Z$ to be equivalent if and only if they are linearly dependent. This is an equivalence relation on $\mathbb{Z}_q^k \setminus Z$, and there are $(p^k - 1)/(p - 1)$ equivalence classes of $\mathbb{Z}_q^k \setminus Z$ generated by this relation [2]. Now if a parity check matrix H is formed, by choosing one element from each equivalence class, then the resulting code will have length $n = (p^k - 1)/(p - 1)$. Further the minimum distance of the code is 3, as no two columns of H are linearly dependent and some three columns will be dependent. The rank of the matrix H is k , as rows of H are linearly independent and there are k rows in H . The Hamming codes so constructed are obviously the natural analogs of Hamming codes over fields.

3.2 Reed Solomon Codes Over $\mathbb{Z}_{p^r}$ [2]

To construct natural analogs of Reed Solomon (RS) codes over fields, Blake [2] considered them over $\mathbb{Z}_q$, $q = p^r$, p is an odd prime. Let $\alpha \in \mathbb{Z}_q$ be such that α is a primitive element of $\mathbb{Z}_p$, and let $k = p - 2$, $m = p^{r-1}$. Define the matrix

$$H = \begin{bmatrix} 1 & \alpha^m & \alpha^{2m}, & \dots, & \alpha^{km} \\ 1 & \alpha^{m+1} & \alpha^{2(m+1)}, & \dots, & \alpha^{k(m+1)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \alpha^{m+d-2} & \alpha^{2(m+d-2)}, & \dots, & \alpha^{k(m+d-2)} \end{bmatrix}$$

If we take any $d - 1$ columns of the matrix then the corresponding matrix is of the form

$$\begin{bmatrix} \alpha^{j_1 m} & \alpha^{j_2 m}, & \dots, & \alpha^{j_{d-1} m} \\ \alpha^{j_1(m+1)} & \alpha^{j_2(m+1)}, & \dots, & \alpha^{j_{d-1}(m+1)} \\ \vdots & \vdots & & \vdots \\ \alpha^{j_1(m+d-2)} & \alpha^{j_2(m+d-2)} & \dots, & \alpha^{j_{d-1}(m+d-2)} \end{bmatrix}$$

The determinant of the above matrix is a Vander Monde determinant

$$\alpha^{(j_1+j_2+\dots+j_{d-1})m} \begin{vmatrix} 1 & 1, & \dots, & 1 \\ \alpha^{j_1} & \alpha^{j_2}, & \dots, & \alpha^{j_{d-1}} \\ \vdots & \vdots & & \vdots \\ \alpha^{j_1(d-2)} & \alpha^{j_2(d-2)}, \dots, & \alpha^{j_{d-1}(d-2)} \end{vmatrix}$$

which is equal to $\alpha^{(j_1+j_2+\dots+j_{d-1})m} \prod_{i>k} (\alpha^{j_i} - \alpha^{j_k})$

The matrix is non-singular if and only if the above expression is non-zero, which is true if and only if α and $1 - \alpha^j$ are units. Now since α is a primitive element of $\mathbb{Z}_p$, therefore, $1, \alpha, \dots, \alpha^{p-2}$ are all the non zero elements of $\mathbb{Z}_p$ and $\alpha^j \neq 1$, for $j = 1, 2, \dots, p-2$. Equivalently $1 - \alpha^j \neq 0$, for $j = 1, 2, \dots, p-2$. Since every non zero element of $\mathbb{Z}_p$ is a unit in $\mathbb{Z}_p$, and hence a unit in $\mathbb{Z}_q$, therefore, if we take α a primitive element of $\mathbb{Z}_p$ then every $d-1$ columns of the matrix H are linearly independent. Further, H has $d-1$ rows therefore rank of H is $d-1$.

The null space of H over $\mathbb{Z}_q$, is a code of length $n = (p-1)$, distance d and dimension $n - (d-1) = n - d + 1$. These codes over $\mathbb{Z}_q$ are the analogs of Reed Solomon codes over finite fields.

3.3 BCH Codes Over $\mathbb{Z}_{p^r}$ [2]

To find a BCH code over $\mathbb{Z}_q$, $q = p^r$, consider an extension ring of $\mathbb{Z}_q$ which can be obtained by adjoining to $\mathbb{Z}_q$, a root of a polynomial which is irreducible over $\mathbb{Z}_q$. Let $\mathbb{Z}_q(\delta)$ be such an extension ring of $\mathbb{Z}_q$ where δ is a root of a polynomial irreducible over $\mathbb{Z}_q$. Let $\alpha \in \mathbb{Z}_q(\delta)^*$ of order nm be such that $1 - \alpha^j$, $j = 1, 2, \dots, n-1$ are units in $\mathbb{Z}_q(\delta)$. Now define a BCH code over $\mathbb{Z}_q$ of length n and designed distance d as the null space over $\mathbb{Z}_q$, of the matrix

$$H = \begin{bmatrix} 1 & \alpha^m & \alpha^{2m}, & \dots, & \alpha^{(n-1)m} \\ 1 & \alpha^m & \alpha^{2(m+1)}, & \dots, & \alpha^{(n-1)(m+1)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \alpha^{m+d-2} & \alpha^{2(m+d-2)}, \dots, & \alpha^{(n-1)(m+d-2)} \end{bmatrix}$$

The above construction depends on the elements $\alpha \in \mathbb{Z}_q(\delta)^*$ such that $1 - \alpha^j$, $j = 1, 2, \dots, n-1$ are units in $\mathbb{Z}_q(\delta)$. This is related to the structure of $\mathbb{Z}_q(\delta)^*$. Therefore to proceed systematically in this direction, the knowledge of the structure of $\mathbb{Z}_q(\delta)^*$ is required [2].

3.4 Abelian Codes Over $\mathbb{Z}_m$

Now we consider a construction of Abelian codes over $\mathbb{Z}_m$ where m is an arbitrary positive integer. This construction was given by Spiegel [13]. We know

that an Abelian code of length n over $\mathbb{Z}_m$ is an ideal of the group ring $\mathbb{Z}_m G$ where G is a finite Abelian group of order n . Further if $m = \prod_{i=1}^s p_i^{e_i}$ is the canonical factorization of m then by the Chinese Remainder Theorem, we have

$$\mathbb{Z}_m G = \prod_{i=1}^s \mathbb{Z}_{p_i^{e_i}} G \quad (2)$$

Therefore, like cyclic codes, to study Abelian codes over $\mathbb{Z}_m$, it is sufficient to consider them over $\mathbb{Z}_{p^r}$. We assume that $(n, p) = 1$.

Definition: Let R be a commutative ring with identity. Then the radical of R , denoted $Rad(R)$, is defined to be the intersection of all maximal ideals of R .

Definition: A ring R is called a semisimple ring if $Rad(R) = 0$.

Theorem (1) (Maschke): Let G be a finite group and let F be a field whose characteristic does not divide the order of G . Then the group algebra FG is semisimple.

We consider Spiegel's [13] construction of Abelian codes over $\mathbb{Z}_{p^r}$. In this construction, codes over $\mathbb{Z}_{p^r}$ are obtained by using codes over $\mathbb{Z}_p$. Since $(n, p) = 1$, therefore, by Maschke's theorem $\mathbb{Z}_p G$ is a commutative semisimple ring and hence can be expressed as a direct product of fields. Let $\mathbb{Z}_p G = \prod_{i=1}^N F_i$. Since $\mathbb{Z}_p G$ is a finite ring of characteristic p , therefore F_i also has characteristic p , $i = 1, 2, \dots, N$. Let F_i has p^{n_i} elements. The multiplicative group F_i^* of F_i forms a cyclic group, so there exists a primitive m_i th root of unity ξ_{m_i} over $\mathbb{Z}_p$, such that $F_i = \mathbb{Z}_p(\xi_{m_i})$ with $(m_i, p) = 1$. Therefore,

$$\mathbb{Z}_p G \cong \prod_{i=1}^N \mathbb{Z}_p(\xi_{m_i}) \quad (3)$$

Each F_i is an ideal of $\mathbb{Z}_p G$, and as F_i is a field, all these ideals are minimal ideals. Therefore there are N minimal non-zero ideals of $\mathbb{Z}_p G$, namely, $F_1, F_2, \dots, F_N$, and any ideal of $\mathbb{Z}_p G$ is a direct product of some of these ideals. Therefore there are 2^N ideals in $\mathbb{Z}_p G$. The above procedure immediately fails for p^r if $r > 1$, as $\mathbb{Z}_{p^r} G$ is not semisimple for $r > 1$, and therefore, it cannot be expressed as a direct product of fields. However, a similar result is true. The following is a special case of a more general result [13].

Let G be a finite group of order n with $(n, p) = 1$. Then

$$\mathbb{Z}_{p^r} G \cong \prod_{i=1}^N \mathbb{Z}_{p^r}(\xi_{m_i}) \quad (4)$$

where ξ_{m_i} 's are same as in (3).

There are $r+1$ ideals $\mathbb{Z}_{p^r}(\xi_{m_i}), p\mathbb{Z}_{p^r}(\xi_{m_i}), \dots, p^r\mathbb{Z}_{p^r}(\xi_{m_i}) = 0$ of $\mathbb{Z}_{p^r}(\xi_{m_i})$. Further, every ideal of $\mathbb{Z}_{p^r}G$ is a direct product of ideals of $\mathbb{Z}_{p^r}(\xi_{m_i})$. Thus there are total $(r+1)^N$ ideals of $\mathbb{Z}_{p^r}G$.

The ideals of $\mathbb{Z}_{p^r}G$ can be found explicitly as follows. Find a minimal ideal of $\mathbb{Z}_pG$. This is a finite field say F_1 . Let F_1 contains p^{n_1} elements. Find an integer m_1 such that $p^{n_1} = 1 \pmod{m_1}$ and $p^m \neq 1 \pmod{m_1}$ for $1 < m < m_1$. In other words F_1 is the smallest field over $\mathbb{Z}_p$ containing a primitive m_1 th root of unity. Then $F_1 = \mathbb{Z}_p(\xi_{m_1})$. Similarly find minimal ideals $F_2, F_3, \dots, F_N$ of $\mathbb{Z}_pG$ such that $F_i = \mathbb{Z}_p(\xi_{m_i})$. Then find $r+1$ ideals $\mathbb{Z}_{p^r}(\xi_{m_i}), p\mathbb{Z}_{p^r}(\xi_{m_i}), \dots, p^r\mathbb{Z}_{p^r}(\xi_{m_i}) = 0$ of $\mathbb{Z}_{p^r}(\xi_{m_i})$ for each $i, 1 \leq i \leq N$. Now using (4), all $(r+1)^N$ ideals (Abelian codes) of $\mathbb{Z}_{p^r}G$ can be obtained. As an ideal (Abelian code) of $\mathbb{Z}_mG$, is a direct product of ideals (Abelian codes) of $\mathbb{Z}_{p^{e_i}}G$; following the procedure illustrated above, Abelian codes over $\mathbb{Z}_m$ can be obtained knowing the minimal ideals of $\mathbb{Z}_pG$.

In his second paper [14], extending the work of Blake [2], Spiegel has given a construction of BCH codes over $\mathbb{Z}_{p^r}$ by using group algebras over p -adic fields and their residue class fields.

We have seen that Abelian codes are a generalization of cyclic codes. Further, if $m = \prod_{i=1}^s p_i^{e_i}$, an Abelian code over $\mathbb{Z}_m$ can actually be expressed as a direct product of cyclic codes over $\mathbb{Z}_{p^{e_i}}$, $1 \leq i \leq s$. A factorization of $\mathbb{Z}_mG$, induced by the factorization of $\mathbb{Z}_m$, is given in (2). We now describe another factorization of $\mathbb{Z}_mG$ induced by the factorization of G into cyclic groups. This construction was given by Wasan [16]. Since G is a finite Abelian group, therefore by the fundamental theorem of Abelian groups, G can be expressed as direct product of cyclic groups. Let the factorization of G into cyclic groups be given by $G = \prod_{j=1}^t C_{q_j^{k_j}}$, where $C_{q_j^{k_j}}$ is a cyclic group of order $q_j^{k_j}$. Then it can be shown [16] that

$$\mathbb{Z}_mG = \prod_{i=1}^s \prod_{j=1}^t \mathbb{Z}_{p^{e_i}} C_{q_j^{k_j}}$$

Now an ideal of $\mathbb{Z}_mG$ is a direct product of ideals of $\mathbb{Z}_{p^{e_i}} C_{q_j^{k_j}}$, and an ideal of $\mathbb{Z}_{p^{e_i}} C_{q_j^{k_j}}$ is a cyclic code of length $q_j^{k_j}$ over $\mathbb{Z}_{p^{e_i}}$. Therefore, an Abelian code over $\mathbb{Z}_m$ of length n is a direct product of cyclic codes over $\mathbb{Z}_{p^{e_i}}$. Now if A_{ij}

are ideals (cyclic codes) of $\mathbb{Z}_{p^{e_i}} C_{q_j^{k_j}}$, $1 \leq i \leq s$, $1 \leq j \leq t$, then $\prod_{i=1}^s \prod_{j=1}^t A_{ij}$ is an ideal (Abelian code) of length n over $\mathbb{Z}_m$, and its minimum distance is $\min_{i,j} d_{ij}$, where d_{ij} is the minimum distance of the cyclic code A_{ij} .

4. Codes Over Some Galois Rings

In this section we discuss some recent constructions of linear and cyclic codes over $\mathbb{Z}_{p^r}$, p a prime, and in particular, codes over $\mathbb{Z}_4$. For this, we first consider some basic properties of finite rings and polynomials over them. All the rings, which we consider here, are commutative rings with identity. Further, all rings, except polynomial rings, are finite.

4.1 Some Basic Notions Of Finite Commutative Rings

Theorem (2): [7, Theorem VI.2] *Let R be a commutative ring. Then R decomposes uniquely, up to order of summands, as a direct product of local rings.*

Therefore, most questions of finite commutative rings reduce to the questions of finite local rings. Further, this decomposition of finite commutative rings induces a natural decomposition of polynomial rings over them, into polynomial rings over local rings. In this section R denotes a finite commutative local ring with maximal ideal M and residue field $K = R/M$. We denote the natural projection $R[x] \rightarrow K[x]$ by μ . Here we collect some basic theory of $R[x]$.

Definition: Let $(f(x))$ be the ideal generated by $f(x) \in R[x]$. If $f(x)$ and $g(x)$ are in $R[x]$, then

- i) $f(x)$ is **regular** if $f(x)$ is not a zero divisor. Equivalently $\mu f(x) \neq 0$.
- ii) $f(x)$ and $g(x)$ are called **coprime** if $R[x] = (f(x)) + (g(x))$.

Definition: An ideal I in R is called a primary ideal if $I \neq R$ and whenever $xy \in I$, either $x \in I$ or $y^n \in I$ for some positive integer n . A polynomial $f(x) \in R[x]$ is called a primary polynomial if $(f(x))$ is a primary ideal.

Theorem (3): (Hensel's Lemma) [7, Theorem XIII.4]

Let $\overline{f(x)} \in \overline{R[x]}$ and $\overline{\mu f(x)} = \overline{g_1(x)} \overline{g_2(x)} \cdots \overline{g_n(x)}$ where $\overline{g_1(x)}, \overline{g_2(x)}, \dots, \overline{g_n(x)}$ are pairwise coprime polynomials. Then there exist $g_1(x), g_2(x), \dots, g_n(x)$ in $R[x]$ such that

- a) $\overline{g_1(x)}, \overline{g_2(x)}, \dots, \overline{g_n(x)}$ are pairwise coprime.
- b) $\mu g_i(x) = \overline{g_i(x)}$, $1 \leq i \leq n$.
- c) $f(x) = g_1(x) g_2(x) \cdots g_n(x)$.

The next theorem tells that every regular polynomial in $R[x]$ has a monic representative.

Theorem (4): [7, Theorem XIII.6] *Let $f(x)$ be a regular polynomial in $R[x]$.*

Then there is a monic polynomial $f^*(x)$ with $\mu f^*(x) = \mu f(x)$ and for an element

$a \in R, f(a) = 0$ if and only if $f^*(a) = 0$. Further there is a unit $v(x)$ in $R[x]$ with $f^*(x) = v(x)f(x)$.

Let J denote the set of all polynomials $f(x) \in R[x]$ such that $\mu f(x)$ has distinct zeros in the algebraic closure of K . The next theorem tells about the irreducible polynomials in $R[x]$.

Theorem (5): [7, Theorem XIII.7] Let $f(x)$ be a regular polynomial in $R[x]$,

then

- a) If $\mu f(x)$ is irreducible in $K[x]$ then $f(x)$ is irreducible.
- b) If $f(x)$ is irreducible then $\mu f(x) = \delta g(x)^n$ for some positive integer n where $\delta \in K$ and $g(x)$ is a monic irreducible polynomial in $K[x]$.
- c) If $f(x)$ is in J then $f(x)$ is irreducible if and only if $\mu f(x)$ is irreducible.

A regular polynomial in $R[x]$ does not necessarily factorizes uniquely into irreducible polynomials. However, there is a similar result true.

Theorem (6): [7, Theorem XIII.11] Let $f(x)$ be a regular polynomial in $R[x]$.

Then

- a) $f(x) = \delta g_1(x)g_2(x) \dots g_n(x)$ where δ is a unit and $g_1(x), g_2(x), \dots, g_n(x)$ are regular primary coprime polynomials.
- b) If $f(x) = \delta g_1(x)g_2(x) \dots g_n(x) = \beta h_1(x)h_2(x) \dots h_m(x)$ where δ, β are units and $\{g_i(x)\}$ and $\{h_j(x)\}$ are regular primary coprime polynomials then $n = m$ and after renumbering $(g_i(x)) = (h_i(x)), 1 \leq i \leq n$.

The study of irreducible polynomials over general rings is difficult because of the presence of zero divisors. More appropriate polynomials for studying the structure of rings are basic irreducible polynomials.

Definition: A polynomial $f(x) \in R[x]$ is called basic irreducible if $\mu f(x)$ is irreducible in $K[x]$.

Now if we take our ring to be $\mathbb{Z}_m$, with $m = \prod_{i=1}^s p_i^{e_i}$, then we know that $\mathbb{Z}_m = \prod_{i=1}^s \mathbb{Z}_{p_i^{e_i}}$. Therefore, the study of $\mathbb{Z}_m[x]$ basically reduces to the study of $\mathbb{Z}_{p^r}[x]$, p a prime. This is a special case of $R[x]$, and we now specialize our results to this case. First of all $\mathbb{Z}_{p^r}$ is a local ring with the maximal ideal $(p) = p\mathbb{Z}_{p^r}$ and the residue field $\mathbb{Z}_{p^r}/(p) \cong \mathbb{Z}_p$. For $f(x) \in \mathbb{Z}_{p^r}[x]$, $\mu f(x) = f(x) \pmod{p}$. Further, $f(x) \in \mathbb{Z}_{p^r}[x]$ is regular (i. e. not a zero divisor), if and only if $\mu f(x) \neq 0$, if and only if $f(x) \pmod{p} \neq 0$, if and only if there is some coefficient a_i of $f(x)$ such the $(a_i, p) = 1$. A polynomial $f(x) \in \mathbb{Z}_{p^r}[x]$ is basic irreducible if $f(x) \pmod{p}$ is irreducible in $\mathbb{Z}_p[x]$. Hensel's lemma for $\mathbb{Z}_{p^r}[x]$ can be described

similarly. Further, it can be shown that a basic irreducible polynomial in $\mathbb{Z}_{p^r}[x]$ is primary [6].

The following important result can be obtained either by using Hensel's Lemma or by direct construction [3].

Theorem (7): [3] If $f_1(x) \in \mathbb{Z}_p[x]$ is a monic irreducible divisor of $x^n - 1$ over $\mathbb{Z}_p$, then there is a unique monic irreducible polynomial $f(x) \in \mathbb{Z}_{p^r}[x]$ which divides $x^n - 1$ over $\mathbb{Z}_{p^r}$ such that $f(x) \pmod{p} = f_1(x)$.

The polynomial $f(x)$ in the above theorem is called the Hensel's lift of the polynomial $f_1(x)$.

For studying cyclic codes of length n over a finite field $\mathbb{F}_q$, we work in an extension $\mathbb{F}_q^m$, of $\mathbb{F}_q$ that contains an n th root of unity. Similarly, for studying cyclic codes of length n over $\mathbb{Z}_{p^r}$, we consider the extension $GR(p^r, m)$ of $\mathbb{Z}_{p^r}$, called a Galois ring over $\mathbb{Z}_{p^r}$ of degree m , which contains an n th root of unity.

Definition: Let p be a prime and $f(x) \in \mathbb{Z}_{p^r}[x]$ is a monic basic irreducible polynomial of degree m . Then the Galois ring of degree m , over $\mathbb{Z}_{p^r}$, is defined by $GR(p^r, m) = \mathbb{Z}_{p^r}[x]/(f(x))$.

We have, $GR(p^r, 1) \cong \mathbb{Z}_{p^r}$ and $GR(p, m) \cong \mathbb{F}_{p^m}$. The polynomial $f(x)$ divides $x^{p^m-1} - 1$. The Galois ring $\mathcal{R} = GR(p^r, m)$ is a local ring with the maximal ideal $(p) = p\mathcal{R}$ and the residue field $\mathcal{R}/p\mathcal{R} \cong \mathbb{F}_{p^m}$. If ξ is a root of $f(x)$, then $\mathcal{R} \cong \mathbb{Z}_{p^r}(\xi)$.

Before considering the codes over the ring $\mathbb{Z}_{p^r}$, we consider a special case, which has received much attention in recent years; the codes over $\mathbb{Z}_4$. The reason for considering this special case first than the general case is two fold. First, most research on codes over rings, in recent years, has been motivated by the results obtained for codes over $\mathbb{Z}_4$. Further, in most cases the results obtained for codes over $\mathbb{Z}_{p^r}$, are natural generalizations of codes over $\mathbb{Z}_4$.

4.2 Codes Over $\mathbb{Z}_4$

Some well known non-linear binary codes have more codewords than any binary linear code with the same length and minimum distance. These codes are Kerdock codes, Preparata codes, Goethal Codes and Delsarte-Goethal codes. Further, Kerdock and Preparata codes behave, as they are duals of each other. More precisely, the weight distribution of one is the weight distribution of the

MacWilliams transform of the other. This fact remained a mystery for several years before it was proved in a breakthrough paper [4] in 1994, by Hammons, Kumar, Calderbank, Sloane and Sole, that Kerdock and Preparata codes are in fact the binary images of two linear codes over $\mathbb{Z}_4$, which are duals of each other in $\mathbb{Z}_4$. This paper triggered an intense research in this area, and as a result several papers on codes over $\mathbb{Z}_4$ have been published. Further, the results in this paper were so motivating that the coding theorists started a fresh search of good codes over arbitrary finite rings, and much of the research in codes over rings, in recent years, can be attributed to this paper. The cyclic nature of Kerdock codes had already been shown by Nechaev [9].

As usual a linear code over $\mathbb{Z}_4$ or a $\mathbb{Z}_4$ -linear code of length n is a submodule of $\mathbb{Z}_4^n$. As we have remarked earlier that such a submodule need not be a free module, therefore, it is difficult to find the generator matrices of such a code. However good generator matrices have been obtained for $\mathbb{Z}_4$ -linear codes. A matrix whose rows generate a $\mathbb{Z}_4$ -linear code will be called a generator matrix for the code.

Let $\mathcal{C}$ be a $\mathbb{Z}_4$ linear code. Then there exist $k_1 + k_2$ codewords $\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_{k_1}, \mathbf{c}_{k_1+1}, \dots, \mathbf{c}_{k_1+k_2}$ such that every codeword in $\mathcal{C}$ can be expressed uniquely as $\mathbf{c} = \sum_{i=1}^{k_1} a_i \mathbf{c}_i + \sum_{i=k_1+1}^{k_1+k_2} a_i \mathbf{c}_i$ where $a_i \in \mathbb{Z}_4$ for $1 \leq i \leq k_1$ and $a_i \in \mathbb{F}_2$ for $k_1 + 1 \leq i \leq k_1 + k_2$. After a suitable permutation of coordinates, a generator matrix for $\mathcal{C}$ can be written as

$$G = \begin{bmatrix} I_{k_1} & A & B \\ O & 2I_{k_2} & 2C \end{bmatrix}$$

where A, C are matrices with entries from $\mathbb{F}_2$, and B is a matrix with entries from $\mathbb{Z}_4$. O is a $k_1 \times k_2$ zero matrix. Such a generator matrix is called in standard form.

The code contains $4^{k_1} 2^{k_2}$ codewords and said to be of the type $4^{k_1} 2^{k_2}$. As a module, the code is free if and only if $k_2 = 0$.

The generator matrix in standard form for $\mathcal{C}^\perp$ is given by

$$G^\perp = \begin{bmatrix} -B^T - C^T A^T & C^T & I_{n-k_1-k_2} \\ 2A^T & 2I_{k_2} & O \end{bmatrix}$$

where O is the $k_2 \times (n - k_1 - k_2)$ zero matrix. $\mathcal{C}^\perp$ has the type $4^{n-k_1-k_2} 2^{k_2}$.

The machinery of getting binary codes from $\mathbb{Z}_4$ linear codes is a map $\phi: \mathbb{Z}_4 \rightarrow \mathbb{F}_2^2$, called the gray map, and is defined by $\phi(0) = 00$, $\phi(1) = 01$, $\phi(2) = 11$ and $\phi(3) = 10$. This map is extended component wise to a map, which is also

denoted by ϕ , from $\mathbb{Z}_4^n$ to $\mathbb{F}_2^{2n}$. Thus the gray image $\phi(\mathcal{C})$ of a $\mathbb{Z}_4$ code $\mathcal{C}$ of length n is a binary code of length $2n$. It is easy to see that the Lee weight of any vector in $\mathbb{Z}_4^n$ is equal to the Hamming weight of its image in $\mathbb{F}_2^{2n}$. Thus the gray map ϕ is an isometry from $\mathbb{Z}_4^n$ with Lee distance to $\mathbb{F}_2^{2n}$ with Hamming distance.

4.2.1 Cyclic codes over $\mathbb{Z}_4$

A cyclic code of length n over $\mathbb{Z}_4$ is an ideal of the residue class ring $\mathcal{R}_n = \mathbb{Z}_4[x]/(x^n - 1)$. Therefore, to study cyclic codes over $\mathbb{Z}_4$, we need to study the structure of ideals of $\mathcal{R}_n$. We have already noted that a basic irreducible polynomial in $\mathbb{Z}_{p^r}[x]$ is primary; in particular, a basic irreducible polynomial in $\mathbb{Z}_4[x]$ is primary. Therefore, by factorization theorem for polynomial rings over finite commutative local rings with identity, any regular polynomial in $\mathbb{Z}_4[x]$ factorizes uniquely into basic irreducible and coprime polynomials. Further, we have the following important result.

Theorem (8): [5] *Let n be odd. Then $x^n - 1 = f_1(x) f_2(x) \dots f_k(x)$ where $f_i(x)$ are unique monic irreducible pairwise coprime polynomials in $\mathbb{Z}_4[x]$. Further, $x^n - 1 = \mu f_1(x) \mu f_2(x) \dots \mu f_k(x)$ is a factorization of $x^n - 1$ into irreducible polynomials in $\mathbb{F}_2[x]$.*

Proof: $\mu(x^n - 1) = (x^n - 1) \pmod{2} = x^n - 1 \in \mathbb{F}_2[x]$. Let $x^n - 1 = h_1(x) h_2(x) \dots h_k(x)$ be the factorization of $x^n - 1$ into irreducible polynomials in $\mathbb{F}_2[x]$. Since n is odd the factors $h_i(x)$ are coprime. By Hensel's Lemma there exist polynomials $g_1(x), g_2(x), \dots, g_k(x)$ in $\mathbb{Z}_4[x]$ such that $\mu g_i(x) = h_i(x)$, $1 \leq i \leq k$, and $x^n - 1 = g_1(x) g_2(x) \dots g_k(x)$ in $\mathbb{Z}_4[x]$.

Now $\mu(x^n - 1) \neq 0$ implies that $x^n - 1$ is a regular polynomial in $\mathbb{Z}_4[x]$.

Also since a basic irreducible polynomial in $\mathbb{Z}_4[x]$ is primary and each $g_i(x)$ is basic irreducible, therefore, each $g_i(x)$ is primary. Hence, by the factorization theorem for regular polynomials, the above factorization is unique up to multiplication by units.

As n is odd, $x^n - 1$ has distinct zeros in its splitting field over $\mathbb{F}_2$. Hence $x^n - 1 \in J$ and therefore each $g_i(x)$ is in J . Now $g_i(x) \in J$ and $\mu g_i(x) = h_i(x)$ is irreducible in $\mathbb{F}_2[x]$, therefore by Theorem (5) each $g_i(x)$ is irreducible in $\mathbb{Z}_4[x]$. Further, since each $g_i(x)$ is regular, therefore there exist a monic polynomial $f_i(x) \in \mathbb{Z}_4[x]$

such that $\mu f_i(x) = \mu g_i(x) = h_i(x)$ and $f_i(x) = v_i(x)g_i(x)$ where $v_i(x)$ is a unit in $\mathbb{Z}_4[x]$. Therefore, $x^n - 1 = a(x)f_1(x) f_2(x) \dots f_k(x)$, where $a(x)$ is a unit. Further, since $\mu f_i(x) = h_i(x)$ is irreducible, therefore by Theorem(5) each $f_i(x)$ is irreducible. Since $x^n - 1$ is monic and each $f_i(x)$ is monic and irreducible, we must have $a(x) = 1$, and hence $x^n - 1 = f_1(x) f_2(x) \dots f_k(x)$. Thus $x^n - 1$ factorizes uniquely into monic coprime irreducible polynomials. ♦

In order to factor $x^n - 1$ in $\mathbb{Z}_4[x]$, we first factor $\mu(x^n - 1) = (x^n - 1) \pmod{2} = x^n - 1$ in $\mathbb{F}_2[x]$ and then use Hensel's lift of each factor to determine the factorization of $x^n - 1$ in $\mathbb{Z}_4[x]$. One method of finding the Hensel's lift of a polynomial $f_1(x) \in \mathbb{F}_2[x]$ to $\mathbb{Z}_4[x]$ is the Graeffe's method [4, 5]. Let $f_1(x) = e(x) + o(x)$, where $e(x)$ contains only even powers of x and $o(x)$ contains only odd powers of x . Then the Hensel's lift $f(x) \in \mathbb{Z}_4[x]$ of $f_1(x)$ is obtained by $f(x) = \pm(e(x)^2 - o(x)^2)$, where the sign $\pm$ is chosen in such a way that the coefficient of the highest power of x is 1. This is illustrated by the following example.

Example 1: $x^7 - 1 = (x - 1)(x^3 + x + 1)(x^3 + x^2 + 1)$ is the factorization of $x^7 - 1$ into irreducible polynomials in $\mathbb{F}_2[x]$. If we take $f_1(x) = x^3 + x + 1$, then $e(x) = 1$ and $o(x) = x^3 + x$. Therefore, $f(x^2) = -(1 - (x^3 + x)^2) = x^6 + 2x^4 + x^2 - 1$ and thus $f(x) = x^3 + 2x^2 + x - 1$ is a factor of $x^7 - 1$ in $\mathbb{Z}_4[x]$.

To find the structures of ideals of $\mathcal{R}_n$, we first need to know the ideal structure of $\mathbb{Z}_4[x]/(f(x))$ where $f(x)$ is a basic irreducible polynomial.

Theorem (9): [5, 11] If $f(x)$ is a basic irreducible polynomial, then $R = \mathbb{Z}_4[x]/(f(x))$ has only three ideals (0) , $(2) = (2s(x) + (f(x)), s(x) \in \mathbb{Z}_4[x])$ and $(1) = R$.

Proof: Let I be a non zero ideal of $\mathbb{Z}_4[x]/(f(x))$. Let $g(x) + (f(x))$ be a non zero element of I . Since $\mu f(x)$ is irreducible, therefore $(\mu f(x), \mu g(x)) = 1$ or $\mu f(x)$. If $(\mu f(x), \mu g(x)) = 1$ then by the Euclidean algorithm there exist polynomials $a(x), b(x) \in \mathbb{Z}_4[x]$ such that $\mu a(x) \mu f(x) + \mu b(x) \mu g(x) = 1$. Therefore, $a(x)f(x) + b(x)g(x) = 1 + 2s(x)$ for some $s(x) \in \mathbb{Z}_4[x]$. Now $(1 + 2s(x))(a(x)f(x) + b(x)g(x)) = (1 + 2s(x))^2 = 1$, therefore, I contains an invertible element and hence $I = R$. If $(\mu f(x), \mu g(x)) = \mu f(x)$ then there exists $a(x) \in \mathbb{Z}_4[x]$ such that $\mu g(x) = \mu a(x) \mu f(x)$. Therefore, $g(x) = a(x) f(x) + 2s(x)$ for some $s(x) \in \mathbb{Z}_4[x]$. In this

case, $g(x) + (f(x)) \in (2)$. We can assume that $I \subseteq (2)$. As $I \neq 0$, there exists $h(x) \in \mathbb{Z}_4[x]$, such that $2h(x) + (f(x)) \in I$ with $2h(x) \notin (f(x))$. If $\mu h(x) \in (\mu f(x))$ then $h(x) = a(x)f(x) + 2s(x)$ for some $a(x), s(x) \in \mathbb{Z}_4[x]$. But then $2h(x) = 2a(x)f(x) \in (f(x))$, a contradiction. Therefore $\mu h(x) \notin (\mu f(x))$. Since $\mu f(x)$ is irreducible, therefore $(\mu f(x), \mu h(x)) = 1$. As before, there exist some $a(x), b(x), s(x) \in \mathbb{Z}_4[x]$ such that $a(x)f(x) + b(x)h(x) = 1 + 2s(x)$. Then $2a(x)f(x) + 2b(x)h(x) = 2$, which shows that $2 + (f(x)) \in I$ and so $(2) \subseteq I$. Thus $I = (2)$. ♦

Theorem (10): [5, 11] Let n is odd and let $x^n - 1 = f_1(x) f_2(x) \dots f_k(x)$ be a factorization of $x^n - 1$ into pairwise monic irreducible polynomials in $\mathbb{Z}_4[x]$. Let $\hat{f}_i(x) = \prod_{j \neq i} f_j(x)$. Then every ideal of $\mathcal{R}_n = \mathbb{Z}_4[x]/(x^n - 1)$ is a direct sum of $(\hat{f}_i(x))$'s and $(2\hat{f}_j(x))$'s.

Proof: By Theorem (8) such a unique factorization exists. Now $\mathcal{R}_n = \mathbb{Z}_4[x]/(x^n - 1) = \mathbb{Z}_4[x]/(\prod_{i=1}^k f_i(x))$. Since $f_i(x)$'s are pairwise coprime, therefore $(\prod_{i=1}^k f_i(x)) = \bigcap_{i=1}^k (f_i(x))$, and hence by the Chinese Remainder Theorem $\mathbb{Z}_4[x]/(x^n - 1) \cong \prod_{i=1}^k \mathbb{Z}_4[x]/(f_i(x))$.

Now if I is any ideal of $\mathcal{R}_n$, then $I = \bigoplus_{i=1}^k I_i$ where I_i is an ideal of $\mathbb{Z}_4[x]/(f_i(x))$. From Theorem (11), $I_i = 0$ or $\mathbb{Z}_4[x]/(f_i(x))$ or $(2 + f_i(x))$. Now $\mathbb{Z}_4[x]/(f_i(x))$ corresponds to the ideal $(\hat{f}_i(x))$, and $(2 + f_i(x))$ corresponds to the ideal $(2\hat{f}_i(x))$ in the ring $\mathcal{R}_n$. In all cases I is a sum of some $(\hat{f}_i(x))$'s and some $(2\hat{f}_j(x))$'s. ♦

The next theorem tells us about the generator polynomials of cyclic codes over $\mathbb{Z}_4$.

Theorem (11): [5, 11] Let $\mathcal{C}$ be a cyclic code over $\mathbb{Z}_4$ of odd length n . Then there exist unique monic polynomials $f(x), g(x)$ and $h(x)$ such that $x^n - 1 = f(x)g(x)h(x)$ and $\mathcal{C} = (f(x)g(x)) \oplus (2f(x)h(x))$.

Proof: Let $x^n - 1 = f_1(x) f_2(x) \dots f_k(x)$ where $f_i(x)$ are pairwise coprime monic basic irreducible polynomials over $\mathbb{Z}_4$. Then $\mathcal{C}$ is a direct sum of $(\hat{f}_i(x))$'s and some $(2\hat{f}_j(x))$'s i.e. $\mathcal{C} = \sum_{i=1}^a (\hat{f}_i(x)) \oplus \sum_{i=a+1}^b (2\hat{f}_i(x))$ for some $1 \leq a, b \leq k$. Let $f(x) = \prod_{i=b+1}^k f_i(x)$, $g(x) = \prod_{i=a+1}^b f_i(x)$ and $h(x) = \prod_{i=1}^a f_i(x)$. Then $x^n - 1 = f(x)g(x)h(x)$.

Now $\hat{f}_i(x)$ is a multiple of $f(x)g(x)$ for $1 \leq i \leq a$, therefore $\sum_{i=1}^a \hat{f}_i(x) \subseteq (f(x)g(x))$. But $(f(x)g(x))$ has size $4^{n-\deg fg} = 4^{\deg h}$ which also happen to be the size of $\sum_{i=1}^a \hat{f}_i(x)$. Therefore $\sum_{i=1}^a \hat{f}_i(x) = (f(x)g(x))$. Similarly it can be shown that $\sum_{i=a+1}^b (2\hat{f}_i(x)) = (2f(x)h(x))$. The uniqueness of $f(x)$, $g(x)$ and $h(x)$ follows from the unique factorization of $x^n - 1$ into monic basic irreducible pairwise coprime polynomials. Clearly the size of $\mathcal{C}$ is $4^{\deg h} 2^{\deg g}$ ♦

Now we generalize our results for codes over $\mathbb{Z}_{p^r}$, p a prime.

4.3 Codes Over $\mathbb{Z}_{p^r}$

Codes over $\mathbb{Z}_{p^r}$, p a prime, are natural generalization of codes over $\mathbb{Z}_4$, and they are also the counterparts of codes over fields. As usual, a linear code over $\mathbb{Z}_{p^r}$ of length n is a submodule of $\mathbb{Z}_{p^r}^n$. As a natural generalization of a generator matrix of a linear code over $\mathbb{Z}_4$, a generator matrix of a linear code $\mathcal{C}$ over $\mathbb{Z}_{p^r}$, after a suitable permutation of the co-ordinates, can be written in the form

$$G = \begin{bmatrix} I & A_{01} & A_{02} & A_{03} & \dots & A_{0,r-1} & A_{0,r} \\ 0 & pI & pA_{12} & pA_{13} & \dots & pA_{1,r-1} & pA_{1,r} \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & p^{r-1}I & p^{r-1}A_{r-1,r} \end{bmatrix}$$

where the columns are grouped into blocks of sizes $k_0, k_1, \dots, k_r$ and k_i 's are non-negative integers adding to n . Therefore, $\mathcal{C}$ consists of all words $[v_0, v_1, \dots, v_{r-1}]G$ where each v_i is a vector of length k_i with components from $\mathbb{Z}_{p^{r-i}}$ and $\mathcal{C}$ contains p^k codewords where $k = \sum_{i=0}^{r-1} (r-i)k_i$ [3].

Here we say that $\mathcal{C}$ has the type $1^{k_0} p^{k_1} (p^2)^{k_2} \dots (p^{r-1})^{k_{r-1}}$. This is to be noted that this definition is different from what is used for $\mathbb{Z}_4$ codes.

In the same way we can find a generator matrix for $\mathcal{C}^\perp$. The

corresponding generator matrix of $\mathcal{C}^\perp$ is of the form

$$G^\perp = \begin{bmatrix} B_{0r} & B_{0,r-1} & \cdots & B_{03} & B_{02} & B_{01} & I \\ pB_{1r} & pB_{1,r-1} & \cdots & pB_{13} & pB_{12} & pI & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots \\ p^{r-1}B_{r-1,r} & p^{r-1}I & \cdots & 0 & 0 & 0 & 0 \end{bmatrix}$$

where the column blocks have the same size as in G . The dual code $\mathcal{C}^\perp$, therefore, contains $p^{k_\perp}$ codewords, where $k_\perp = \sum_{i=1}^r ik_i$, and has the type $1^{k_r} p^{k_{r-1}} (p^2)^{k_{r-2}} \cdots (p^{r-1})^{k_1}$ [3].

4.3.1 Cyclic Codes Over $\mathbb{Z}_{p^r}$

Let p be a prime and $q = p^r$. As usual a cyclic code of length n over $\mathbb{Z}_q$ is an ideal in the residue class ring $\mathcal{R}_n = \mathbb{Z}_q[x]/(x^n - 1)$. We have a similar result for the factorization of $x^n - 1$ over $\mathbb{Z}_q$ as we had for $\mathbb{Z}_4$.

Theorem (12): [6] *If p does not divide n , then $x^n - 1 = f_1(x) f_2(x) \cdots f_k(x)$, where $f_i(x)$ are regular, primary, pairwise-coprime polynomials in $\mathbb{Z}_q[x]$. The factorization is unique up to order of multiplication, and multiplication by units.*

Most of the following results are straightforward generalizations to $\mathbb{Z}_q$, of the similar results over $\mathbb{Z}_4$.

Theorem (13): [6] *If $f(x) \in \mathbb{Z}_q[x]$ is a basic irreducible polynomial then the ideals of $\mathbb{Z}_q[x]/(f(x))$ are precisely $(0), (1), (p + f(x)), \dots, (p^{r-1} + f(x))$.*

The next theorem gives the structure of ideals of $\mathcal{R}_n = \mathbb{Z}_q[x]/(x^n - 1)$.

Theorem (14): [6] *Let p does not divide n and $x^n - 1 = f_1(x) f_2(x) \cdots f_k(x)$ be the factorization of $x^n - 1$ into basic irreducible pairwise coprime polynomials in $\mathbb{Z}_q[x]$. Then every ideal in $\mathcal{R}_n$ is a direct sum of ideals $(p^j \hat{f}_i(x))$ where $0 \leq j \leq r-1, 1 \leq i \leq k$ and for $1 \leq i \leq k$ $\hat{f}_i(x) = (x^n - 1) / f_i(x) = \prod_{j \neq i} f_j(x)$.*

The following important theorem gives us the generator polynomials of cyclic codes over $\mathbb{Z}_q$.

Theorem (15): [3, 6] Let p does not divide n and, $\mathcal{C}$ is a cyclic code over $\mathbb{Z}_q$.

Then there exist polynomials $f_0(x) f_1(x) \dots f_{r-1}(x)$ such that $f_{r-1}(x) \mid f_{r-2}(x) \mid \dots \mid f_0(x) \mid (x^n - 1)$ and $\mathcal{C} = (f_0(x), pf_1(x), \dots, p^{r-1}f_{r-1}(x))$.

In fact $\mathcal{R}_n = \mathbb{Z}_q[x] / (x^n - 1)$ is a principal ideal ring [3, 6], and $\mathcal{C}$ defined above, is generated by the polynomial $g(x) = f_0(x) + pf_1(x) + \dots + p^{r-1}f_{r-1}(x)$ [3].

Cyclic codes over $\mathbb{Z}_p$ can be used to obtain cyclic codes over $\mathbb{Z}_q$. We know from Theorem (7), that if $f_1(x) \in \mathbb{Z}_p[x]$ is a monic irreducible divisor of $x^n - 1$ over $\mathbb{Z}_p$, then there is a unique monic irreducible polynomial $f(x) \in \mathbb{Z}_q[x]$ which divides $x^n - 1$ over $\mathbb{Z}_q$ and $f(x) \pmod{p} = f_1(x)$. To construct a cyclic code over $\mathbb{Z}_q$, we find a monic polynomial $f_1(x) \in \mathbb{Z}_p[x]$ which is a divisor of $x^n - 1$ over $\mathbb{Z}_p$. This is a generator polynomial of a cyclic code of length n over $\mathbb{Z}_p$. Now defining the Hensel's lift of this polynomial to be the generator polynomial of a cyclic code, we get a cyclic code of length n over $\mathbb{Z}_q$. For example, a BCH code over $\mathbb{Z}_q$ can be obtained by defining its generator polynomial to be the Hensel's lift of a generator polynomial of a BCH code over $\mathbb{Z}_p$. This construction of BCH codes over $\mathbb{Z}_q$ is essentially the construction given by Shankar [12]. Shankar [12] has then given a construction of BCH codes over $\mathbb{Z}_m$, where m is an arbitrary positive integer. In the study of codes over rings, in recent years, Shankar's construction of cyclic codes has played a very important role.

5. CONCLUSION

In this paper we have considered some classical and some recent constructions of linear and cyclic codes over the integer residue ring $\mathbb{Z}_m$. The constructions of codes over $\mathbb{Z}_m$, basically reduces to the corresponding constructions of codes over $\mathbb{Z}_{p^r}$ where p is a prime. In particular, we have briefly discussed some $\mathbb{Z}_4$ linear codes.

Acknowledgement: The authors greatly acknowledge the financial support given by Centre for Development of Advanced Computing (C-DAC), Noida, India, to present the paper in the conference.

REFERENCES

1. I. F. Blake, *Codes over certain rings*, Inform. Control, **20**, (1972), 396 – 404.
2. I. F. Blake, *Codes over integer residue rings*, Inform. Control, **29**, (1975), 295 – 300.
3. A. R. Calderbank and N. J. A. Sloane *Modular and p -adic cyclic codes*, Designs Codes and Cryptography, **6**, (1995), 21-35.
4. A. R. Hammons Jr., P. V. Kumar, A. R. Calderbank, N. J. A. Sloane and P. Sole, *The $\mathbb{Z}_4$ - linearity of Kerdock, Preparata, Goethals and related codes*, IEEE Trans. Inform. Theory, **40**, (1994), 301 – 319.
5. W. Cary Huffman and Vera Pless, *Fundamentals of Error Correcting Codes*, Cambridge University Press, 2003.
6. P. Kanwar and Sergio R. Lopez-Permouth *Cyclic codes over the integers modulo p^m* , Finite Fields and Their Applications, **3**, (1997), 334-352.
7. Bernard R. Macdonald, *Finite Rings with Identity*, Marcel Dekker , New York, 1974.
8. F. J. Macwilliams and N. J. A. Sloane *The Theory of Error Correcting Codes*, North Holland Publishers, 1977.
9. A. A. Nechaev, *The kerdock code in cyclic form*, Diskret. Mat. **1** (1989), 123 – 139. English translation in Discrete Math. Appl., **1** (1991), 365 – 384.
10. G. H. Norton and A. Salagean-Mandache, *On the structures of linear and cyclic codes over finite chain rings*, Applicable Algebra in Engineering Communication and Computing, **10**, (2000), 489 – 506
11. Vera S. Pless and Zhongqiang Qian, *Cyclic codes and quadratic residue codes over $\mathbb{Z}_4$* , IEEE Trans. Inform. Theory, **42**, (1996), 1594 - 1600.
12. Priti Shankar, *On BCH codes over arbitrary integer rings*, IEEE Trans. Inform. Theory, **25**, (1979), 480 - 483.
13. Eugene Spiegel, *Codes over $\mathbb{Z}_m$* , Inform. Control, **35**, (1977), 48 - 51.
14. Eugene Spiegel, *Codes over $\mathbb{Z}_m$, Revisited*, Inform. Control, **37**, (1978), 100 – 104.
15. Z. X. Wan, *Cyclic codes over Galois rings*, Algebra Colloquium, **6**, (1999), 291-304
16. S. K. Wasan, *On codes over $\mathbb{Z}_m$* , IEEE Trans. Inform. Theory, **28**, (1982), 117 - 120.
17. J. A. Wood, *Duality for modules over finite rings and applications to coding theory*, American Journal of Mathematics, **121** (1999), 555 - 575

Maheshanand

Centre for Development of Advanced Computing (C-DAC),
C-56/1, Anusandhan Bhavan, Sector-62, Noida – 201307, India
E-Mail : maheshanand@cdacnoida.in

Siri Krishan Wasan

Department Of Mathematics, Jamia Millia Islamia, Delhi-110025, India
E-Mail : skwasan@yahoo.com

THE EXISTENCE OF A SOLUTION OF THE NON-HOMOGENEOUS SYSTEMS OVER AN ELEMENTARY DIVISOR DOMAIN ON SOLVABLE MODULE

IKRAR PRAMUDYA, SRI WAHYUNI,
ARI SUPARWANTO AND WIDODO

Abstract. Suppose R is ring with identity and $\mathcal{M}$ is a R -module. And further we form the non-homogeneous equation systems: $Ax = x_0$ where $A \in R^{n \times n}$, $\det(A) \neq 0$, and $x_0 \in \mathcal{M} \setminus \{0\}$. The solution of equation systems on $\mathcal{M}$ doesn't always exist. In this paper, we study the properties of R and $\mathcal{M}$ such that the solution of equation systems exists. We also investigate the characterization of its solution space and find the relevant examples. The systems has a solution, if R is an elementary divisor domain and $\mathcal{M}$ is solvable module, that is, equation $p x = x_0$ where $p \in R \setminus \{0\}$ and $x_0 \in \mathcal{M} \setminus \{0\}$ always has a solution in $\mathcal{M}$.

This study works in the algebraic framework. Therefore, a solution of the equation systems will be found by implementation theorems in algebra, especially theorems in ring theory, module theory, or linear algebra. One of the most important result is the existence of the solution of equation systems: $Ax = x_0$ where $A \in \mathbf{R}[\mathbf{d}, \sigma, \sigma^{-1}]^{n \times n}$, $\det(A) \neq 0$, and $x_0 \in \mathcal{L}_{e, \text{pol}}^n$

in $\mathcal{L}_{e, \text{pol}}^n$ ¹. In this case, each $p \in \mathbf{R}[\mathbf{d}, \sigma, \sigma^{-1}]$ can be considered as delay-differential

operator on $\mathcal{L}_{e, \text{pol}}$. In general, characteristic of the solution space of equation systems can

be identified completely because the solution space of equation systems: $Ax = x_0$ equals to

$P(\mathcal{B}_{\mathcal{L}}^{\mathcal{L}_{e, \text{pol}}}(D, v_0))$ for $P \in \text{GL}(n, \mathcal{H})$ where $\mathcal{H} := \{ \frac{p}{\phi} \mid p \in \mathbf{R}[\mathbf{d}, \sigma, \sigma^{-1}], \phi \in \mathbf{R}[\mathbf{d}] \setminus \{0\}, \frac{p}{\phi} \in H(\mathbf{C}) \}$ and $\mathcal{B}_{\mathcal{L}}^{\mathcal{L}_{e, \text{pol}}}(D, v_0)$ be the solution space of equation systems: $Dx = v_0$ where D is

diagonal matrix and $v_0 \in \mathcal{L}_{e, \text{pol}}^n$.

Key words: an elementary divisor domain, delay-differential operator, solvable module

1. INTRODUCTION

Suppose $\mathcal{M}$ is module over ring R . We observed the equation systems of the form: $Dx = x_0$ with $D = (d_1, \dots, d_n)$ is diagonal matrix over R , $\det(D) \neq 0$, and $x_0 \in \mathcal{M}^n$. If the solution space of the equation systems is denoted by $\mathcal{B}_{\mathcal{M}}(A, x_0)$ then

$$\begin{aligned} \mathcal{B}_{\mathcal{M}}(A, x_0) &= \{x \in \mathcal{M}^n \mid Dv = x_0\} \\ &= \{x \in \mathcal{M}^n \mid d_i x_i = x_{0i}\} \end{aligned}$$

¹ $\mathcal{L}_{e, \text{pol}} := \text{Span}_{\mathbf{C}} \{t^k e^{\lambda t} \mid k \in \mathbf{N}_0, \lambda \in \mathbf{C}\}$

$$= \{x \in \mathcal{M}^n \mid x_i \in \mathcal{B}_{\mathcal{M}}(d_i, x_{0i})\}$$

$$\text{which } \mathcal{B}_{\mathcal{M}}(d_i, x_{0i}) = \{x_i \in \mathcal{M} \mid d_i x_i = x_{0i}\}.$$

The set $\mathcal{B}_{\mathcal{M}}(A, x_0)$ is not always exist. The set $\mathcal{B}_{\mathcal{M}}(A, x_0)$ is not empty, if for each i ($1 \leq i \leq n$) hold $\mathcal{B}_{\mathcal{M}}(d_i, x_{0i}) \neq \emptyset$. In this paper we discuss the properties A , R , and $\mathcal{M}$ in order the solution space of equation systems: $Ax = x_0$ with $A \in R^{n \times n}$ and $x_0 \in \mathcal{M}^n$ is not empty.

One of the definitions in ring theory explains that in an elementary division domain R for each $A \in R^{m \times n}$ there are $P \in GL(m, R)$ and $Q \in GL(n, R)$ such that $PAQ = D$ with D is diagonal matrix. The conditions can be used as an analyze tools for characterization of the solution space of the systems.

The next problem is how to determine the conditions of $\mathcal{M}$ such that for $p \in R \setminus \{0\}$ and $v_0 \in \mathcal{M}$ hold $\mathcal{B}_{\mathcal{M}}(p, v_0) \neq \emptyset$. To simplify, in this paper we assumed $\mathcal{M}$ having that property.

This research is important because its results can be applied to solve many problems. For example, to determine the solution or the solution space of the delay-differential equation systems, especially commensurate delay.

Definition 1.1. [2, p. 107] *Let R be an arbitrary ring with identity (not necessarily commutative). A left R -module is an abelian group M together with a scalar multiplication map*

$$\cdot : R \times M \mapsto M$$

that satisfy the following axioms (as is customary we will write am in place of $\cdot(a, m)$ for the scalar multiplication of $m \in M$ by $a \in R$). In these axioms, a, b are arbitrary elements of R and m, n are arbitrary elements of M .

- (1) $a(m+n) = am + an$
- (2) $(a+b)m = am + bm$
- (3) $(ab)m = a(bm)$
- (4) $1m = m$

In this paper all of module is considered as left module.

Definition 1.2. [1, p.36] *Let R be a ring. R is called elementary divisor domain if every matrix $A \in R^{n \times m}$ is equivalent to a diagonal matrix where each diagonal element divides the next one. Precisely, there exist $P \in GL(n, R)$ and $Q = GL(m, R)$ such that $PAQ = \text{diag}_{n \times m}(d_1, \dots, d_r, 0, \dots, 0)$ with $r = \text{rk}(A)$ and $d_i \in R \setminus \{0\}$ satisfying $d_i \mid d_{i+1}$ for $i = 1, \dots, r-1$. The diagonal elements are the invariant factors of A and hence unique up to units of R .*

2. MAIN RESULTS

To discuss solution and the solution space of the systems: $Ax = x_0$ directly is not easy. We need some knowledge of concept and elementer theory algebra and by that we can solve its problem more easily. In this study R -module $\mathcal{M}$ is considered as solvable module over an elementary divisor domain except there is especially notes. The following definitions and theorems are needed to solve the research problem.

Definition 2.3. *Let $\mathcal{M}$ be a R -module, let $r \in R \setminus \{0\}$, and let $m_0 \in \mathcal{M}$.*

- (1) *Define $\mathcal{B}_{\mathcal{M}}(r, m_0) := \{x \in \mathcal{M} \mid rx = m_0\}$.*
- (2) *$\mathcal{M}$ is called solvable, if for each $r \in R \setminus \{0\}$ and $m_0 \in \mathcal{M}$ hold $\mathcal{B}_{\mathcal{M}}(r, m_0) \neq \emptyset$.*

Definition 2.4. Let $\mathcal{M}$ be a R -module, $A \in R^{n \times n} \setminus \{0\}$, and let $x_0 \in \mathcal{M}^n$. Define $\mathcal{B}_{\mathcal{M}}(A, x_0) := \{x \in \mathcal{M}^n \mid Ax = x_0\}$.

Proposition 2.5. If D is a diagonal matrix in $R^{n \times n}$, $\det(D) \neq 0$, and $x_0 \in \mathcal{M}^n$, then $\mathcal{B}_{\mathcal{M}}(D, x_0) = \{x \in \mathcal{M}^n \mid x_i \in \mathcal{B}_{\mathcal{M}}(d_i, x_{0i}); i = 1, \dots, n\} \neq \emptyset$.

Proof. Since $\det(D) \neq 0$, then $d_i \neq 0$ ($1 \leq i \leq n$) such that $\mathcal{B}_{\mathcal{M}}(d_i, x_{0i}) \neq \emptyset$. Consequently, $\mathcal{B}_{\mathcal{M}}(D, x_0) = \{x \in \mathcal{M}^n \mid x_i \in \mathcal{B}_{\mathcal{M}}(d_i, x_{0i}); i = 1, \dots, n\} \neq \emptyset$.

Theorem 2.6. Let $\mathcal{M}$ be a solvable R -module and let $x_0 \in \mathcal{M}^n$. If $D = \text{diag}(d_1, \dots, d_n) \in R^{n \times n}$ and $\det(D) \neq 0$, then exist $g_0 \in \mathcal{M}^n$ such that $(\text{Ker}(d_1))^n + \{g_0\} \subseteq \mathcal{B}_{\mathcal{M}}(D, x_0)$.

Proof. Since $\det(D) \neq 0$, then $d_j \neq 0$ for each j ($1 \leq j \leq n$). And then, exist $g_{0j} \in \mathcal{M}$ such that $d_j g_{0j} = x_{0j}$, since $\mathcal{M}$ is a solvable module. Choice $g_0 := [g_{01} \cdots g_{0n}]^t$ and then form the set of $(\text{Ker}(d_1))^n + \{g_0\}$. Clear, $Dg_0 = x_0$. If $y \in (\text{Ker}(d_1))^n + \{g_0\}$, that is, $y = y_0 + g_0$ where $y_0 \in (\text{Ker}(d_1))^n$.

$$\begin{aligned} Dy &= D(y_0 + g_0) \\ &= Dy_0 + Dg_0 \\ &= [d_1 y_{01} \cdots d_n y_{0n}]^t + x_0. \end{aligned}$$

Since $y_0 \in (\text{Ker}(d_1))^n$, then $d_1 y_{0i} = 0$ ($1 \leq i \leq n$). Furthermore, $d_{i+1} = p_i d_1$ for some $p_i \in R$, because $d_i \mid d_{i+1}$ ($1 \leq i \leq n-1$). Hence, $d_1 y_{01} = 0$, $d_2 y_{02} = p_1 d_1 y_{02} = 0$, ..., $d_n y_{0n} = p_{n-1} d_1 y_{0n} = 0$. $\therefore [d_1 y_{01} \cdots d_n y_{0n}]^t = 0$ or $Dy = x_0$.

Theorem 2.7. The solution of equation: $rx = v_0$ where $r \in R \setminus \{0\}$ and $v_0 \in \mathcal{M}$ that is, $\mathcal{B}_{\mathcal{M}}(r, v_0) = \mathcal{B}_{\mathcal{M}}(r, 0) + \mathcal{B}_{\mathcal{M}}(r, v_0)$.

Proof. Let $x \in \mathcal{B}_{\mathcal{M}}(r, v_0)$. If we write $x = 0 + x$, we get $x \in \mathcal{B}_{\mathcal{M}}(r, 0) + \mathcal{B}_{\mathcal{M}}(r, v_0)$ because $0 \in \mathcal{B}_{\mathcal{M}}(r, 0)$ and $x \in \mathcal{B}_{\mathcal{M}}(r, v_0)$. Conversely, if $x \in \mathcal{B}_{\mathcal{M}}(r, 0) + \mathcal{B}_{\mathcal{M}}(r, v_0)$, we have $x = y_0 + y_1$ which $y_0 \in \mathcal{B}_{\mathcal{M}}(r, 0)$ and $y_1 \in \mathcal{B}_{\mathcal{M}}(r, v_0)$. Since $rx = r(y_0 + y_1) = ry_0 + ry_1 = 0 + v_0 = v_0$, then $x \in \mathcal{B}_{\mathcal{M}}(r, v_0)$.

Theorem 2.8. If $\mathcal{B}_{\mathcal{M}}(A, x_0)$ is the set of all partial solution of equation systems: $Ax = x_0$ where $A \in R^{n \times n}$ and $x_0 \in \mathcal{M}^n$ then $\mathcal{B}_{\mathcal{M}}(A, x_0) = \mathcal{B}_{\mathcal{M}}(A, 0) + \mathcal{B}_{\mathcal{M}}(A, x_0)$.

Proof. The inclusion “ $\subseteq$ ”, let $v \in \mathcal{B}_{\mathcal{M}}(A, x_0)$. Since $v = 0 + v$ with $0 \in \mathcal{B}_{\mathcal{M}}(A, 0)$ and $v \in \mathcal{B}_{\mathcal{M}}(A, x_0)$ implies $v \in \mathcal{B}_{\mathcal{M}}(A, 0) + \mathcal{B}_{\mathcal{M}}(A, x_0)$. As for “ $\supseteq$ ”, let $v \in \mathcal{B}_{\mathcal{M}}(A, 0) + \mathcal{B}_{\mathcal{M}}(A, x_0)$, so that $v = y_0 + y_1$ where $y_0 \in \mathcal{B}_{\mathcal{M}}(A, 0)$ and $y_1 \in \mathcal{B}_{\mathcal{M}}(A, x_0)$. Since $Av = A(y_0 + y_1) = Ay_0 + Ay_1 = 0 + x_0 = x_0$. Thus $v \in \mathcal{B}_{\mathcal{M}}(A, x_0)$.

Theorem 2.9.

Let $A \in R^{n \times n}$ and $x_0 \in \mathcal{M}^n$. For each $P \in R^{n \times n}$ holds $\mathcal{B}_{\mathcal{M}}(A, x_0) \subseteq \mathcal{B}_{\mathcal{M}}(PA, Px_0)$.

Proof. Let $x \in \mathcal{B}_{\mathcal{M}}(A, x_0)$, we have $Ax = x_0$. Hence, we get $PAx = Px_0$. In other words, $x \in \mathcal{B}_{\mathcal{M}}(PA, Px_0)$. Since for each $x \in \mathcal{B}_{\mathcal{M}}(A, x_0)$ implies $x \in \mathcal{B}_{\mathcal{M}}(PA, Px_0)$, then $\mathcal{B}_{\mathcal{M}}(A, x_0) \subseteq \mathcal{B}_{\mathcal{M}}(PA, Px_0)$.

Theorem 2.10. *Let $A \in R^{n \times n}$ and $x_0 \in \mathcal{M}^n$. For each $P \in GL(n, R)^2$ holds $\mathcal{B}_{\mathcal{M}}(A, x_0) = \mathcal{B}_{\mathcal{M}}(PA, Px_0)$.*

Proof. From **Theorem 2.9.**, we can conclude that $\mathcal{B}_{\mathcal{M}}(A, x_0) \subseteq \mathcal{B}_{\mathcal{M}}(PA, Px_0)$. As for “ $\supseteq$ ”, let $x \in \mathcal{B}_{\mathcal{M}}(PA, Px_0)$, we have $PAx = Px_0$. There exist $P^{-1} \in GL(n, R)$ such that $P^{-1}P = I_n$, since $P \in GL(n, R)$. Consequently, $Ax = P^{-1}PAx = P^{-1}Px_0 = x_0$. Hence, we have $x \in \mathcal{B}_{\mathcal{M}}(A, x_0)$. For each $x \in \mathcal{B}_{\mathcal{M}}(PA, Px_0)$ implies $x \in \mathcal{B}_{\mathcal{M}}(A, x_0)$, that is, $\mathcal{B}_{\mathcal{M}}(PA, Px_0) \subseteq \mathcal{B}_{\mathcal{M}}(A, x_0)$.

Theorem 2.11. *Let $A \in R^{n \times n}$, $\det(A) \neq 0$, and $x_0 \in \mathcal{M}^n$. If $P \in GL(n, R)$ then $\mathcal{B}_{\mathcal{M}}(AP, x_0) = P^{-1}(\mathcal{B}_{\mathcal{M}}(A, x_0))$.*

Proof. Let $x \in \mathcal{B}_{\mathcal{M}}(AP, x_0)$, we have $APx = x_0$. In other words, $Px \in \mathcal{B}_{\mathcal{M}}(A, x_0)$. $x = P^{-1}u$ with $u \in \mathcal{B}_{\mathcal{M}}(A, x_0)$. For each $x \in \mathcal{B}_{\mathcal{M}}(AP, x_0)$ implies $x \in P^{-1}(\mathcal{B}_{\mathcal{M}}(A, x_0))$, that is, $\mathcal{B}_{\mathcal{M}}(AP, x_0) \subseteq P^{-1}(\mathcal{B}_{\mathcal{M}}(A, x_0))$. Conversely, if $x \in P^{-1}(\mathcal{B}_{\mathcal{M}}(A, x_0))$, then $x = P^{-1}v$ for some $v \in \mathcal{B}_{\mathcal{M}}(A, x_0)$. Because $APx = AP P^{-1}v = Av = x_0$, $x \in \mathcal{B}_{\mathcal{M}}(AP, x_0)$. For each $x \in P^{-1}(\mathcal{B}_{\mathcal{M}}(A, x_0))$ implies $x \in \mathcal{B}_{\mathcal{M}}(AP, x_0)$, that is, $P^{-1}(\mathcal{B}_{\mathcal{M}}(A, x_0)) \subseteq \mathcal{B}_{\mathcal{M}}(AP, x_0)$.

2.1 The Solution Space of The Homogeneous Systems

Observed the systems: $Ax = 0$ where $A \in R^{n \times n}$ and $\det(A) \neq 0$. The form of the systems is called the homogeneous systems. The relevant properties of the solution space on $\mathcal{M}^n$ are as the following theorems.

Theorem 2.12.

$\mathcal{B}_{\mathcal{M}}(A, 0) = Q(\mathcal{B}_{\mathcal{M}}(D, 0))$ for some diagonal matrix D and $Q \in GL(n, R)$.

Proof. Observed the systems: $Ax = 0$ where $A \in R^{n \times n}$ and $\det(A) \neq 0$. Because R is an elementary divisor domain, There are exist $P, Q \in GL(n, R)$ such that $PAQ = D$ and D is diagonal matrix. Because $\det(PAQ) = \det(P)\det(A)\det(Q)$ and $\det(A) \neq 0$, $\det(D) \neq 0$. Consequently, $D = \text{diag}(d_1, \dots, d_n)$; $d_i \neq 0$ ($1 \leq i \leq n$). From $PAQ = D$, we have $PA = DQ^{-1}$. By applying **Theorem 2.10.** and **2.11.**, we get

$$\begin{aligned} \mathcal{B}_{\mathcal{M}}(A, 0) &= \mathcal{B}_{\mathcal{M}}(PA, P0) \\ &= \mathcal{B}_{\mathcal{M}}(DQ^{-1}, 0) \\ &= Q(\mathcal{B}_{\mathcal{M}}(D, 0)). \end{aligned}$$

Theorem 2.13.

There exist $d \in R$ and $Q \in GL(n, R)$ such that $Q(\mathcal{M}_{\tau(d)})^n \subseteq \mathcal{B}_{\mathcal{M}}(A, 0)$.

Proof. Because R is an elementary divisor domain, there exist some $P, Q \in GL(n, R)$ such that $PAQ = D$ where D is diagonal matrix; $d_i | d_{i+1}$ ($1 \leq i \leq n-1$). Therefore, $d_{i+1} = a_i d_i$ for some $a_i \in R$. If $d = d_1$, for each i ($1 \leq i \leq n-1$) holds $\mathcal{B}_{\mathcal{M}}(d_1, 0) \subseteq \mathcal{B}_{\mathcal{M}}(d_i, 0)$. Consequently, $(\mathcal{B}_{\mathcal{M}}(d, 0))^n = (\mathcal{M}_{\tau(d)})^n \subseteq \mathcal{B}_{\mathcal{M}}(D, 0)$. Hence, we have $Q((\mathcal{M}_{\tau(d)})^n) \subseteq Q(\mathcal{B}_{\mathcal{M}}(D, 0)) = \mathcal{B}_{\mathcal{M}}(A, 0)$.

² The set of all invertible matrices in $R^{n \times n}$

2.2 The Solution Space of The Non-homogeneous Systems

Observed the equation systems: $Ax = x_0$ where $A \in R^{n \times n}$, $\det(A) \neq 0$, and $x_0 \in \mathcal{M}^n \setminus \{0\}$. That systems is called the non-homogen systems. The relevant properties of the solution space on $\mathcal{M}^n$ are as the following theorems.

Theorem 2.14. $\mathcal{B}_{\mathcal{M}}(A, x_0) = Q(\mathcal{B}_{\mathcal{M}}(D, x_0))$ for some $P, Q \in GL(n, R)$ where D is diagonal matrix. Moreover, there exist $d \in R$ and $g_0 \in \mathcal{M}^n$ such that $Q(\mathcal{M}_{\tau(d)}^n + \{g_0\}) \subseteq \mathcal{B}_{\mathcal{M}}(A, x_0)$.

Proof. There are $P, Q \in GL(n, R)$ such that $PAQ = D$ where D is diagonal matrix; $d_i | d_{i+1}$ ($1 \leq i \leq n-1$). Because $\det(PAQ) = \det(P)\det(A)\det(Q)$ and $\det(A) \neq 0$, $\det(D) \neq 0$. Consequently, $D = \text{diag}(d_1, \dots, d_n)$; $d_i \neq 0$ ($1 \leq i \leq n$). From $PAQ = D$, we have $PA = DQ^{-1}$. By applying **Theorem 2.10.** and **2.11.**, we get

$$\begin{aligned}\mathcal{B}_{\mathcal{M}}(A, x_0) &= \mathcal{B}_{\mathcal{M}}(PA, Px_0) \\ &= \mathcal{B}_{\mathcal{M}}(DQ^{-1}, Px_0) \\ &= Q(\mathcal{B}_{\mathcal{M}}(D, Px_0)).\end{aligned}$$

Because $d_i \neq 0$ ($1 \leq i \leq n$), $\mathcal{B}_{\mathcal{M}}(d_i, v_{0i}) \neq \emptyset$. Consequently, $\mathcal{B}_{\mathcal{M}}(A, x_0) = Q(\mathcal{B}_{\mathcal{M}}(D, Px_0)) \neq \emptyset$. Furthermore, because $d_i | d_{i+1}$ ($1 \leq i \leq n-1$), we have $\mathcal{B}_{\mathcal{M}}(d_1, v_{01}) \subseteq \mathcal{B}_{\mathcal{M}}(d_i, v_{0i})$ for each i ($1 \leq i \leq n-1$). If $d = d_1$, according to **Theorem 2.6.**, we get $(\ker(d_1))^n + \{g_0\} = (\mathcal{M}_{\tau(d)}^n + \{g_0\}) \subseteq \mathcal{B}_{\mathcal{M}}(D, Px_0)$ for some $g_0 \in \mathcal{M}^n$. Hence, we have $Q(\mathcal{M}_{\tau(d)}^n + \{g_0\}) \subseteq Q(\mathcal{B}_{\mathcal{M}}(D, Px_0)) = \mathcal{B}_{\mathcal{M}}(A, x_0)$.

2.3 Algorithm

The following activities are steps to determine the solution space of the equation systems.

- 1) Check wheter or not $\det(A) \neq 0$.
- 2) Determine $P, Q \in GL(n, R)$ such that $PAQ = D$ where D is diagonal matrix.
- 3) Compute $v_0 = Px_0$.
- 4) Determine a partial solution of equation: $d_i x_i = \text{row}_i(P)x_0$, $i = 1, \dots, n$ in $\mathcal{M}$.
Hence, we get $\mathcal{B}_{\mathcal{M}}(D, Px_0)$.
- 5) Compute $\mathcal{B}_{\mathcal{M}}(A, x_0) = Q(\mathcal{B}_{\mathcal{M}}(D, Px_0))$.

3. ITS APPLICATIONS

The following definitions and theorems are usefull to understanding **Example 3.21. – 3.23.** as application of the research results.

Theorem 3.15. [2, p. 307-308] Let R be a principal ideal ring and let $A \in R^{m \times n}$. Then there is $P \in GL(m, R)$ and $Q \in GL(n, R)$ such that $PAQ = \begin{pmatrix} D_r & 0 \\ 0 & 0 \end{pmatrix}$ where $r = \text{rank}(A)$ and $D_r = \text{diag}(s_1, \dots, s_r)$ with $s_i \neq 0$ and $s_i | s_{i+1}$ for $1 \leq i \leq r-1$.

Theorem 3.16. [1, p. 13] For each $\phi \in \mathbf{R}[\mathbf{d}] \setminus \{0\}$ and $f \in \mathcal{L} := C^\infty(\mathbf{R}, \mathbf{C})^3$ there is $g \in \mathcal{L}$ such that $\phi g = f$.

Definition 3.17. [1, p. 10]

(1) Denote by $H(\mathbf{C})$ the ring of entire function on the full complex plane.
 (2) For a subset $S \subseteq H(\mathbf{C})$ define the variety $\mathcal{V}(S)$ as the set of all common zeros of S , thus $\mathcal{V}(S) := \{\lambda \in \mathbf{C} \mid f(\lambda) = 0 \text{ for all } f \in S\}$. In Case, $S = \{f_1, \dots, f_n\}$ is finite, we simply write $\mathcal{V}(f_1, \dots, f_n)$ for $\mathcal{V}(S)$.

(3) For $q = \frac{p}{\phi} \in Q(\mathbf{R}[\mathbf{d}])[\sigma, \sigma^{-1}]$; $p = \sum_{i=0}^N \sum_{j=m}^M p_{ij} \mathbf{d}^i \sigma^j \in \mathbf{R}[\mathbf{d}, \sigma, \sigma^{-1}]$ and $\phi \in \mathbf{R}[\mathbf{d}] \setminus \{0\}$, denote by $q^* \in M(\mathbf{C})$ the meromorphic function given by

$$q^*(s) = \frac{\sum_{i=0}^N \sum_{j=m}^M p_{ij} s^i e^{-js}}{\phi(s)} = \frac{p(s, e^{-s})}{\phi(s)} \text{ for } s \in \mathbf{C} \setminus \mathcal{V}(\phi).$$

In case, q^* is entire, we call the set $\mathcal{V}(q^*)$ the characteristic variety and its elements the characterisitic zeros of q .

(4) For $f \in H(\mathbf{C})$ and $\lambda \in \mathbf{C}$ let $\text{ord}_\lambda(f) := \min\{k \in \mathbf{N}_0 \mid f^{(k)}(\lambda) \neq 0\}$ denote the multiplicity of λ as a zero of f . If $f = 0$, we put $\text{ord}_\lambda(f) = \infty$ for $\lambda \in \mathbf{C}$.

(5) $\mathcal{H} := \{\frac{p}{\phi} \mid p \in \mathbf{R}[\mathbf{d}, \sigma, \sigma^{-1}], \phi \in \mathbf{R}[\mathbf{d}] \setminus \{0\}, \frac{p^*}{\phi^*} \in H(\mathbf{C})\}$

(6) $\mathcal{H}_0 := \{\frac{p}{\phi} \mid p \in \mathbf{R}[\mathbf{d}, \sigma, \sigma^{-1}], \phi \in \mathbf{R}[\mathbf{d}] \setminus \{0\}, \frac{p^*}{\phi^*} \in H(\mathbf{C})\}$

Theorem 3.18. [1, p. 13] For $p \in \mathbf{R}[\mathbf{d}, \sigma, \sigma^{-1}]$ and $\phi \in \mathbf{R}[\mathbf{d}] \setminus \{0\}$ satysfying $\text{Ker}(\phi) \subseteq \text{Ker}(p)$ on $\mathcal{L}$ if and only if $\frac{p^*}{\phi^*} \in H(\mathbf{C})$.

Theorem 3.19. [1, p. 13] The operator $p \in \mathbf{R}[\mathbf{d}, \sigma, \sigma^{-1}] \setminus \{0\}$ is a surjective endomorphism on the space of exponential polynomials $\mathcal{L}_{e, \text{pol}}$. More precisely, let $a := \text{ord}_\lambda(p^*) \geq 0$. Then, for all $t^k e^{\lambda t} \in \mathcal{L}_{e, \text{pol}}$ there exist constants $\alpha_0, \dots, \alpha_{l+a} \in \mathbf{C}$ with $\alpha_{l+a} \neq 0$ such that $p\left(\sum_{i=0}^{l+a} \alpha_i t^i e^{\lambda t}\right) = t^k e^{\lambda t}$.

Theorem 3.20. [1, p. 36] Let R be any of the rings $\mathcal{H}$ or $\mathcal{H}_0$. Then R is an elementary divisor domain.

Example 3.21.

Let be the equation systems: $Ax = x_0$ where $A \in \mathbf{R}[\mathbf{d}]^{n \times n}$, $\det(A) \neq 0$, and $x_0 \in \mathcal{L}^n$. The existence of its solution space is guaranteed, because $\mathbf{R}[\mathbf{d}]$ is a principal ideal domain, $\mathbf{R}[\mathbf{d}]$ is also an elementary divisor domain, and $\mathcal{L}$ is a solvable module over $\mathbf{R}[\mathbf{d}]$. Applying **Theorem 3.20.** and **2.14.**, we have $\mathcal{B}_\mathcal{L}(A, x_0) = Q(\mathcal{B}_\mathcal{L}^{\mathcal{L}_{e, \text{pol}}}(D, Px_0))$ where D is diagonal matrix.

Example 3.22.

Let be the equation systems: $Ax = x_0$ where $A \in \mathbf{R}[\mathbf{d}, \sigma]^{n \times n}$, $\det(A) \neq 0$, and $x_0 \in \mathcal{L}_{e, \text{pol}}^n$. $\mathcal{H}_0$ is an elementary divisor domain (**Theorem 3.20.**). By applying **Theorem 3.19.**, $\mathcal{L}_{e, \text{pol}}$ as a solvable module over $\mathcal{H}_0$ can be showed. Because

³ The set of all smooth function from $\mathbf{R}$ to $\mathbf{C}$.

$\mathbf{R}[\mathbf{d}, \sigma] \subseteq \mathcal{H}_0$, we get $\mathcal{B}_{\mathcal{L}_{e, \text{pol}}}(A, x_0) = Q(\mathcal{B}_{\mathcal{L}_{e, \text{pol}}}(D, Px_0))$ for some $P, Q \in \text{GL}(n, \mathcal{H}_0)$ where D is diagonal matrix (**Theorem 2.14.**).

Example 3.23.

Let be the equation systems: $Ax = x_0$ where $A \in \mathbf{R}[\mathbf{d}, \sigma, \sigma^{-1}]^{n \times n}$, $\det(A) \neq 0$, and $x_0 \in \mathcal{L}_{e, \text{pol}}^n \cdot \mathcal{H}$ is an elementary divisor domain (**Theorem 3.20.**). By applying **Theorem 3.19.**, $\mathcal{L}_{e, \text{pol}}$ as a solvable module over $\mathcal{H}$ can be showed. Because $\mathbf{R}[\mathbf{d}, \sigma, \sigma^{-1}] \subseteq \mathcal{H}$, we get $\mathcal{B}_{\mathcal{L}_{e, \text{pol}}}(A, x_0) = Q(\mathcal{B}_{\mathcal{L}_{e, \text{pol}}}(D, Px_0))$ for some $P, Q \in \text{GL}(n, \mathcal{H})$ where D is diagonal matrix (**Theorem 2.14.**).

4. CONCLUDING REMARKS

This research results show that a characteristic of the solution space of the equation systems can be identified via an algebra approach. One original contribution of this work is application of module theory and systems over ring to solve the delay-diferensial equation systems. The characteristic of the solution space of the equation systems that previously required complex proofs can now be done in a straightforward and almost trivial manner.

Acknowledgement

This research is a part of A Graduate Team Research Program (Program Hibah Penelitian Tim Pascasarjana / HPTP) financed by The Directorate of Higher Education, The Ministry of National Education under contract No. LPP-UGM/LPPM-UGM/588/2007, March 29, 2007.

REFERENCES

1. H. Gluesing-Luerssen, *Linear Delay-Differential Systems with Commensurate Delays: An Algebraic Approach*, Springer-Verlag Berlin Heidelberg, Germany, 2002.
2. W.A. Adkins and S.H. Weintraub, *Algebra an Approach via Module Theory*, Springer-Verlag, New York, USA, 1992.

Ikrar Pramudya : PhD student in Gadjah Mada University.
E-mail: ikrarpramudya@yahoo.com.sg

Sri Wahyuni : Department of Mathematics, Faculty of Mathematics and Natural Sciences Gadjah Mada University, Yogyakarta, Indonesia.
E-mail: swahyuni@ugm.ac.id

Ari Suparwanto : Department of Mathematics, Faculty of Mathematics and Natural Sciences Gadjah Mada University, Yogyakarta, Indonesia.
E-mail: ari_suparwanto@yahoo.com

Widodo: Department of Mathematics, Faculty of Mathematics and Natural Sciences Gadjah Mada University, Yogyakarta, Indonesia.
E-mail: widodo@math.web.id

ON EDGE MAGIC TOTAL LABELING OF CATERPILLAR-LIKE TREES

KIKI A. SUGENG AND DENNY R. SILABAN

Abstract. Let $G = (V, E)$ be a simple, connected and undirected graph with v vertices and e edges. Let α be a bijection from $V \cup E$ to the set $\{1, 2, \dots, v + e\}$. An *edge magic total labeling* is a bijection α such that if xy is an edge of G then the weight of edge $\alpha(x) + \alpha(y) + \alpha(xy) = k$, for some integer constant k . A *super edge magic total labeling* is an edge magic total labeling α which $\alpha(V) = \{1, \dots, v\}$. A *caterpillar* is a graph derived from a path by hanging any number of leaves from the vertices of the path. In this paper we construct a super edge magic total labeling of a special classes of tree that we called caterpillar-like tree. Caterpillar-like tree is a tree which is constructed by embedded a caterpillar in a grid and then apply some edge replacements (elementary transformation) on the original backbone path of the caterpillar.

Key words and Phrases: Edge magic total labeling, super edge magic total labeling, caterpillar-like tree.

1. INTRODUCTION

In this paper, we consider simple finite undirected graph $G = (V, E)$. The set of vertices and edges denote by V and E respectively. Let G has v vertices and e edges. Let α be a bijection from $V \cup E$ to the set $\{1, 2, \dots, v + e\}$. An *edge magic total labeling* is a bijection α such that if xy is an edge of G then the weight of edge $\alpha(x) + \alpha(y) + \alpha(xy) = k$, for some integer constant k . A *super edge magic total labeling* is an edge magic total labeling α which $\alpha(V) = \{1, \dots, v\}$. A graph that has a (super) edge magic total labeling is called a (*super*) *edge magic total graph*.

2000 *Mathematics Subject Classification:* 05C78.

Sedláček introduced the concept of magic labeling in 1963 [8]. The notion of super edge-magic labeling was defined by Enomoto *et al.* in [5]. The famous conjecture that still has not revealed until now has been introduced by Kotzig and Rosa [7] as follows.

Conjecture 1 *Every tree is edge magic.*

and later on Enomoto *et al.* [5] proposed a stronger conjecture.

Conjecture 2 *Every tree is super edge magic.*

Since then, many papers have been published on super edge magic total labeling on several sub classes of trees, such as paths, caterpillars, banana trees and path-like trees. See [1, 4, 6] for details. In this paper, we generalised the idea of path-like trees to caterpillar-like trees.

2. PATH-LIKE TREES

Let P_n be a path with $V(P_n) = \{x_i : i = 1 \dots n\}$ and $E(P_n) = \{x_i x_{i+1} : i = 1 \dots n-1\}$. Let f be a bijection $f : V(P_n) \cup E(P_n) \rightarrow \{1, 2, \dots, 2n-1\}$ where f is a super magic total labeling. The path can be embedded in the two dimensional grid and consider a set of transformations which keep the edge-magic property of the paths. Let $P_n, n \geq 4$. Consider the ordered set of subpaths $S_1, S_2, \dots, S_t$ which are the longest straight segments in the embedding and the end of S_j is the beginning of S_{j+1} for any $j = 1, 2, \dots, t-1$. Suppose that $S_j = P_2$ for some j , $1 < j < t$, $V(S_j) = \{u_0, v_0\}$, $u_0 \in V(S_{j-1}) \cup V(S_j)$, $v_0 \in V(S_j) \cup V(S_{j+1})$ and that some vertex u of S_{j-1} is at distance one in the grid to a vertex v of S_{j+1} . The distance of u_0 and $u \in S_{j-1}$ is equal to the distance of v_0 and $v \in S_{j+1}$. An elementary transformation of the path consists in replacing the edge $u_0 v_0$ by a new edge uv . A tree T of order n is called *path-like tree* when it can be obtained from some embedding of P_n in the two dimensional grid by a set of elementary transformations. This concept of path-like tree was introduced by Barrientos in [2].

Bača, Lin and Muntaner-Batle [1] and Ngurah, Baskoro and Simanjuntak [4] in separate papers proved the following theorem.

Theorem 1 *All path-like trees are super edge magic total graph.*

3. CATERPILLAR-LIKE TREES

Let G be a caterpillar with centers $\{c_1, c_2, \dots, c_r\}$. We called the subgraph BP_r of G with $V(BP_r) = \{c_1, c_2, \dots, c_r\}$ and $E(BP_r) = \{c_i c_{i+1} : i = 1, \dots, r-1\}$ as a *backbone path* for caterpillar G . Using generalisation of path-like tree on a backbone path of caterpillar G we obtain a super edge magic total labeling on new subclass of trees that we called caterpillar-like trees.

The following Lemma will be needed to proof the main theorem.

Lemma 1 [3] *A graph G with p vertices and q edges is super edge magic graph if and only if there exists a bijective function $f : V(G) \cup E(G) \rightarrow \{1, 2, \dots, p+q\}$ such that the set $S = \{f(x) + f(y) | xy \in E(G)\}$ consists of q consecutive integers. In such a case, f extends to a super edge magic total labeling of G with magic constant $c = p + q + s$ where $s = \min(S)$ and $S = \{f(x) + f(y) | xy \in E(G)\} = \{c - (p+1), c - (p+2), \dots, c - (p+q)\}$.*

In the next theorem we prove that a subclass of tree that we called caterpillar-like tree is a super edge magic total graph.

Theorem 2 *All Caterpillar-like trees are super edge magic total graph.*

Proof. Let G be a caterpillar with c_i as center vertices, $i = 1, \dots, r$ and s as number leaves of every center. Let BP_r be a backbone path of order r . Thus $V(BP_r) = \{c_1, c_2, \dots, c_r\}$ and $E(BP_r) = \{c_i c_{i+1} : i = 1, \dots, r-1\}$. Let v_j^i be leaves of center c_i , $i = 1, 2, \dots, r$, $j = 1, \dots, s$.

Label the center vertices as follows.

$$\begin{aligned} f(c_i) &= 1 + \frac{(i-1)(s+1)}{2}, & i \text{ odd} \\ f(c_i) &= \frac{(r+i)(s+1)}{2}, & r \text{ even, } i \text{ even} \\ f(c_i) &= 1 + \frac{(r+i-1)(s+1)}{2}, & r \text{ odd, } i \text{ even} \end{aligned}$$

Label the leaves vertices as follows.

For i odd :

$$\begin{aligned} f(v_j^i) &= 1 + \frac{(i+r-2)(s+1)}{2} + j, & r \text{ odd} \\ f(v_j^i) &= 1 + \frac{(i+r-3)(s+1)}{2} + s + j, & r \text{ even} \end{aligned}$$

For i even :

$$f(v_j^i) = 1 + \frac{(i-2)(s+1)}{2} + j.$$

From Lemma 1, we obtain a super edge magic total labeling for caterpillar since the edge weight $W_f = \{f(x) + f(y), xy \in E\}$ form a consecutive integers. From the labeling f above, we have the following property

$$f(c_i) + f(c_{i+1}) = f(c_{i-p}) + f(c_{i+p+1}), \quad i = 1, \dots, r, p = 1, \dots, i. \quad (1)$$

Without loss of generalisation, assume that i and p are odds. Then $i - p$ is even and $i + p + 1$ is odd.

Case 1: If r is odd then

$$\begin{aligned} f(c_i) + f(c_{i+1}) &= 1 + \frac{(i-1)(s+1)}{2} + 1 + \frac{(r+i+1-1)(s+1)}{2} \\ &= 2 + \frac{(s+1)(i-1+r+i)}{2} \\ &= 2 + \frac{(s+1)(2i+r-1)}{2} \\ &= 2 + \frac{(s+1)(r+i-p-1+i+p)}{2} \\ &= 1 + \frac{(r+i-p-1)(s+1)}{2} + 1 + \frac{(i+p+1-1)(s+1)}{2} \\ &= f(c_{i-p}) + f(c_{i+p+1}). \end{aligned} \quad (2)$$

Case 2: If r is even then

$$\begin{aligned} f(c_i) + f(c_{i+1}) &= 1 + \frac{(i-1)(s+1)}{2} + \frac{(r+i+1)(s+1)}{2} \\ &= 1 + \frac{(s+1)(i-1+r+i+1)}{2} \\ &= 1 + \frac{(s+1)(r+i-p+i+p)}{2} \\ &= 1 + \frac{(s+1)(2i+r)}{2} \\ &= 1 + \frac{(s+1)(r+i-p+i+p)}{2} \\ &= \frac{(r+i-p)(s+1)}{2} + 1 + \frac{(i+p+1-1)(s+1)}{2} \\ &= f(c_{i-p}) + f(c_{i+p+1}). \end{aligned}$$

Embed BP_r on a two dimensional grid and let $S_1, S_2, \dots, S_k$ be subpaths of BP_r which are maximal straight segment in the embedding. Suppose that $S_j = P_2$ for some $j \in \{1, 2, \dots, k\}$. An elementary edge transformation is a process to move

edge $c_j c_{j+1} \in E(S_j)$ to be edge $c_{j-p} c_{j+1-p}$. Using Equation [1] then there is a guarantee that the movements (elementary edge transformation) do not affect the edge weight set W_f . Let $T_0 = BP_r, T_1, T_2, \dots, T_t = T$ be the series of trees obtained by successively applying the appropriate elementary transformations of BP_r to obtain a caterpillar-like tree T . The caterpillar-like tree T will have the same edge weight set as BP_r . Thus the caterpillar-like tree T is an edge magic total graph. $\square$

Figure 1 shows two examples of caterpillar-like trees on 21 centers with two leaves on every center.

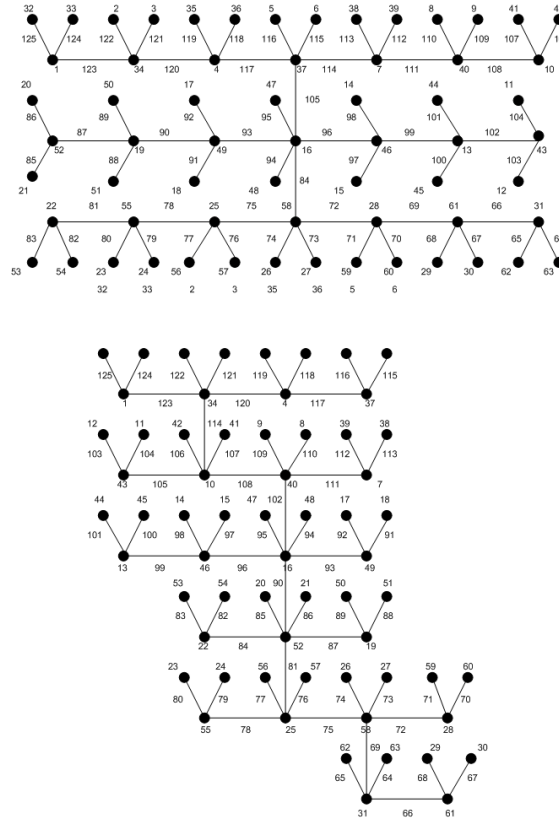


Figure 1: Examples of caterpillar-like tree.

4. CONCLUSION

In this paper we construct a super edge magic total labeling on a new subclass of tree. However, the original conjecture for the generalised tree is still open. Thus

there are still open problems that we can explore more, especially to find a more subclass of tree that has super edge magic total labelings.

REFERENCES

1. Bača, Y. Lin and F.A. Muntaner-Batle, Super edge-antimagic labelings of the path-like trees, *Utilitas Mathematica*, to appear.
2. C. Barrientos, Difference Vertex Labelings, Ph. D. Thesis, Universitat Politècnica de Catalunya (2004).
3. R.M. Figuerora-Centeno, R. Ichisima and F.A. Muntaner-Batle, The place of super edge magic labeling among other classes of labelings, *Discrete Math.* **231** (2001) 153-168.
4. AAG Ngurah, E.T. Baskoro, R. Simanjuntak, On (super) edge magic total labeling of graph, Seminar Nasional Matematika, Jakarta, Indonesia Nov. 2005.
5. H. Enomoto, A.S. Llado, T. Nakamigawa and G. Ringel, Super edge-magic graphs, *SUT J. Math.* **34** (1998) 105-109.
6. J. Gallian, A dynamic survey of graph labeling, *The Electronic Journal of Combinatorics* **6** (2003), DS 6.
7. A.Kotzig and A.Rosa, Magic valuations of finite graphs, *Canad. Math.Bull.* **13** (1970), 451-461.
8. J. Sedláček, *Problem 27*, in Theory of Graphs and Its Applications, *Proc. Symposium Smolenice*, June (1963) 163-167.

KIKI A. SUGENG AND DENNY R. SILABAN:

Department of Mathematics, University of Indonesia, Depok 16424, Indonesia.

E-mail: {kiki,denny-rs}@ui.edu.

SEMIPRIME SUBMODULES CHARACTERIZATIONS

SUPRAPTO AND INDAH EMILIA WIJAYANTI

Abstract. In this paper, we characterize a semiprime submodule based on properties in [3]. Our main purpose is to obtain a characterization of a semiprime submodules related to the annihilator of its submodules and the annihilator of every element of R -module M as a semiprime ideal in its base ring.

Key words : *semiprime submodules, semiprime ideal*

1. INTRODUCTION

The characterization of a semiprime module related to its annihilator is discussed. We denote R as a commutative and unital ring, and by a module we mean a right R -module. The ideal of a ring is denoted by " $\triangleleft$ " and the submodule of a module is denoted by " $<$ ". A submodule $K < M$ is called prime if for any $\beta \in M$ and any $t \in R$, if $\beta Rt \subseteq K$, then $\beta \in K$ or $t \in \text{ann}_R(M/K)$, where

$$\text{ann}_R(M/K) = \{r \in R \mid Mr \subseteq K\}$$

is an annihilator ideal of M/K in R .

We observe the characterization of a semiprime submodule related to the annihilator of submodules and the annihilator of elements in module M . The annihilators are semiprime ideals in the ring R . The submodule $K < M$ is called semiprime if for any $\beta \in M$ and $s \in R$, if $\beta sRs \subseteq K$, then $\beta s \in K$. A right R -modul M is called prime if for every $0 \neq K < M$, $\text{ann}_R(K) = \text{ann}_R(M)$.

2. SEMIPRIME SUBMODULES

Definition 2.1. Let a ring R with unit element. For any right R -module M , a submodule $K < M$ is called a prime submodule if for every $\alpha \in M \setminus K$ and every $t \in R$, $\alpha Rt \subseteq K$, then $t \in \text{ann}_R(M/K)$. A right R -module M is called a prime if $0 < M$ is a prime submodule.

Definition 2.2. Let R be a ring with unit element. A right ideal $L \triangleleft R$ is semiprime if for any $s \in R \setminus L$, then $sRs \not\subseteq L$. (Its contraposition is if $sRs \subseteq L$, then $s \in L$).

Theorem 2.3. (Dauns 1.3, [1]) For a right R -module M , a submodule $K < M$, and the ideal annihilator $J =: \text{ann}_R(M/K)$ in R , the following are all equivalent :

1. $K < M$ is a prime.
2. $\forall \alpha \in M \setminus K$ and $\forall t \in R$, $\alpha Rt \subseteq K \Rightarrow t \in J$.
3. For every a submodule $0 \neq Q \subset M/K$, $\text{Ann}_R(Q) = J$.
4. $\forall V < M$, $\forall A \triangleleft R$, $VA \subseteq K \Rightarrow V \subseteq K$ or $A \subseteq J$.
5. $\forall K < W \leq M$, $K \neq W$ and $\forall J \triangleleft B \triangleleft R$, $J \neq B \Rightarrow WB \not\subseteq K$.

Proposition 2.4. (Dauns 1.5, [1]) An R -module M is prime if and only if every submodule $0 \neq K < M$, satisfies $\text{ann}_R(K) = \text{ann}_R(M)$.

Proposition 2.5. (Suprpto 2.6, [3]) If an R -module M is prime, then every submodule $K < M$ is prime.

Definition 2.6. Let M be a right R -module. For a submodule $K < M$ and for arbitrary $\alpha, \beta, \gamma \in M$ and $s \in R$, then a submodul $K < M$ is called :

1. *strong semiprime* if $\alpha sRs \subseteq K$, then $\alpha Rs \subseteq K$.
2. *semiprime* if $\beta sRs \subseteq K$, then $\beta s \in K$.
3. *weak semiprime* if $\gamma RsRs \subseteq K$, then $\gamma Rs \subseteq K$.

very weak semiprime if $\gamma RsRs \subseteq K$, then $\gamma s \in K$.

This statement is clear : $1 \Rightarrow 2 \Rightarrow 3 \Rightarrow 4$. Furthermore : *prime* $\Rightarrow$ *strong semiprime*.

Theorem 2.7. Let M be a right R -module, a submodule $0 \neq K < M$ and R is a commutative ring with unit element. The following are all equivalent :

1. $K < M$ is semiprime.
2. $\forall \beta \in M$ and $\forall s \in R$, $\beta sRs \subseteq K \Rightarrow \beta s \in K$.
3. For any submodule $0 \neq Q < M/K$, $\text{ann}_R(Q)$ is a semiprime ideal in R .
4. $\forall 0 \neq \beta \in M$, $\text{ann}_R(\beta + K)$ is a semiprime ideal in R .

Proof :

(1 $\Leftrightarrow$ 2) It is clear by Definition 2.6.

(2 $\Rightarrow$ 3) Take any $0 \neq (\beta + K) \in M/K$ then $0 \neq (\beta + K)R = (\beta R + K) \subseteq M/K$. Take any $s \in R \setminus \text{ann}_R(\beta R + K)$ that

satisfies $sRs \subseteq \text{ann}_R(\beta R + K)$ then $\beta RsRs \subseteq K$. Since R is a commutative ring, then $\beta RsRs = \beta sRRs = \beta sRs \subseteq K$. Furthermore $\beta s \in K$ or $s \in \text{ann}_R(\beta R + K)$.

($2 \Leftarrow 3$) By Definition 2.2, for any $s \in R \setminus \text{ann}_R(\beta R + K)$ if $sRs \subseteq \text{ann}_R(\beta R + K)$, then $s \in \text{ann}_R(\beta R + K)$. Or if $\beta RsRs \subseteq K$, then $\beta Rs \subseteq K$. Take any $\beta \in M$ and $s \in R$ that satisfies $\beta RsRs \subseteq K$. By hypothesis, so $\beta Rs \subseteq K$. Take $1 \in R$, then $\beta 1s = \beta s \in K$.

($3 \Rightarrow 4$) Take any $0 \neq (\beta + K) \in M/K$ and any $s \in R \setminus \text{ann}_R(\beta + K)$ that satisfies $sRs \subseteq \text{ann}_R(\beta + K)$ then $\beta sRs \subseteq K$. Since $K < M$ is a submodule, then $(\beta sRs)R \subseteq K$. Since R is a commutative ring, then $(\beta sRs)R = \beta sRsR = \beta sRRs = \beta sRs \subseteq K$. By hypothesis, then $\beta s \in K$, so $s \in \text{ann}_R(\beta + K)$.

($3 \Leftarrow 4$) Take any $0 \neq (\beta R + K) \subseteq M/K$ and any $s \in R \setminus \text{ann}_R(\beta R + K)$ that satisfies $sRs \subseteq \text{ann}_R(\beta R + K)$ so $\beta RsRs \subseteq K$. Since R is a commutative ring, then $\beta RsRs = \beta sRRs = \beta sRs \subseteq K$, so $\beta s \in K$. Since $K < M$ is submodule, then $(\beta s)R \subseteq K$. So $(\beta s)R = \beta sR = \beta Rs \subseteq K$ or $s \in \text{ann}_R(\beta R + K)$.

3. CONCLUDING REMARKS

The commutativity of the ring R with unit element, give the possibilities to get the characterizations of semiprime submodules. Moreover, this statement will give us the opportunities to observe the characterization of semiprime modules related to the annihilator of submodules and the annihilator of elements in the module.

Acknowledgement. I would like to thank Indah Emilia Wijayanti, Ph.D. for the suggestions and encouragement to inspire the author in research.

REFERENCES

1. Dauns, J., *Prime modules*, Journal für Mathematik, vol. 298, pg. 156-181, 1978.
2. R. Wisbauer, *"Foundations of modules and ring theory"*. A handbook for study and research, Gordon and Breach Science Publishers, Philadelphia, PA, 1991.
3. Suprpto dan Wijayanti, I.E., *Modul Semiprime*, A paper presented in Seminar Nasional Mahasiswa S3 Matematika, ITB, Bandung, 2007.
4. Wijayanti, I.E., "Coprimeness Modules and Comodules" Ph.D. Thesis, University of Düsseldorf, Germany, 2006.

SUPRAPTO, I.E. WIJAYANTI

SUPRAPTO : Junior High School (SMP) 1 Banguntapan, Bantul, Yogyakarta,
55197, Indonesia.

E-mail : suprpto_72@yahoo.com

I.E. WIJAYANTI : Department of Mathematics, Faculty of Mathematics and
Natural Sciences, Gadjah Mada University, Yogyakarta 55281, Indonesia.

E-mail : ind_wijayanti@yahoo.com

ON PRIME $\mathcal{L}$ -SUBCOMODULES

INDAH EMILIA WIJAYANTI

Abstract. The study of prime ideals and prime submodules has been established well in module theory. One of the techniques used to construct the primeness is defining first the 'multiplication' in a module to define prime submodules (modules). Meanwhile the notion of primeness of $\mathcal{L}$ -ideals and $\mathcal{L}$ -submodules have been studied by some authors, which is based on the classical ring theory. In our work, we give another point of view defining prime $\mathcal{L}$ -submodules. The motivation is the idea of Dauns which used the term of annihilator to obtain the definition of primeness. Then we define $\mathcal{L}$ -annihilator as a tool to define prime $\mathcal{L}$ -submodules. Our further objective is to transfer the results of primeness $\mathcal{L}$ -submodules to $\mathcal{L}$ -subcomodules.

Key words and Phrases: prime submodules, annihilator, prime $\mathcal{L}$ -submodules, $\mathcal{L}$ -annihilators, prime $\mathcal{L}$ -subcomodules.

1. INTRODUCTION

In module theory, prime submodules has been studied and these fact motivates us to define and investigate prime $\mathcal{L}$ -submodules, where $\mathcal{L}$ is complete Heyting algebra (complete lattice). The notions and results could be formulated by transferring the corresponding knowledge from modules to comodules. So we concern to construct notion of primeness in $\mathcal{L}$ -subcomodules.

In study of fuzzy algebra, the notion of primeness of fuzzy ideals has been studied by Mordeson and Malik in [7], which is based on the classical ring theory. We refer to [7] and [6] for basic notions of fuzzy algebras.

The study of prime submodules has been established well in module theory, for example in [3], [12], [14] and [9]. Similar with the technique in ring theory, one defined first the 'multiplication' in a module to define prime submodules (modules) (see [3]). Using these technique, Acar defined prime fuzzy submodules and investigated the related properties (see [2]).

In our work, we give another point of view defining prime fuzzy submodules, that is transferring the idea of Dauns, by using an annihilator. Some studies of primeness of modules related with these approach were doing by Wisbauer ([12], [14]) and Wijayanti ([9]). Based on the results in module theory we observe the properties of prime $\mathcal{L}$ -submodules. We also concern by investigating the primeness of any $\mathcal{L}$ -ideal of R by viewing as an $\mathcal{L}$ -submodule of R as well. The relationship between the primeness of an $\mathcal{L}$ -submodule and the primeness of a submodule of N is also the case.

Since we need the *fuzzification* of terms quotient modules and annihilators, we extend the quotient $\mathcal{L}$ -subrings into quotient $\mathcal{L}$ -submodules and define $\mathcal{L}$ -annihilators using the product of elements in $\mathcal{L}$ -ideals of the base ring. Then we define prime $\mathcal{L}$ -submodules by these $\mathcal{L}$ -annihilator (see [11]).

The theory of coalgebras over fields and their comodules is discussed in various text books (see for example Abe [1], Sweedler [8] and Dăscălescu et.al. [5]). The study of coalgebras over commutative rings is presented in Brzeziński-Wisbauer [4].

As well known, definition of comodules is a dualization of definition of modules. We adopt the analogous way of defining $\mathcal{L}$ -submodules to obtain a definition of $\mathcal{L}$ -subcomodules. First we construct a comultiplication to any element in the set of all $\mathcal{L}$ -submodules as a dual of a product determined in $\mathcal{L}$ -submodules. By these comultiplication we give a basic concept of $\mathcal{L}$ -subcomodules (see [10]).

Our objective is to transfer the results in $\mathcal{L}$ -submodules to $\mathcal{L}$ -subcomodules.

2. QUOTIENT $\mathcal{L}$ -SUBMODULES AND $\mathcal{L}$ -ANNIHILATORS

Let $\mathcal{L}$ be a complete Heyting algebra, that is a complete Lattice such that for all $A \subseteq L$ and for all $b \in \mathcal{L}$ hold

$$\begin{aligned}\bigvee \{a \wedge b \mid a \in A\} &= (\bigvee \{a \mid a \in A\}) \wedge b, \\ \bigwedge \{a \vee b \mid a \in A\} &= (\bigwedge \{a \mid a \in A\}) \vee b.\end{aligned}$$

The meet, join and partial ordering of $\mathcal{L}$ will be written as $\wedge$, $\vee$ and $\leq$ respectively. The closed interval $[0, 1]$ together with the operations $\min$, $\max$ and $\leq$ form a complete Heyting lattice. By an $\mathcal{L}$ -subset of X , we mean a function from X to L and the set of all $\mathcal{L}$ -subset of X is denoted by $\mathcal{L}^X$.

When $Y \subseteq X$ and $a \in \mathcal{L}$, we define $a_Y \in \mathcal{L}^X$ as follows :

$$a_Y = \begin{cases} a & \text{for } x \in Y \\ 0 & \text{for } x \notin Y \end{cases}$$

In particular, if Y is a singleton, say $\{y\}$, then $a_{\{y\}}$ is often referred to as an $\mathcal{L}$ -point.

Throughout we assume R be a commutative ring with unit 1 and N is a left R -module. Denote 0_N is the zero element of N .

Let $\xi \in \mathcal{L}^R$. Then μ is called an $\mathcal{L}$ -ideal of R if for any $x, y \in R$ it satisfies the conditions :

- (a) $\xi(x - y) \geq \xi(x) \wedge \xi(y)$.
- (b) $\xi(xy) \geq \xi(x) \vee \xi(y)$.

We denote the set of all $\mathcal{L}$ -ideals of R by $\mathcal{LI}(R)$.

Consider now several operations of $\mathcal{L}$ -subset of a module induced by addition and scalar multiplication.

Definition 2.1. Let $\mu, \nu \in \mathcal{L}^N$. For any $x \in N$ and $r \in R$ define

$$(\mu + \nu)(x) := \vee \{ \mu(y) + \nu(y) \mid y, z \in N, y + z = x \} \quad (1)$$

$$(-\mu)(x) := \mu(-x) \quad (2)$$

$$(r\mu)(x) := \vee \{ \mu(y) \mid y \in N, ry = x \}. \quad (3)$$

Then $\mu + \nu$ is called *sum* of μ and ν , and $r\mu$ is called *product* of r and μ . We recall the notion of $\mathcal{L}$ -submodules.

Definition 2.2. An $\mathcal{L}$ -submodule of N is an $\mathcal{L}$ -subset $\mu \in \mathcal{L}^N$ such that

- (a) $\mu(0_N) = 1$.
- (b) $\mu(rx) \geq \mu(x)$ for all $r \in R$ and $x \in N$.
- (c) $\mu(x + y) \geq \mu(x) \wedge \mu(y)$ for all $x, y \in N$.

When $\mathcal{L} = [0, 1]$, an $\mathcal{L}$ -submodule is called a *fuzzy submodule*. We denote the set of all $\mathcal{L}$ -submodules of N by $\mathcal{L}(N)$.

Following the construction of quotient $\mathcal{L}$ -subrings in Section 3.2 of [7], we define quotient $\mathcal{L}$ -submodules as follow.

Definition 2.3. Let $\mu \in \mathcal{L}(N)$ and $x \in N$. Then $\mu(0_N)_{\{x\}} + \mu$ is called a coset of μ .

Let $x, y \in N$. Then

$$\begin{aligned} (\mu(0_N)_{\{x\}} + \mu)(y) &= \vee \{ \mu(0_N)_{\{x\}}(u) \wedge \mu(v) \mid u, v \in N, u + v = y \} \\ &= \mu(0_N)_{\{x\}}(x) \wedge \mu(x - y) \\ &= \mu(x - y). \end{aligned}$$

We write $x + \mu$ for $\mu(0_N)_{\{x\}} + \mu$ and conclude that $(x + \mu)(y) = \mu(x - y)$ for all $y \in N$.

Lemma 2.4. Let $\mu \in \mathcal{L}(N)$ and $x, y \in N$. Then $x + \mu = y + \mu$ if and only if $\mu(x - y) = \mu(0_N)$.

Proof. Suppose $x + \mu = y + \mu$, then $(x + \mu)(x) = (y + \mu)(x)$ and $(x + \mu)(y) = (y + \mu)(y)$. Hence $\mu(0_N) = \mu(y - x) = \mu(x - y)$.

Conversely, let $\mu(0_N) = \mu(x - y)$. By definition we have $(x + \mu)(y) = \mu(0_N)$ and for any $z \in N$ holds

$$\begin{aligned} (x + \mu)(z) &= \mu(z - x) = \mu(z - y + y - x) \\ &\geq \mu(z - y) \wedge \mu(y - x) \\ &= \mu(z - y) = (y + \mu)(z). \end{aligned}$$

Thus $x + \mu \supseteq y + \mu$. Similarly we obtain $y + \mu \supseteq x + \mu$. Let $N/\mu := \{x + \mu \mid x \in N\}$.

We define the addition and scalar multiplication on N/μ as follows :

$$\begin{aligned} (x + \mu) + (y + \mu) &:= (x + y) + \mu, \\ r(x + \mu) &:= rx + \mu, \end{aligned}$$

for any $x, y \in N$ and $r \in R$. We now show that those operations are well defined. Let $x, y, u, v \in N$ and $r, s \in R$. Suppose $x + \mu = u + \mu$ and $y + \mu = v + \mu$. Then $\mu(x - u) = \mu(y - v) = \mu(0_N)$. Hence

$$\mu(x + y - u - v) \geq \mu(x - u) \wedge \mu(y - v) = \mu(0_N).$$

Thus $\mu(x + y - u - v) = \mu(0_N)$, i.e. $(x + y) + \mu = (u + v) + \mu$. Now

$$\mu(rx - ru) = \mu(r(x - u)) \geq \mu(x - u) = \mu(0_N).$$

Thus $\mu(rx - ru) = \mu(0_N)$, i.e. $rx + \mu = ru + \mu$.

Clearly, $-(x + \mu) = (-x) + \mu$ and $0_N + \mu = \mu$ is the zero element of N/μ . To show that our definitions satisfy the axioms of an R -module, we consider :

- (a) $r((x + \mu) + (y + \mu)) = r((x + y) + \mu) = (r(x + y)) + \mu = (rx + ry) + \mu = (rx + \mu) + (ry + \mu)$.
- (b) $(r + s)(x + \mu) = (rx + sx) + \mu = (rx + \mu) + (sx + \mu)$.
- (c) $(rs)(x + \mu) = (rs(x)) + \mu = (r(sx)) + \mu = r(sx + \mu)$.
- (d) $1_R(x + \mu) = (1_R x) + \mu = x + \mu$.

It proved that N/μ is an R -module.

Definition 2.5. Let $\mu \in \mathcal{L}(N)$. Then the R -module $N/\mu = \{x + \mu \mid x \in N\}$ is called the *quotient module* of N by μ .

For any $\mu \in \mathcal{L}(N)$ put

$$\mu_* := \{x \in N \mid \mu(x) = \mu(0_N)\}.$$

If $\mathcal{L}$ is regular, then μ_* is a submodule of N (see Section 4.2 of [7]).

Lemma 2.6. *Let $\mu \in \mathcal{L}(N)$. If $\mathcal{L}$ is regular, then $N/\mu \simeq N/\mu_*$.*

Proof. Define $f : N \rightarrow N/\mu$ by $x \mapsto x + \mu$. Then f is a surjective homomorphism and $\text{Ker } f = \mu_*$, since for any $x \in \text{Ker } f$, $f(x) = x + \mu = 0_N + \mu$ implies $\mu(x - 0_N) = \mu(0_N)$. Thus $N/\mu \simeq N/\mu_*$.

Similarly we can define $N/U = \{n + \mu \mid n \in N, \mu \in U\}$ as a quotient $\mathcal{L}$ -submodule, where $U \subseteq \mathcal{L}(N)$.

In order to define a multiplication between an element in $\mathcal{L}$ -subset of ring R and an element in $\mathcal{L}$ -subset of module N , we recall the two kind of multiplication below.

Let $\xi \in \mathcal{L}^R$ and $\mu \in \mathcal{L}^N$. For any $x \in N$, recall the following products :

$$(\xi\mu)(x) := \vee\{\xi(r) \wedge \mu(y) \mid r \in R, y \in N, ry = x\},$$

$$(\xi \odot \mu)(x) := \vee\{\wedge_{i=1}^n (\xi(r_i) \wedge \mu(x_i)) \mid r_i \in R, x_i \in N, 1 \leq i \leq n, n \in \mathbb{N}, \sum_{i=1}^n r_i x_i = x\}.$$

The former multiplication is a special case of the latter one, that is when $n = 1$.

For any $\mu \in \mathcal{L}(N)$, put

$$\mathcal{L} - \text{Ann}(\mu) := \{\xi \in \mathcal{LI}(R) \mid \xi \odot \mu = 0\}.$$

We call these set as an $\mathcal{L}$ -annihilator of μ . Similarly, for any $U \subseteq \mathcal{L}(N)$ we define

$$\mathcal{L} - \text{Ann}(U) := \{\xi \in \mathcal{LI}(R) \mid \xi \odot \mu = 0, \forall \mu \in U\}.$$

Let $\xi \in \mathcal{LI}(R)$ and $\mu \in \mathcal{L}(M)$. Then $\xi \odot \mu \in \mathcal{L}(M)$. It is A straightforward property referred from Theorem 4.1.16 of [7]. Suppose $M = R$, if $\xi, \mu \in \mathcal{LI}(R)$, then $\xi \odot \mu \in \mathcal{LI}(R)$.

Lemma 2.7. *Let $\xi, \mu \in \mathcal{L}^R$. If $\mu \neq 0$ and $\xi \odot \mu = 0$, then for any $x \in R$, where $\sum_{i=1}^n r_i x_i = x$, it holds $\xi(r_i) = 0$, for $i = 1, 2, \dots, n$.*

Proof. Following the definition of the multiplication and the assumption, for any $x \in R$ holds

$$(\xi \odot \mu)(x) = \vee\{\wedge_{i=1}^n (\xi(r_i) \wedge \mu(x_i)) \mid r_i, x_i \in R, 1 \leq i \leq n, n \in \mathbb{N}, \sum_{i=1}^n r_i x_i = x\} = 0.$$

Since $\mu(x_i) \neq 0$ for $i = 1, 2, \dots, n$, $\vee\{\wedge_{i=1}^n (\xi(r_i))\} = 0$. Thus $(\xi(r_i) = 0$ for $i = 1, 2, \dots, n$.

3. PRIME $\mathcal{L}$ -SUBMODULES

According to the Definition 1.2.1 of Wijayanti [9], it gives us a possibility to define the primeness of an $\mathcal{L}$ -submodule as follow.

Definition 3.1. Let $\mu \in \mathcal{L}(N)$. Then μ is called a *prime $\mathcal{L}$ -submodule* if for any nonzero submodule of N/μ has the same $\mathcal{L}$ -annihilator, that is $\mathcal{L} - \text{Ann}(N/\mu)$.

The subset $U \subseteq \mathcal{L}(N)$ is a *prime set* if all the elements of U are prime $\mathcal{L}$ -submodule of N . The quotient $\mathcal{L}$ -submodule N/U is called *prime* if U is a prime set of $\mathcal{L}(N)$.

Proposition 3.2. For a module N and $\mu \in \mathcal{L}(N)$ the following are equivalent:

- (a) N/μ is a prime module.
- (b) $\mathcal{L} - \text{Ann}(X/\mu) = \mathcal{L} - \text{Ann}(N/\mu)$ for any submodule X/μ of N/μ .

Proof. It is known that N/μ is a prime module means μ is a prime $\mathcal{L}$ -submodule of N . Then by definition we have a trivial biimplication from (a) to (b).

We recall the $\mathcal{L}$ -prime ideal of the ring R following Definition 3.5.1 and Theorem 3.5.2 of [7]. An $\mathcal{L}$ -ideal ξ of ring R is called prime if ξ is non-constant and for any $\mu, \nu \in \mathcal{LI}(R)$, if $\mu \odot \nu \subseteq \xi$, then $\mu \subseteq \xi$ or $\nu \subseteq \xi$. The quotient $\mathcal{L}$ -subring R/μ is prime if μ is prime.

Proposition 3.3. Let R be a ring and N be an R -module. If for any $\mathcal{L}$ -submodule ν of N , $\mathcal{L} - \text{Ann}(\nu) \subseteq \mathcal{L} - \text{Ann}(\mu)$ or $\mathcal{L} - \text{Ann}(N/\nu) \subseteq \mathcal{L} - \text{Ann}(\mu)$, then $R/\mathcal{L} - \text{Ann}(\mu)$ is a prime ring.

Proof. Take any two $\mathcal{L}$ -ideals ν, η of R with $\nu \odot \eta \subset \mathcal{L} - \text{Ann}(\mu)$. Then $\nu \odot \eta \odot \mu = 0$.

Since $\eta \odot \mu$ is an $\mathcal{L}$ -submodule, by assumption $\mathcal{L} - \text{Ann}(\eta \odot \mu) = \mathcal{L} - \text{Ann}(\mu)$ or $\mathcal{L} - \text{Ann}(N/(\eta \odot \mu)) = \mathcal{L} - \text{Ann}(\mu)$.

If $\eta \odot \mu = \mu$, then $\mathcal{L} - \text{Ann}(N/(\eta \odot \mu)) \neq \mathcal{L} - \text{Ann}(\mu)$, hence $\nu \subset \mathcal{L} - \text{Ann}(\eta \odot \mu) = \mathcal{L} - \text{Ann}(\mu)$.

If $\eta \odot \mu \neq \mu$ and $\nu \not\subset \mathcal{L} - \text{Ann}(\mu)$, then $\nu \subset \mathcal{L} - \text{Ann}(\eta \odot \mu) \not\subset \mathcal{L} - \text{Ann}(\mu)$. Thus $\eta \subset \mathcal{L} - \text{Ann}(N/(\eta \odot \mu)) = \mathcal{L} - \text{Ann}(\mu)$.

Corollary 3.4. Let R be a ring, N an R -module, and $\mu \in \mathcal{LI}(R)$. If $\mathcal{L} - \text{Ann}(\mu)$ is singleton, say $\{\nu\}$, and for any $\xi \in \mathcal{LI}(R)$, $\mathcal{L} - \text{Ann}(\xi) = \{\nu\}$ or $\mathcal{L} - \text{Ann}(R/\xi) = \{\nu\}$, then R/ν is a prime ring.

4. COALGEBRAS AND COMODULES

The theory of coalgebras over fields and their comodules is discussed in various text books (see for example Abe [1], Sweedler [8] and Dăscălescu et.al. [5]). The study of coalgebras over commutative rings is presented in Brzeziński-Wisbauer [4].

An R -coalgebra is an R -module C with R -linear maps

$$\Delta : C \rightarrow C \otimes_R C \text{ and } \varepsilon : C \rightarrow R$$

called (*coassociative*) *coproduct* and *counit* respectively, with the properties

$$(I_C \otimes \Delta) \circ \Delta = (\Delta \otimes I_C) \circ \Delta \text{ and } (I_C \otimes \varepsilon) \circ \Delta = I_C = (\varepsilon \otimes I_C) \circ \Delta.$$

We use Sweedler's notation, that is for any $c \in C$, we write $\Delta(c) = \sum c_1 \otimes c_2$. For any coalgebra C and R -algebra A , $\text{Hom}_R(C, A)$ is an R -algebra by the convolution product, that is for any $f, g \in \text{Hom}_R(C, A)$ and $c \in C$ define

$$f * g = \mu \circ (f \otimes g) \circ \Delta, \text{ i.e. } f * g(c) = \sum f(c_1)g(c_2).$$

In particular, $C^* = \text{Hom}_R(C, R)$ is called the *dual algebra* of C .

A *right C -comodule* is an R -module M with an R -linear map $\varrho^M : M \rightarrow M \otimes_R C$ called a *right C -coaction*, with the properties

$$(I_M \otimes \Delta) \circ \varrho^M = (\varrho^M \otimes I_C) \circ \varrho^M \text{ and } (I_M \otimes \varepsilon) \circ \varrho^M = I_M.$$

To denote the action of ϱ^M on elements of M we write $\varrho^M(m) = \sum m_0 \otimes m_1$.

An R -submodule K of a right C -comodule M is called a *C -subcomodule* of M provided K has a right comodule structure such that the inclusion is a comodule morphism. In general, the fact that K is an R -submodule of M does not imply that $K \otimes_R C$ is a submodule of $M \otimes_R C$, since the tensor functor is not left exact. However, if K is a C -pure R -submodule of M then $K \otimes_R C \subset M \otimes_R C$, and K is a subcomodule of M provided $\varrho^M(K) \subset K \otimes_R C \subset M \otimes_R C$.

An R -submodule $D \subset C$ is called a *subcoalgebra* provided D has a coalgebra structure and the inclusion map is a coalgebra morphism. Any coalgebra C can be considered as a right C -comodule, and there is a close relationship between comodules of C and modules of C^* .

Any $M \in \mathbf{M}^C$ is a (unital) left C^* -module by

$$\rightarrow : C^* \otimes_R M \rightarrow M, \quad f \otimes m \mapsto (I_M \otimes f) \circ \varrho^M(m) = \sum m_0 f(m_1)$$

and any morphism $h : M \rightarrow N$ in $\mathbf{M}^C$ is a left C^* -module morphism, i.e.,

$$\text{Hom}^C(M, N) \subset {}_{C^*}\text{Hom}(M, N).$$

C is a subgenerator in $\mathbf{M}^C$, that is all C -comodules are subgenerated by C as C -comodules and C^* -modules.

Denote by ${}_{C^*}\mathbf{M}$ the category of left C^* -modules. The above observation shows that there is a faithful functor from $\mathbf{M}^C$ to ${}_{C^*}\mathbf{M}$. $\mathbf{M}^C$ is a full subcategory of ${}_{C^*}\mathbf{M}$ when all the (left) C^* -linear maps between right C -comodules arise from (right) C -comodule morphisms, or when $\mathbf{M}^C$ is isomorphic to $\sigma[{}_{C^*}C]$, that is the full subcategory of left C^* -modules whose objects are ${}_{C^*}C$ -subgenerated. The necessary and sufficient condition such that $\mathbf{M}^C \simeq \sigma[{}_{C^*}C]$ is the coalgebra C should be a locally projective module (Section 4.3 of [4]).

According to 3.7 of [4], coproducts of comodules in $\mathbf{M}^C$ is again a comodule.

5. PRIME $\mathcal{L}$ -SUBCOMODULES

We define first the comultiplication (or coaction) of an $\mathcal{L}$ -submodules and tensor product of two $\mathcal{L}$ -submodules as follow.

Definition 5.1. Let M, K be a C -comodule, $\mu \in \mathcal{L}^M$ and $\nu \in \mathcal{L}^K$. For any $m \in M$ and $x \in M \otimes_R K$ we define

$$\begin{aligned}\mu^M(m) &:= \vee \{ \wedge_{i \in I} (\mu_1(m_{\underline{0}}) \wedge \mu_2(m_{\underline{1}})) \mid \mu_1 \in \mathcal{L}^M, \mu_2 \in \mathcal{L}^C, \\ &\quad \rho^M(m) = \sum m_{\underline{0}} \otimes m_{\underline{1}} \}. \\ (\mu \otimes \nu)(x) &:= \vee \{ \wedge_{i \in I} (\mu(y) \wedge \nu(z)) \mid y \in M, z \in N, y \otimes z = x \}.\end{aligned}$$

Then we define the notion of $\mathcal{L}$ -subcomodules.

Definition 5.2. An $\mathcal{L}$ -subcomodule of M is an $\mathcal{L}$ -submodule $\mu \in \mathcal{L}^M$ such that for all $m \in M$ hold

$$\mu^M(m) \geq \mu_1(m_{\underline{0}}) \wedge \mu_2(m_{\underline{1}}),$$

where $\rho^M(m) = \sum m_{\underline{0}} \otimes m_{\underline{1}}$.

We denote the set of all $\mathcal{L}$ -subcomodules of M by $\mathcal{L}_C(M)$. It is known that a coalgebra C is also a C -comodule, and the Δ is its comultiplication or coaction. Therefore, for any coalgebra C and for all $c \in C$ we have as a special case of Definition 5.1. :

$$\mu^C(c) := \vee \{ \mu_1(c_{\underline{1}}) \wedge \mu_2(c_{\underline{2}}) \mid \mu_1, \mu_2 \in \mathcal{L}^C, \Delta(c) = \sum c_{\underline{1}} \otimes c_{\underline{2}} \}.$$

An $\mathcal{L}$ -subcomodule of C is then an $\mathcal{L}$ -submodule $\mu \in \mathcal{L}^C$ such that for all $c \in C$ hold

$$\mu^C(c) \geq \mu_1(c_{\underline{1}}) \wedge \mu_2(c_{\underline{2}}),$$

where $\Delta(c) = \sum c_{\underline{1}} \otimes c_{\underline{2}}$.

Lemma 5.3. Let $M = \bigoplus_{i \in I} M_i$, M_i a comodule and $\mu_i \in \mathcal{L}_C(M_i)$ for all $i \in I$. Then $\bigoplus_{i \in I} \mu_i \in \mathcal{L}_C(M)$, where for all $(m_i)_{i \in I} \in M$ holds

$$\left(\bigoplus_{i \in I} \mu_i \right)^M (m_i)_{i \in I} = \wedge_{i \in I} \mu_i^M (m_i).$$

Proof. Following Definition 5.1., we prove in similar way. Let $\mu_i \in \mathcal{L}(M_i)$ for all $i \in I$. An $\mathcal{L}$ -subset $\mu \in \mathcal{L}^M$ given by for any $m = (m_i)_{i \in I}$, where $m_i \neq 0$ for finitely many $i \in I$. The coproduct is denoted by

$$\mu^M = (\bigoplus_{i \in I} \mu_i)^M := \wedge_{i \in I} \mu_i^M,$$

where for all $(m_i)_{i \in I} \in M$ holds

$$(\bigoplus_{i \in I} \mu_i)^M(m_i)_{i \in I} = \wedge_{i \in I} \mu_i^M(m_i)$$

as needed.

For any $\mu, \nu \in \mathcal{L}^M$, it is known that the addition is again belong to $\mathcal{L}^M$ ([7], Theorem 4.1.13). As a direct corollary of Lemma 5.3. we obtain

Corollary 5.4. *Let $\mu, \nu \in \mathcal{L}_C(M)$. Then $\mu + \nu \in \mathcal{L}_C(M)$.*

Lemma 5.5. *Let $\mu \in \mathcal{L}_C(M)$ and $\nu \in \mathcal{L}_C(C)$. By defining*

$$(\mu \otimes \nu)^{M \otimes_R C} := \mu^M \otimes \nu^C$$

it implies $\mu \otimes \nu \in \mathcal{L}_C(M \otimes_R C)$.

Proof. For any $x \in M \otimes_R C$ we obtain :

$$\begin{aligned} (\mu \otimes \nu)^{M \otimes_R C}(x) &= (\mu^M \otimes \nu^C)(x) \\ &= \vee \{ \mu^M(y) \wedge \nu^C(z) \mid y \in M, z \in C, y \otimes z = x \} \\ &\geq \{ \mu_1(y_0) \wedge \mu_2(y_1) \} \wedge \{ \nu_1(z_0) \wedge \nu_2(z_1) \} \\ &= \mu_1(y_0) \wedge \nu_1(z_0) \wedge \mu_2(y_1) \wedge \nu_2(z_1) \\ &= (\mu_1^M \otimes \nu_1^C)(x_0) \wedge (\mu_2^M \otimes \nu_2^C)(x_1) \end{aligned}$$

We also can determine M as a left C^* -module, so by Definition 2.1. we obtain the following. Let $\mu, \nu \in \mathcal{L}^M$. For any $m \in M$ and $f \in C^*$ define

$$(\mu + \nu)(m) := \vee \{ \mu(n) + \nu(t) \mid y, z \in N, n + t = m \} \quad (4)$$

$$(-\mu)(m) := \mu(-m) \quad (5)$$

$$(f\mu)(m) := \vee \{ \mu(n) \mid n \in M, f \rightharpoonup n = m \}. \quad (6)$$

An $\mathcal{L}$ -submodule of M as a left C^* -module is an $\mathcal{L}$ -subset $\mu \in \mathcal{L}^M$ such that

(a) $\mu(0_M) = 1$.

(b) $\mu(f \rightharpoonup m) \geq \mu(m)$ for all $f \in C^*$ and $m \in M$.

(c) $\mu(m + n) \geq \mu(m) \wedge \mu(n)$ for all $m, n \in M$.

We denote the set of all $\mathcal{L}$ -submodules of M as a left C^* -module by $\mathcal{L}_{C^*}(M)$, and the notation $\mathcal{L}_{C^*}$ -submodule means an $\mathcal{L}$ -submodule of M by considering it as a left C^* -module. Since any right C -comodule is a left C^* -module, we conclude that any $\mathcal{L}$ -subcomodule is an $\mathcal{L}_{C^*}$ -submodule but the converse is not always true.

Moreover we adopt the method presented in the previous sections to define prime $\mathcal{L}$ -subcomodules. Throughout, the right comodule M is seen as a left C^* -module.

Definition 5.6. Let $\mu \in \mathcal{L}(M)$. Then μ is called a prime $\mathcal{L}$ -subcomodule if for any nonzero C^* -submodule of M/μ has the same $\mathcal{L}_{C^*}$ -annihilator, that is $\mathcal{L}_{C^*} - \text{Ann}(M/\mu)$.

In other words, $\mu \in \mathcal{L}(M)$ is a prime $\mathcal{L}$ -subcomodule if it is prime $\mathcal{L}_{C^*}$ -module. The subset $K \subseteq \mathcal{L}_{C^*}(M)$ is a prime set if all the elements of K are prime $\mathcal{L}_{C^*}$ -submodule of M . The quotient $\mathcal{L}_{C^*}$ -submodule M/K is called prime if K is a prime set of $\mathcal{L}_{C^*}(M)$. As a special case, $\mu \in \mathcal{L}(C)$ is a prime $\mathcal{L}$ -subcoalgebra if it is prime $\mathcal{L}_{C^*}$ -module.

From Proposition 3.2. we obtain the following

Proposition 5.7. For a comodule M and $\mu \in \mathcal{L}_{C^*}(M)$ the following are equivalent:

- (a) M/μ is a prime C^* -module.
- (b) $\mathcal{L}_{C^*} - \text{Ann}(K/\mu) = \mathcal{L}_{C^*} - \text{Ann}(M/\mu)$ for any C^* -submodule K/μ of M/μ .

Notice that M/μ is not necessary a comodule, hence in Proposition 5.7. we can not say that the M/μ is a prime $\mathcal{L}$ -comodule.

Proposition 5.8. Let C be a coalgebra and M a right C -comodule. If for any $\mathcal{L}_{C^*}$ -submodule ν of M , $\mathcal{L}_{C^*} - \text{Ann}(\nu) \subseteq \mathcal{L}_{C^*} - \text{Ann}(\mu)$ or $\mathcal{L}_{C^*} - \text{Ann}(M/\nu) \subseteq \mathcal{L}_{C^*} - \text{Ann}(\mu)$, then $C^*/\mathcal{L}_{C^*} - \text{Ann}(\mu)$ is a prime algebra.

Proof. Apply Proposition 3.3. by replace the ring R by the dual algebra C^* .

Corollary 5.9. Let M be a right C -comodule, and $\mu \in \mathcal{LI}(C^*)$. If $\mathcal{L} - \text{Ann}(\mu)$ is singleton, say $\{\nu\}$, and for any $\xi \in \mathcal{LI}(C^*)$, $\mathcal{L} - \text{Ann}(\xi) = \{\nu\}$ or $\mathcal{L} - \text{Ann}(C^*/\xi) = \{\nu\}$, then C^*/ν is a prime ring.

Proof. Apply Corollary 3.3..

6. CONCLUDING REMARKS

This work is a very beginning study of prime $\mathcal{L}$ -submodules and prime $\mathcal{L}$ -subcomodules which is defined by an $\mathcal{L}$ -annihilator. Of course it is needed to validate the definition, whether it is a generalization of the previous definition of fuzzy prime ideals or fuzzy prime submodules. We still concern with this work by continuing the investigation to the primeness of any $\mathcal{L}$ -ideal of R by viewing as an $\mathcal{L}$ -submodule of R as well. The relationship between the primeness of an $\mathcal{L}$ -submodule and the primeness of a submodule of N is also the case.

REFERENCES

1. E. ABE, *Hopf Algebras*, Cambridge University Press, 1980.
2. U. ACAR, 2005, On $\mathcal{L}$ -Fuzzy Prime Submodules, *Hacetatepe Journal of Mathematics and Statistics*, **34**, 17–25, 2005.
3. L. BICAN, P. JAMBOR, T. KEPKA, P. NEMEC, Prime and Coprime Modules, *Fundamenta Mathematicae*, **107**, 33–44, 1980.
4. T. BRZEZIŃSKI, R. WISBAUER, *Corings and Comodules*, Cambridge University Press, 2003.
5. S. DĂSCĂLESCU, C. NĂSTĂSESCU, Ș. RAIANU, *Hopf Algebras : An Introduction*, FACULTY OF MATHEMATICS, UNIVERSITY OF BUCHAREST, ROMANIA, 2000.
6. W.B.V. KANDASAMY, *Smarandache Fuzzy Algebra*, Dept. of Mathematics, Indian Institut of Technology, Madras, 2003.
7. J.N MORDESON, D.S. MALIK, *Fuzzy Commutative Algebra*, World Scientific, Singapore, 1998.
8. M.E. SWEEDLER, *Hopf Algebra*, Benjamin, New York, 1969.
9. I.E. WIJAYANTI, *Coprime Modules and Comodules*, Dissertation, University of Düsseldorf, Germany, 2006.
10. I.E. WIJAYANTI, $\mathcal{L}$ -Subcomodules, to be appeared in *JPMS*, UNY, 2007.
11. I.E. WIJAYANTI, Definition of Prime $\mathcal{L}$ -Submodules, submitted in *JMS*, ITB, 2007.
12. R. WISBAUER, On Prime Modules and Rings, *Comm. Algebra*, **11(20)**, 2249–2265, 1983.
13. R. WISBAUER, *Grundlagen der Modul- und Ringtheorie*, Verlag Reinhard Fischer, München, 1988.
14. R. WISBAUER, *Modules and Algebras : Bimodule Structure and Group Actions on Algebras*, Addison Wesley Longman Limited, England, 1996.

INDAH E. WIJAYANTI: Department of Mathematics, Gadjah Mada University, Yogyakarta, Indonesia.

E-mail: indah.wijayanti@yahoo.com

PART II

Analysis

ON CERTAIN CLASSES OF CLOSE-TO-CONVEX FUNCTIONS

K. AL-SHAQSI AND MASLINA DARUS

Abstract. In this paper we define a class of close-to-convex functions by using the generalized of Ruscheweyh derivatives operator D_λ^n were introduced by authors and defined by $D_\lambda^n f(z) = D_\lambda \left(\frac{z(z^{n-1}f(z))^n}{n!} \right)$, where $D_\lambda f(z) = (1 - \lambda)f(z) + \lambda z f'(z)$, $n \in \mathbb{N}_0$ and $\lambda \geq 0$. We establish containment property, integral transform and coefficient inequalities for this class.

Key words and Phrases: Derivatives operator, Convex functions, Starlike functions, Close-to-convex functions, Subordinate.

1. INTRODUCTION

Let $\mathcal{A}$ denote the class of functions of the form:

$$f(z) = z + \sum_{k=2}^{\infty} a_k z^k, \quad (1)$$

which are analytic in the unite disk $\mathbb{U} = \{z : |z| < 1\}$.

Also let $\mathcal{S}$ denotes the subclass of $\mathcal{A}$ consisting of all functions which are univalent in $\mathbb{U}$. Finally, let P denote the class of functions of the form $p(z) = 1 + p_1 z + \cdots$ analytic in $\mathbb{U}$ which satisfy the condition $\Re\{P(z)\} > 0$.

The authors introduced following differential operator (see [2]):

$$\begin{aligned} D_\lambda^0 f(z) &= (1 - \lambda)f(z) + \lambda z f'(z) = D_\lambda f(z), \quad \lambda \geq 0, \\ D_\lambda^1 f(z) &= (1 - \lambda)z f'(z) + \lambda z (z f'(z))', \\ D_\lambda^n f(z) &= D_\lambda \left(\frac{z(z^{n-1} f(z))^n}{n!} \right), \quad (n \in \mathbb{N}_0 = \mathbb{N} \cup \{0\}). \end{aligned} \quad (2)$$

If f is given by (1), then we see that

$$D_\lambda^n f(z) = z + \sum_{k=2}^{\infty} [1 + \lambda(k-1)] C(n, k) a_k z^k,$$

where

$$C(n, k) = \binom{k+n-1}{n} = \frac{\prod_{j=1}^{k-1} (j+n)}{(k-1)!}, \quad k \geq 2.$$

Let S^* and C denote the subclasses of $\mathcal{S}$ consisting of starlike and convex functions, respectively.

A functions $f \in \mathcal{S}$ belongs to the class K of close-to-convex if and only if for some $g \in S^*$,

$$\Re \frac{z f'(z)}{g(z)} > 0, \quad z \in \mathbb{U}.$$

Definition 1.1. [2] Let $M_\lambda^n(\phi)$ be the class of functions $f \in \mathcal{A}$ for which

$$\frac{z(D_\lambda^n f(z))'}{D_\lambda^n f(z)} \prec \phi(z), \quad \phi(z) \in P, \quad z \in \mathbb{U},$$

where $\prec$ denotes the subordination between analytic functions (see [1]).

In [2] the authors showed that $M_\lambda^{n+1}(\phi) \subset M_\lambda^n(\phi)$ for each $n \in \mathbb{N}_0$ and for $\lambda \geq 0$. It is clear that $M_0^0(\phi) \equiv S^*(\phi)$ was introduced and studied by Ma and Minda [16]. When $\phi(z) = (1 + Az)/(1 + Bz)$, $(-1 \leq B < A \leq 1)$, we denote the subclasses $S^*(\phi)$ by $S^*[A, B]$ (see [14] and [8]). For $0 \leq \alpha < 1$, let $S^*(\alpha) := S^*[1 - 2\alpha, -1]$ and $C(\alpha) := C[1 - 2\alpha, -1]$. Finally, $M_0^n((1+z)/(1-z)) \equiv R_n$ was studied by Singh and singh [9] (see also [10]), when $\lambda = 0$ and $\phi(z) = (1 + (1 - 2\alpha)z)/(1 - z)$ we get the class $R_n(\alpha)$ which was studied by Ahuja [6].

Definition 1.2. A function $f \in \mathcal{A}$ belongs to the class $K_\lambda^n(\phi)$ if and only if there exists $g \in M_\lambda^n(\phi)$ such that

$$\frac{z(D_\lambda^n f(z))'}{D_\lambda^n g(z)} \prec \phi(z) \quad \phi(z) \in P, \quad z \in \mathbb{U},$$

Note that the class $K_0^0((1+z)/(1-z)) \equiv K$ is the class of close-to-convex functions (see [15]), the class $K_0^1((1+z)/(1-z)) \equiv QC$ the class of quasi-convex functions introduced in [3], also Noor [4] studied the class $K_0^n((1+z)/(1-z)) \equiv K_n$. Finally, we note that the class $K_0^n(\Psi) \equiv K_n(A, B, \alpha)$ where $\Psi = [1 + (B + (A - B)(1 - \alpha))z]/(1 + Bz)$ was introduced and studied by Kwon and Cho [7]. In this paper we shall establish the containment property $K_\lambda^{n+1}(\phi) \subset K_\lambda^n(\phi)$. We also study integral transforms and some application for $K_\lambda^n(\phi)$.

To derive our theorems, we need the following lemma.

Lemma 1.3. [5] *Let β, γ be complex numbers. Let $h \in \mathcal{A}$ be convex univalent in $\mathbb{U}$ with $h(0) = 1$ and $\Re[\beta h(z) + \gamma] > 0$, $z \in \mathbb{U}$ and $q \in \mathcal{A}$, with $q(z) \prec h(z)$, $z \in \mathbb{U}$. If $p \in P$ is analytic in $\mathbb{U}$ then*

$$p(z) + \frac{zp'(z)}{\beta q(z) + \gamma} \prec h(z) \Rightarrow p(z) \prec h(z).$$

Lemma 1.4. [12] *Let ψ be convex and g be starlike in $\mathbb{U}$. Then for F analytic in $\mathbb{U}$ with $F(0) = 1$, $\frac{\psi * F_g}{\psi * g}$ is contained in the convex hull of $F(\mathbb{U})$.*

Lemma 1.5. [13] *If $p(z) = 1 + c_1z + c_2z^2 + \dots$ is a function with positive real part, then for any complex number μ , $|c_2 - \mu c_1^2| \leq 2\max\{1, |2\mu - 1|\}$ and the result is sharp for the functions given by*

$$p(z) = \frac{1+z^2}{1-z^2}, \quad p(z) = \frac{1+z}{1-z}.$$

2. THE CLASS $K_\lambda^n(\phi)$

We first prove that all functions in $K_\lambda^n(\phi)$ are close-to-convex and hence univalent.

Theorem 2.1. *For $\lambda \geq 0$ and $n \in \mathbb{N}_0$, then*

$$K_\lambda^{n+1}(\phi) \subset K_\lambda^n(\phi)$$

Proof. Let $f \in K_\lambda^{n+1}(\phi)$. Then

$$\frac{z(D_\lambda^{n+1}f(z))'}{D_\lambda^{n+1}g(z)} \prec \phi(z).$$

Set

$$\frac{z(D_\lambda^n f(z))'}{D_\lambda^n g(z)} = p(z) \quad (3)$$

one can easily verify the identity

$$z(D_\lambda^n f(z))' = (n+1)D_\lambda^{n+1}f(z) - nD_\lambda^n f(z). \quad (4)$$

By using (3) in (4) and then differentiating, we get

$$(n+1)z(D_\lambda^{n+1}f(z))' = p(z)z(D_\lambda^n g(z))' + zp'(z)D_\lambda^n g(z) + nz(D_\lambda^n f(z))' \quad (5)$$

Now apply (4) for the function g , and use (5) to obtain

$$\begin{aligned} (n+1)z(D_\lambda^{n+1}f(z))' &= p(z)\left[(n+1)D_\lambda^{n+1}g(z) - nD_\lambda^n g(z)\right] \\ &\quad + zp'(z)D_\lambda^n g(z) + np(z)D_\lambda^n g(z), \end{aligned}$$

or

$$\frac{z(D_\lambda^{n+1}f(z))'}{D_\lambda^{n+1}g(z)} = p(z) + \frac{zp'(z)}{(n+1)q(z)}$$

where $q(z) = \frac{D_\lambda^{n+1}g(z)}{D_\lambda^n g(z)}$. Hence by applying Lemma 1.3 we obtain the required result.

Remark 2.2. Let $f \in K_\lambda^n(\phi)$, f given by (1) and $D_\lambda^n f(z)$ given by (2). Thus by Definition 1.2 it follows that $f \in K_\lambda^n(\phi)$ if and only if $D_\lambda^n f(z) \in K(\phi)$.

Theorem 2.3. If ψ convex and $f \in K_\lambda^n(\phi)$, then $\psi * f \in K_\lambda^n(\phi)$.

Proof. For $f \in K_\lambda^n(\phi)$, there exist $g \in M_\lambda^n(\phi)$ such that $\Re \frac{z(D_\lambda^n f(z))'}{D_\lambda^n g(z)} > 0$. Since $g \in M_\lambda^n(\phi)$ and $\psi \in C$ it follows from [12] that $\psi * g \in M_\lambda^n(\phi)$.

Now

$$\frac{z[D_\lambda^n(\psi * f(z))']}{D_\lambda^n(\psi * g(z))} = \frac{z[\psi(z) * D_\lambda^n f(z)]'}{\psi(z) * D_\lambda^n g(z)} = \frac{\psi(z) * \frac{z(D_\lambda^n f(z))'}{D_\lambda^n g(z)} D_\lambda^n g(z)}{\psi(z) * D_\lambda^n g(z)}$$

Applying Lemma 1.4 with $F(z) = \frac{z(D_\lambda^n f(z))'}{D_\lambda^n g(z)}$ and $D_\lambda^n g(z) \in S^*(\phi)$, we have

$$\Re \frac{z[D_\lambda^n(\psi * f(z))']}{D_\lambda^n(\psi * g(z))} > 0, \quad (z \in \mathbb{U}).$$

3. INTEGRAL OPERATOR

Theorem 3.1. *Let the function $f \in K_\lambda^n(\phi)$ and let $\Re \gamma > 0$, then the function F defined by*

$$F(z) = \frac{\gamma + 1}{z^\gamma} \int_0^z t^{\gamma-1} f(t) dt, \quad (6)$$

belongs to the class $K_\lambda^n(\phi)$.

Proof. Observe that $F(z) = (h * f)(z)$ where $h(z) = \sum_{k=1}^{\infty} \frac{\gamma+1}{\gamma+k} z^k$, $\Re \gamma > 0$. Since h was shown to be convex by Ruscheweyh [11], the result follows immediately from Theorem 2.3.

Theorem 3.2. *Let the function $f \in K_\lambda^n(\phi)$ and let c be real number such that $c > -1$, then the function F defined by*

$$F(z) = \frac{c+1}{z^c} \int_0^z t^{c-1} f(t) dt, \quad (7)$$

belongs to the class $K_\lambda^{n+1}(\phi)$.

Proof. Let

$$G(z) = \frac{c+1}{z^c} \int_0^z t^{c-1} g(t) dt,$$

where $g \in M_\lambda^n(\phi)$. Then from [2], $G \in M_\lambda^{n+1}(\phi)$. From the representation of $F(z)$ and $G(z)$, it follows that

$$z(D_\lambda^n F(z))' = (c+1)D_\lambda^n f(z) - cD_\lambda^n F(z).$$

and

$$z(D_\lambda^n G(z))' = (c+1)D_\lambda^n g(z) - cD_\lambda^n G(z).$$

In [6] it is shown that $D^n f(z) = D^{n+1} F(z)$, and $D^n g(z) = D^{n+1} G(z)$. By using the same technique as in the proof of Theorem 2.1, we get

$$\frac{z(D_\lambda^{n+1} F(z))'}{D_\lambda^{n+1} G(z)} = p(z) + \frac{zp'(z)}{(c+1)q(z)},$$

where $q(z) = \frac{D_\lambda^{n+1} G(z)}{D_\lambda^n G(z)}$. By applying Lemma 1.3 we obtain the required result.

4. COEFFICIENTS INEQUALITIES

Lemma 4.1. *Let $\phi(z) = 1 + B_1z + B_2z^2 + \dots$. If $g(z) = z + \sum_{k=2}^{\infty} b_kz^k$ belongs to $M_{\lambda}^n(\phi)$, then for any complex number μ ,*

$$|b_2| \leq \frac{B_1}{(n+1)(1+\lambda)},$$

and

$$|b_3 - \mu b_2^2| \leq \frac{B_1}{(n+2)(n+1)(1+2\lambda)} \max \left\{ 1, \left| \frac{B_1}{B_2} + \left[\frac{(n+1)(1+\lambda)^2 - (n+2)(1+2\lambda)\mu}{(n+1)(1+\lambda)^2} \right] B_1 \right| \right\}.$$

The result is sharp.

Proof. If $g(z) \in M_{\lambda}^n(\phi)$, then there is a Schwarz function $w(z)$, analytic in $\mathbb{U}$ with $w(0) = 0$ and $|w(z)| < 1$ in $\mathbb{U}$ such that

$$\frac{z(D_{\lambda}^n g(z))'}{D_{\lambda}^n g(z)} = \phi(w(z)).$$

Define the function $p_1(z)$ by

$$p_1(z) = \frac{1 + w(z)}{1 - w(z)} = 1 + c_1z + c_2z^2 + \dots$$

Since $w(z)$ is a Schwarz function, we see that $\Re p_1(z) > 0$ and $p_1(0) = 1$. Define the function $p(z)$ by

$$p(z) = \frac{z(D_{\lambda}^n g(z))'}{D_{\lambda}^n g(z)} = 1 + \gamma_1z + \gamma_2z^2 + \dots,$$

we obtain

$$(n+1)(1+\lambda)b_2 = \gamma_1 \text{ and } (n+2)(n+1)(1+2\lambda)b_3 = (n+1)^2(1+\lambda)^2b_2^2 + \gamma_2.$$

Also we have

$$p(z) = \phi\left(\frac{p_1(z) - 1}{p_1(z) + 1}\right),$$

and from this equation, we obtain

$$\gamma_1 = \frac{1}{2}B_1c_1 \text{ and } \gamma_2 = \frac{1}{2}B_1(c_2 - \frac{1}{2}c_1^2) + \frac{1}{4}B_2c_1^2.$$

Therefore we have

$$b_3 - \mu b_2^2 = \frac{B_1}{2(n+2)(n+1)(1+2\lambda)}[c_2 - vc_1^2],$$

where

$$v = \frac{1}{2} \left(1 - \frac{B_2}{B_1} + \frac{(n+2)(1+2\lambda)\mu - (n+1)(1+\lambda)^2}{(n+1)(1+\lambda)^2} B_1 \right).$$

Our result now follows by an application of Lemma 1.5. The result is sharp for the function defined by

$$\frac{z(D_\lambda^n g(z))'}{D_\lambda^n g(z)} = \phi(z^2),$$

and

$$\frac{z(D_\lambda^n g(z))'}{D_\lambda^n g(z)} = \phi(z).$$

Theorem 4.2. *Let $\phi(z) = 1 + B_1 z + B_2 z^2 + \dots$. If $f(z)$ given by (1) belongs to $K_\lambda^n(\phi)$, then*

$$|a_2| \leq \frac{B_1 + 1}{2(n+1)(1+\lambda)},$$

and

$$|a_3| \leq \frac{2B_1}{3(n+2)(n+1)(1+2\lambda)} \max \left\{ 1, \left| \frac{B_1}{B_2} + \frac{(n+1)}{(1+\lambda)} B_1 \right| \right\} + \frac{2(B_1 + 1)}{3(n+2)(n+1)(1+2\lambda)}.$$

.

Proof. There is exists $g(z) = z + \sum_{k=2}^{\infty} b_k z^k \in M_\lambda^n(\phi)$, and Schwarz function $w(z)$, such that

$$\frac{z(D_\lambda^n f(z))'}{D_\lambda^n g(z)} = \frac{1+w(z)}{1-w(z)} = 1 + \gamma_1 z + \gamma_2 z^2 + \dots$$

Comparing series expansions, we see

$$a_2 = \frac{b_2}{2} + \frac{\gamma_2}{2(n+1)(1+\lambda)},$$

and

$$a_3 = \frac{2b_3}{3} + \frac{2(1+\lambda)b_2\gamma_1}{3(n+2)(1+2\lambda)} + \frac{2\gamma_2}{3(n+2)(n+1)(1+2\lambda)}$$

By applying Lemma 1.5 and Lemma 4.1 ($\mu = 0$), we have

$$|a_3| \leq \frac{2B_1}{3(n+2)(n+1)(1+2\lambda)} \max \left\{ 1, \left| \frac{B_1}{B_2} + \frac{(n+1)}{(1+\lambda)} B_1 \right| \right\} + \frac{2(B_1+1)}{3(n+2)(n+1)(1+2\lambda)}.$$

REFERENCES

1. C. POMMERENKE, *Univalent Functions*, Vandenhoeck and Ruprecht, Gttingen, 1975.
2. K. AL-SHAQSI AND M. DARUS, On univalent functions with respect to k -symmetric points defined by a generalized Ruscheweyh derivatives operator.(Submitted)
3. K. I. NOOR, On certain classes of close-to-convex functions, *Int. J. Math. and Math Sci.* **16**(2), 329–336, 1993.
4. K. I. NOOR AND D. K. THOMAS, On quasi-convex univalent functions, *Int. J. Math. and Math Sci.* **3**, 225–266, 1980.
5. K. PADMANABHAN AND PARVATHAM, Some applications of subordinations, *Bull. Austral. Math. Soc.* **32**, 321–330, 1985.
6. O. AHUJA, Integral operators of certain univalent functions, *Internat. J. Math. Sci.* **8**, 653–662, 1985.
7. O. S. KWON AND N. E. CHO, OA subclass of close-to-convex functions, *Soochow J. Math.* **22**(4), 449–456, 1996.
8. R. M. GOEL AND B. S. MEHROK, On a class of close-to-convex functions, *Indian J. Pure Appl. Math.* **12**(5), 684–658, 1981.
9. R. SINGH AND S. SINGH, Integrals of certain univalent functions, *Proce. Amer. Math. Soc.* **77**, 336–340, 1979.
10. S. OWA, S. FUKUI, K. SAKAGUCHI AND S. OGAWA, An application of the Ruscheweyh derivatives, *Internat. J. Math. Sci.* **9**, 721–730, 1986.
11. S. RUSCHEWEYH AND T. SHIEL-SMALL, Hadamard products of schlicht functions and the Polya-Schoenberg conjecture, *Comment. Math. Helv.* **48**, 119–135, 1973.
12. T. R. REDDY, O. P. JUNEJA AND K. SATHYANARAYANA, A convolution approach to certain subclasses of starlike functions, *Tamkang J. of Math.* **23**(4), 311–320, 1992. 311-320.
13. V. RAVICHANDRAN, M. BOLCAL, Y. Polatoğlu and A. Sen, Certian subclasses of Starlike and Convex functions of Complex Order, *Hacettepe J. Math. Stat.* **34**, 9–15, 2005.
14. W. JANOWSKI, Some extremal problems for certain families of analytic functions, *Ann. Polon. Math.* **28**, 297–326, 1973.
15. W. KAPLAN, Close-to-convex schlicht functions, *Mich. Math. J.* **1**, 169–185, 1952.
16. W. MA AND D. MINDA, A unified treatment of some special classes of univalent functions, in: *Proceedings of the conference on complex analysis*, Z. Li, F. Ren, L. Yang, and S. Zhang(Eds.), Int. Press, 157-169, 1994.

K. AL-SHAQSI : School of Mathematical Sciences, Faculty of Science and Technology,
Universiti Kebangsaan Malaysia, Bangi 43600 Selangor D. Ehsan, Malaysia.

E-mail: ommath@hotmail.com

M. DARUS: School of Mathematical Sciences, Faculty of Science and Technology, Univer-
siti Kebangsaan Malaysia, Bangi 43600 Selangor D. Ehsan, Malaysia.

E-mail: maslina@pkrisc.cc.ukm.my

REPRESENTATION OF SM -OPERATORS ON CLASSICAL BANACH SPACES

$$L_p(E, \Sigma, \mu), 1 < p < \infty$$

MUSLIM ANSORI
 SOEPARNA DARMAWIJAYA, SUPAMA

Abstract. This is a continuation of our results in studying generalization of Hilbert-Schmidt and Carleman operators in Banach spaces. The recent results of the generalization of Hilbert-Schmidt operator were called SM -operator. Infact, almost all of properties of the SM -operator preserve almost all of properties of the Hilbert-Schmidt operators (Darmawijaya, Ansori and Supama, 2006). The generalization can be done by preserving some intrinsic properties of Hilbert spaces involved; for examples, reflexivity and separability. In this paper, the SM – operator is defined by leaving aside orthogonality in Banach spaces. The applications on classical Banach spaces $L_p(E, \Sigma, \mu)$, $1 < p < \infty$ are discussed and some examples are given.

Keywords : Schauder bases, biorthogonal system, Separable and reflexive Banach spaces, Hilbert-Schmidt operator

1. INTRODUCTION

One of the most important classes of bounded operators is the class of Hilbert-Schmidt operators. Let H_1 and H_2 be Hilbert spaces. A bounded operator $A: H_1 \rightarrow H_2$ is called a Hilbert-Schmidt operator if there exists an orthonormal bases $\{e_n\}$ of H_1 such that

$$\sum_{n=1}^{\infty} \|Ae_n\|^2 < \infty.$$

This definition implies that $A: H_1 \rightarrow H_2$ is a Hilbert-Schmidt operator if and only if $A^*: H_2 \rightarrow H_1$ is a Hilbert-Schmidt operator ; in this case

$$\sum_{n=1}^{\infty} \|Ae_n\|^2 = \sum_{n=1}^{\infty} \|A^*d_n\|^2,$$

for every orthonormal bases $\{e_n\}$ of H_1 and $\{d_n\}$ of H_2 (Weidmann, 1980 and Conway, 1990). Now, a question arises whether such an operator can be developed in Banach spaces. The answer is positive whenever we preserve some intrinsic properties of the two Hilbert spaces, for examples, reflexivity and separability. The separability of Banach space X is to guarantee the existence of countable bases of a Banach space X and the reflexivity of a Banach space X is to guarantee that the bases of X is shrinking (Zippin, 1968). Further, Johnson., et

al.,1971, pointed out that the existence of bases in the dual X^* does imply that also X has a bases , see also (Dapa,2000; Morrison,2001). More precisely, if a separable and reflexive Banach space X has a shrinking bases, so does the dual space X^* . For example, ℓ_p , $1 < p < \infty$, has a bases (Schauder bases) but ℓ_∞ has not.

Let the Banach spaces X , Y and Z are reflexive and separable. Let X^* be the dual space of X that is the collection of all continuous linear functionals on X . We always write $\langle x, x^* \rangle$ to stand for $x^*(x)$ and vice versa, for every $x \in X$ and $x^* \in X^*$. A sequence of linearly independent vectors $\{e_n\} \subset X$ is called a Schauder bases of X if for every vector $x \in X$ there is uniquely sequence of scalars $\{\alpha_n\}$ such that

$$x = \sum_{k=1}^{\infty} \alpha_k e_k.$$

We define a sequence of vector $\{e_n^*\} \subset X^*$, which is called biorthogonal system of $\{e_n\}$, as follows:

$$\langle x, e_n^* \rangle = e_n^*(x) = e_n^*\left(\sum_{k=1}^{\infty} \alpha_k e_k\right) = \sum_{k=1}^{\infty} \alpha_k e_n^*(e_k) = \alpha_n,$$

for every $n \in N$. It is true that $e_n^* \in X^*$, for e_n^* is linear and bounded, $\langle e_k, e_n^* \rangle = 0$ for every $k \neq n$ and $\langle e_k, e_n^* \rangle = 1$ for every $k = n$. The sequence $\{e_n^*\}$ forms a bases of the closed subspace $\overline{\{e_n^*\}} \subset X^*$. Especially, we have $\overline{\{e_n^*\}} = X^*$ if and only if $\{e_n\}$ is shrinking, i.e.,

$$\lim_{n \rightarrow \infty} \sum_{k=n}^{\infty} \langle e_k, x^* \rangle e_k^* = 0^*,$$

for every $x^* \in X^*$ (Lindenstrauss and Tzafriri, 1996, Proposition 1.b.1).

Let X and Y be separable and reflexive Banach spaces, $\{e_n^*\} \subset X^*$, $\{d_m^*\} \subset Y^*$, be their respectively biorthogonal system of $\{e_n\} \subset X$, $\{d_m\} \subset Y$, respectively. If $A \in L_c(X, Y)$, where $L_c(X, Y)$ is the collection of continuously linear operators from Banach space X into Banach space Y , the operator $A^* \in L_c(Y^*, X^*)$ is called the adjoint operator of A if for any $x \in X$ and $y^* \in Y^*$, we have

$$\langle Ax, y^* \rangle = \langle x, A^* y^* \rangle.$$

Then, we have

$$Ae_n = \sum_{k=1}^{\infty} d_k^*(Ae_n) d_k = \sum_{k=1}^{\infty} \langle Ae_n, d_k^* \rangle d_k = \sum_{k=1}^{\infty} \langle e_n, A^* d_k^* \rangle d_k,$$

It implies

$$\langle Ae_n, d_k^* \rangle = \langle e_n, A^* d_k^* \rangle.$$

By the fact that,

$$\phi(x)(f) = \widehat{x}(f) = f(x)$$

for every $f \in X^*$, then

$$\langle A^* d_k^*, e_n^{**} \rangle = \langle e_n, A^* d_k^* \rangle.$$

Because of X isometrically isomorphic with X^{**} , then we can take $e_n^{**} \equiv e_n$.

Further, Darmawijaya, Ansori and Supama (2006) defined a new operator $A: X \rightarrow Y$ called an *SM-operator* if

$$\sum_{n=1}^{\infty} \sum_{m=1}^{\infty} |\langle Ae_n, d_m^* \rangle| < \infty,$$

for every orthonormal Schauder bases $\{e_n\} \subset X$, $\{d_m\} \subset Y$. In the paper X and Y was especially limited to admit orthogonality. Because of

$$\sum_{n=1}^{\infty} \sum_{m=1}^{\infty} |\langle Ae_n, d_m^* \rangle| = \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} |\langle e_n, A^* d_m^* \rangle| = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} |\langle e_n, A^* d_m^* \rangle|$$

for any reflexive and separable Banach spaces X, Y , then

$$\|A\| = \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} |\langle Ae_n, d_m^* \rangle| = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} |\langle e_n, A^* d_m^* \rangle| = \|A^*\|.$$

Hence, if A is an *SM-operator*, so does A^* and vice versa. The collection of all *SM-operators* from X into Y denoted by $SM(X, Y)$ equipped with the norm $\|\cdot\|$ forms a Banach space.

Let $(X, \|\cdot\|)$ be a separable and reflexive Banach space with a Schauder bases $\{e_n\}$. For every $x = \sum_{n=1}^{\infty} \langle x, e_n^* \rangle e_n$ in X the expression $\|x\|_0 = \sup_k \left\| \sum_{n=1}^k \langle x, e_n^* \rangle e_n \right\|$ is finite. Evidently, $\|\cdot\|_0$ is a norm on X and $\|x\| \leq \|x\|_0$ for every $x \in X$. Further, X is complete with respect to $\|\cdot\|_0$ and thus by open mapping theorem, the norms $\|\cdot\|$ and $\|\cdot\|_0$ are equivalent (Lindenstrauss and Tzafriri, 1996. Now, by leaving aside orthogonality (orthonormality) of Schauder Bases of X and Y , we have

$$\sum_{n=1}^{\infty} \sup_n \left\| \sum_{k=1}^n \langle Ae_n, d_k^* \rangle d_k \right\|_Y \leq \sum_{n=1}^{\infty} \sum_{k=1}^{\infty} \|\langle Ae_n, d_k^* \rangle d_k\|_Y = \|A\|$$

may not be the same with

$$\sum_{n=1}^{\infty} \sup_k \left\| \sum_{n=1}^k \langle A^* d_k^*, e_n \rangle e_n^* \right\|_{X^*} \leq \sum_{n=1}^{\infty} \sum_{k=1}^{\infty} \|\langle A^* d_k^*, e_n \rangle e_n^*\|_{X^*} = \|A^*\|. \text{ But,}$$

they are automatically the same whenever the bases admit orthonormality or X and Y are Hilbert spaces.

In this paper, we want to implement the norm as the variant of *SM-operator* of the operator $A: X \rightarrow Y$, where $X = L_p(E_1, \Omega_1, \mu_1)$ and $Y = L_q(E_2, \Sigma_{2,E_2}, \mu_2)$, $(E_i, \Sigma_{i,E_i}, \mu_i), i=1,2$ are complete σ -finite measure spaces, $1 < p, q < \infty$ and $\frac{1}{p} + \frac{1}{q} = 1$.

2. PRELIMINARY

Let $\Omega_1, \Omega_2 \neq \emptyset$ and $(\Omega_1, \Sigma_1, \mu_1)$ and $(\Omega_2, \Sigma_2, \mu_2)$ be two complete σ -finite measure spaces, and consider the direct product $\Omega_1 \times \Omega_2$ of Ω_1 and Ω_2 . If $E_1 \subset \Omega_1$ and $E_2 \subset \Omega_2$, we call $E_1 \times E_2$ a rectangle. If $E_1 \in \Sigma_1$ and $E_2 \in \Sigma_2$, we call $E_1 \times E_2$ a measurable rectangle. If $E_1 \times E_2$ is a measurable rectangle, we set $\mu(E_1 \times E_2) = \mu_1(E_1)\mu_2(E_2)$. As a consequence, we also have $\mu(E_1 \times E_2) = \sum_{i \in I} \mu(E_{1i} \times E_{2i})$, where $\{(E_{1i} \times E_{2i})\}_{i \in I}$ is a countable disjoint collection of measurable rectangles whose union is a measurable rectangle $E_1 \times E_2$. In this paper, we always assume that $(\Omega_1, \Sigma_1, \mu_1)$ and $(\Omega_2, \Sigma_2, \mu_2)$ complete σ -finite measure spaces. If Σ_1 and Σ_2 are σ -algebra of sets on Ω_1 and Ω_2 , respectively, and $E_1 \in \Sigma_1, E_2 \in \Sigma_2$ then the family of sets $\Sigma_{1,E_1} = \{F_1 \cap E_1 : F_1 \in \Sigma_1\}$ and $\Sigma_{2,E_2} = \{F_2 \cap E_2 : F_2 \in \Sigma_2\}$, are also σ -algebra of sets on E_1 and E_2 , respectively. Hence, we can construct two complete σ -finite measure subspaces $(E_1, \Sigma_{1,E_1}, \mu_1)$ and $(E_2, \Sigma_{2,E_2}, \mu_2)$. For complete references see (Royden, 1988). Let $X = L_p(E_1, \Sigma_{1,E_1}, \mu_1), Y = L_q(E_2, \Sigma_{2,E_2}, \mu_2)$ where complete σ -finite measure spaces. Let $E_1 \in \Sigma_1$ and $E_2 \in \Sigma_2$, respectively. Then $E_2 \times E_1 \in \Sigma_2 \times \Sigma_1$. The points of $E_2 \times E_1$ can be written in the form $(\bar{x}, \bar{y})$ with $\bar{x} \in E_2, \bar{y} \in E_1$. Assume $k \in L_q(E_2 \times E_1, \Sigma_{2,E_2} \times \Sigma_{1,E_1}, \mu_2 \times \mu_1)$. We define an operator $K : L_p(E_1, \Sigma_{1,E_1}, \mu_1) \rightarrow L_q(E_2, \Sigma_{2,E_2}, \mu_2)$, $1 < p, q < \infty, \frac{1}{p} + \frac{1}{q} = 1$ by

$$(Kf)(\bar{x}) = \int_{E_1} k(\bar{x}, \bar{y}) f(\bar{y}) d\mu_1(\bar{y}) \quad (1.1)$$

for all $\bar{x} \in E_2$. Then we have

$$|(Kf)(\bar{x})| \leq \|f\| \left(\int_A |k(\bar{x}, \bar{y})|^q d\mu_1(\bar{y}) \right)^{\frac{1}{q}}.$$

Let h be the integrable function in $L_q(E_2 \times E_1, \Sigma_{2,E_2} \times \Sigma_{1,E_1}, \mu_2 \times \mu_1)$ defined by

$$h(\bar{x}, \bar{y}) = k(\bar{x}, \bar{y}) f(\bar{y}) g(\bar{x})$$

for every $g \in L_q(E_2, \Sigma_{2,E_2}, \mu_2)$. Therefore by Fubini's Theorem the function

$$g(\bar{x})(Kf)(\bar{x}) = g(\bar{x}) \int_{E_1} K(\bar{x}, \bar{y}) f(\bar{y}) d\mu_1(\bar{y}) = \int_{E_1} h(\bar{x}, \bar{y}) d\mu_1(\bar{y})$$

is a measurable function on E_2 . If we put $E_{2,n} = \{\bar{x} \in E_2 : \|\bar{x}\| \leq n\}$ and $g_n = \chi_{E_{2,n}} Kf$ is measurable for every $n \in \mathbb{N}$. Consequently, Kf is measurable.

Because of $Kf \in L_q(E_2, \Sigma_{2,E_2}, \mu_2)$, then we have

$$\|Kf\| = \left(\int_{E_2} |(Kf)(\bar{x})|^q d\mu_2(\bar{x}) \right)^{\frac{1}{q}} \leq \|f\| \left(\int_{E_2} \int_{E_1} |k(\bar{x}, \bar{y})|^q d\mu_1(\bar{y}) d\mu_2(\bar{x}) \right)^{\frac{1}{q}} < \infty.$$

Since the mapping $f \rightarrow Kf$ is linear, we have defined linear continuous operator from $L_p(E_1, \Sigma_{1,E_1}, \mu_1)$ into $L_q(E_2, \Sigma_{2,E_2}, \mu_2)$, $1 < p, q < \infty, \frac{1}{p} + \frac{1}{q} = 1$. The operator K is induced by the function $k \in L_q(E_2 \times E_1, \Sigma_{2,E_2} \times \Sigma_{1,E_1}, \mu_2 \times \mu_1)$.

For simplicity, we always write $X = L_p(E_1, \Sigma_{1,E_1}, \mu_1)$ as $L_p(E_1)$ and

$Y = L_q(E_2, \Sigma_{2,E_2}, \mu_2)$ as $L_q(E_2)$. The operator $K \in L_c(L_p(E_1), L_q(E_2))$ is in the sense of (1.1).

3. MAIN RESULTS

Based on the previous discussion we start with the following definition.

Definition 1. An operator $K \in L_c(L_p(E_1), L_q(E_2))$ is called an *SM – operator* from $L_p(E_1)$ into $L_q(E_2)$, if

$$\sum_{n=1}^{\infty} \sum_{k=1}^{\infty} \|\langle Ke_n, d_k^* \rangle d_k\|_Y < \infty.$$

for every Schauder bases $\{e_n\}$ of $L_p(E_1)$ and $\{d_k\}$ of $L_q(E_2)$.

It is clear that if A is an *SM – operator*, then the number $\|K\|$:

$$\|K\| = \sum_{n=1}^{\infty} \sum_{k=1}^{\infty} \|\langle Ke_n, d_k^* \rangle d_k\|_Y$$

is nonnegative and it does not depend on the choice of an Schauder bases $\{e_n\}$ of $L_p(E_1)$ and an Schauder bases $\{d_n\}$ of $L_q(E_2)$. Let $SM(L_p(E_1), L_q(E_2))$ be the collection of *SM – operators* from a Banach space $L_p(E_1)$ into a Banach space $L_q(E_2)$.

Theorem 2. Given $L_p(E_1)$, $L_q(E_2)$. Let $\{e_n\}$ and $\{d_n\}$ be an Schauder bases of Banach space $L_p(E_1)$ and $L_q(E_2)$ respectively. Then,

- (i) $\|K\| \leq \|K\|$, for every, $K \in SM(L_p(E_1), L_q(E_2))$
- (ii) $SM(L_p(E_1), L_q(E_2))$ is a Banach space with respect to $\|\cdot\|$,
- (iii) If $K \in SM(L_p(E_1), L_q(E_2))$, then K is compact.

Proof: (i) For every $f \in L_p(E_1)$, we have

$$\begin{aligned} \|Kf\|_Y &= \left\| \sum_{n=1}^{\infty} \langle f, e_n^* \rangle Ke_n \right\|_Y \\ &\leq \sum_{n=1}^{\infty} |\langle f, e_n^* \rangle| \|Ke_n\|_Y \\ &= \sum_{n=1}^{\infty} |\langle f, e_n^* \rangle| \left\| \sum_{j=1}^{\infty} \langle Ke_n, d_j^* \rangle d_j \right\|_Y \\ &\leq \|f\| \sum_{n=1}^{\infty} \sum_{j=1}^{\infty} \|\langle Ke_n, d_j^* \rangle d_j\|_Y \\ &\leq \|f\| \|K\| \end{aligned}$$

which implies $\|K\| \leq \|K\|$. ■

(ii). The space $SM(L_p(E_1), L_q(E_2))$ is a normed space with respect to the norm $\|\cdot\|$, for:

$$(ii.a). \left\| K \right\| = \sum_{n=1}^{\infty} \sum_{k=1}^{\infty} \left\| \langle K e_n, d_k^* \rangle d_k \right\|_Y \geq 0, \text{ for}$$

$$\text{every } K \in SM(L_p(E_1), L_q(E_2)),$$

$$\left\| K \right\| = 0 \Leftrightarrow \sum_{k=1}^{\infty} \left\| \langle K e_n, d_k^* \rangle d_k \right\|_Y = 0 \iff K = O \text{ (null operator),}$$

$$(ii.b). \left\| \alpha K \right\| = \sum_{n=1}^{\infty} \sum_{k=1}^{\infty} \left\| \langle \alpha K e_n, d_k^* \rangle d_k \right\|_Y = |\alpha| \sum_{n=1}^{\infty} \sum_{k=1}^{\infty} \left\| \langle K e_n, d_k^* \rangle d_k \right\|_Y$$

$$= |\alpha| \left\| K \right\|, \text{ for every scalar } \alpha \text{ and } K \in SM(L_p(E_1), L_q(E_2)),$$

and

$$(ii.c). \text{ For every } K_1, K_2 \in SM(L_p(E_1), L_q(E_2)), \text{ we have}$$

$$\begin{aligned} \left\| K_1 + K_2 \right\| &= \sum_{n=1}^{\infty} \sum_{k=1}^{\infty} \left\| \langle (K_1 + K_2) e_n, d_k^* \rangle d_k \right\|_Y \\ &\leq \sum_{n=1}^{\infty} \sum_{k=1}^{\infty} \left\| \langle K_1 e_n, d_k^* \rangle d_k + \langle K_2 e_n, d_k^* \rangle d_k \right\|_Y \\ &= \sum_{n=1}^{\infty} \sum_{k=1}^{\infty} \left(\left\| \langle K_1 e_n, d_k^* \rangle d_k \right\|_Y + \left\| \langle K_2 e_n, d_k^* \rangle d_k \right\|_Y \right) \\ &= \sum_{n=1}^{\infty} \sum_{k=1}^{\infty} \left\| \langle K_1 e_n, d_k^* \rangle d_k \right\|_Y + \sum_{n=1}^{\infty} \sum_{k=1}^{\infty} \left\| \langle K_2 e_n, d_k^* \rangle d_k \right\|_Y \\ &= \left\| K_1 \right\| + \left\| K_2 \right\| \end{aligned}$$

or

$$\left\| K_1 + K_2 \right\| \leq \left\| K_1 \right\| + \left\| K_2 \right\|.$$

The proof of the completeness of the space is as follows. Let $\{K_n\} \subset SM(L_p(E_1), L_q(E_2))$ be an arbitrary Cauchy sequence. Then, for any number $\varepsilon > 0$, there is a positive integer n_0 such that for every two positive integers $m, n \geq n_0$, we have

$$\left\| K_m - K_n \right\| < \frac{\varepsilon}{2}.$$

We want to prove that there is $K \in SM(L_p(E_1), L_q(E_2))$ such that $\lim_{n \rightarrow \infty} \left\| K_n - K \right\| = 0$. Since $L_c(L_p(E_1), L_q(E_2))$ is complete and $\left\| K_m - K_n \right\| \leq \left\| K_m - K_n \right\|$, there is $K \in L_c(L_p(E_1), L_q(E_2))$ such that $\left\| K - K_n \right\| < \varepsilon$, for every $n \geq n_0$ or $\lim_{n \rightarrow \infty} \left\| K_n - K \right\| = 0$. Thus, we have

$$\begin{aligned} \sum_{j=1}^s \left\| \sum_{k=1}^t \langle (K_n - K_m) e_j, d_k^* \rangle d_k \right\| &\leq \sum_{j=1}^s \sum_{k=1}^t \left\| \langle (K_n - K_m) e_j, d_k^* \rangle d_k \right\| \\ &\leq \left\| K_n - K_m \right\| < \frac{\varepsilon}{2} \end{aligned}$$

for any integers s, t and $m, n \geq n_0$. By letting $m \rightarrow \infty$, we have

$$\sum_{j=1}^s \left\| \sum_{k=1}^t \langle (K_n - K) e_j, d_k^* \rangle d_k \right\| \leq \sum_{j=1}^s \sum_{k=1}^t \left\| \langle (K_n - K) e_j, d_k^* \rangle d_k \right\| \leq \frac{\varepsilon}{2}$$

for any integers s, t and $n \geq n_0$. Letting $s \rightarrow \infty$ and $t \rightarrow \infty$, we have

$$\sum_{j=1}^{\infty} \left\| \sum_{k=1}^{\infty} \langle (K_n - K)e_j, d_k^* \rangle d_k \right\| \leq \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} \left\| \langle (K_n - K)e_j, d_k^* \rangle d_k \right\|_Y \leq \frac{\varepsilon}{2} < \varepsilon$$

for $n \geq n_0$. Therefore $K_n - K \in SM(L_p(E_1), L_q(E_2))$ and hence $K = K_n + (K - K_n)$ in $SM(L_p(E_1), L_q(E_2))$. Moreover, $\|K_n - K\| < \varepsilon$ for $n \geq n_0$. Hence $\lim_{n \rightarrow \infty} \|K_n - K\| = 0$.

(iii) If $K \in SM(L_p(E_1), L_q(E_2))$ and $f \in L_p(E_1)$, we have

$$Kf = \sum_{k=1}^{\infty} \langle Kf, d_k^* \rangle d_k$$

and for every positive integer n , we define an operator $K_n : L_p(E_1) \rightarrow L_q(E_2)$:

$$K_n f = \sum_{k=1}^n \langle Kf, d_k^* \rangle d_k.$$

It is clear that $K_n \in L_c(L_p(E_1), L_q(E_2))$ and K_n is a finite rank operator, and $\lim_{n \rightarrow \infty} \|K - K_n\| = 0$. Therefore, K is a compact operator. ■

Theorem 3. Let $L_p(E_1), L_q(E_2)$ and $L_p(E_3)$ be separable and reflexive Banach spaces. If $K \in SM(L_p(E_1), L_q(E_2))$ and $M \in L_c(L_q(E_2), L_p(E_3))$, then $MK \in SM(L_p(E_1), L_p(E_3))$ and $\|MK\| \leq \|M\| \|K\|$.

Proof: For every Schauder bases $\{e_n\}$ of $L_p(E_1)$, $\{d_k\}$ of $L_q(E_2)$ and $\{l_j\}$ of $L_p(E_3)$ we have

$$\begin{aligned} \|MK\| &= \sum_{n=1}^{\infty} \sum_{j=1}^{\infty} \left\| \langle MKe_n, l_j^* \rangle l_j \right\| \\ &\leq \sum_{n=1}^{\infty} \sum_{j=1}^{\infty} \|M\| \left\| \langle Ke_n, l_j^* \rangle l_j \right\| \\ &= \|M\| \sum_{n=1}^{\infty} \sum_{j=1}^{\infty} \left\| \langle Ke_n, l_j^* \rangle l_j \right\| \\ &= \|M\| \|K\|, \end{aligned}$$

i.e., $MK \in SM(L_p(E_1), L_p(E_3))$ and $\|MK\| \leq \|M\| \|K\|$. ■

Let $SM(L_p(E_1))$ and $L_c(L_p(E_1))$ stand for $SM(L_p(E_1), L_p(E_1))$ and $L_c(L_p(E_1), L_p(E_1))$, respectively. Combining the results of Theorem 2 and Theorem 3, we have

Theorem 4. Let $L_p(E_1)$ be a Banach space having a shrinking Schauder bases. Then, $SM(L_p(E_1))$ is a Banach $*$ -algebra and an ideal of $L_c(L_p(E_1))$.

We want to give some conditions for the operator K to be an SM -operator with respect to its inducing function or kernel k . Firstly, we shall consider Schauder bases of $L_p(E_2 \times E_1, \Sigma_{2, E_2} \times \Sigma_{1, E_1}, \mu_2 \times \mu_1)$, $1 < p < \infty$, in the following

theorem.

Theorem 5. Let $(E_1, \Sigma_{1,E_1}, \mu_1)$ and $(E_2, \Sigma_{2,E_2}, \mu_2)$ be complete σ -finite measure spaces, respectively. Let $\{\bar{e}_n\}$ and $\{\bar{f}_m\}$ be Schauder bases of $L_q(E_1, \Sigma_{1,E_1}, \mu_1)$ and $L_q(E_2, \Sigma_{2,E_2}, \mu_2)$, $1 < q < \infty$, respectively, where $\{\bar{e}_n^*\}$ and $\{\bar{f}_m^*\}$ be Schauder bases of $L_p(E_1, \Sigma_{1,E_1}, \mu_1)$ and $L_p(E_2, \Sigma_{2,E_2}, \mu_2)$, $1 < p < \infty$, respectively and $\frac{1}{p} + \frac{1}{q} = 1$ in the biorthogonal system. If we defined $g_{mn}^* \in L_p(E_2 \times E_1, \Sigma_{2,E_2} \times \Sigma_{1,E_1}, \mu_2 \times \mu_1)$ by $g_{mn}^*(\bar{x}, \bar{y}) = e_n^*(\bar{y}) f_m^*(\bar{x})$ for $\bar{x} \in \Omega_2, \bar{y} \in \Omega_1$, then $\{g_{mn}^*\}$ is a Schauder bases of $L_p(E_2 \times E_1, \Sigma_{2,E_2} \times \Sigma_{1,E_1}, \mu_2 \times \mu_1)$.

Proof: it is clear that the functions $\{g_{mn}^*\}$ are in $L_p(E_2 \times E_1, \Sigma_{2,E_2} \times \Sigma_{1,E_1}, \mu_2 \times \mu_1)$ and they form a biorthogonal system. It remain to prove that $\{g_{mn}^*\}$ are total. Let h be an element of $L_q(E_2 \times E_1, \Sigma_{2,E_2} \times \Sigma_{1,E_1}, \mu_2 \times \mu_1)$ such that $\langle h, g_{mn}^* \rangle = 0$ for all $(m, n) \in \mathbb{N} \times \mathbb{N}$. Let $h_{\bar{x}}(\bar{y}) = h(\bar{x}, \bar{y}), \bar{x} \in E_2$. By Fubini's Theorem we have $h_{\bar{x}} \in L_q(E_2, \Sigma_{2,E_2}, \mu_2)$, for all $\bar{x} \in E_1$, and we have

$$\begin{aligned} \|h\|^q &= \int_{E_2} \left(\int_{E_1} |h_{\bar{x}}(\bar{y})|^q d\mu_1(\bar{y}) \right) d\mu_2(\bar{x}) \\ &= \int_{E_2} \|h_{\bar{x}}\|^q d\mu_2(\bar{x}) \\ &\leq \int_{E_2} \left(\sum_{n=1}^{\infty} \|\langle h_{\bar{x}}, e_n^* \rangle e_n\|^q \right) d\mu_2(\bar{x}) \\ &= \int_{E_2} \left(\sum_{n=1}^{\infty} \left(\int_{E_1} |\langle h_{\bar{x}}, e_n^* \rangle e_n(\bar{y})|^q d\mu_1(\bar{y}) \right)^{\frac{1}{q}} \right)^q d\mu_2(\bar{x}) \end{aligned} \quad (1.2)$$

Let us defined k_n by

$$k_n(\bar{x}) = \langle h_{\bar{x}}, e_n^* \rangle e_n(\bar{y}) = \int_{E_1} h(\bar{x}, \bar{y}) e_n^*(\bar{y}) d\mu_1(\bar{y}) e_n(\bar{y}).$$

It is easy to prove that $k_n \in L_q(E_2, \Sigma_{2,E_2}, \mu_2)$. It follows that

$$\begin{aligned} \|k_n\| &\leq \sum_{m=1}^{\infty} \|\langle k_n, f_m^* \rangle f_m\| \\ &= \sum_{m=1}^{\infty} \left(\left| \int_{E_2} k_n(\bar{x}) f_m^*(\bar{x}) d\mu_2(\bar{x}) f_m(\bar{x}) \right|^q \right)^{\frac{1}{q}} \\ &= \sum_{m=1}^{\infty} \left(\left| \int_{E_2} \left(\int_{E_1} h(\bar{x}, \bar{y}) e_n^*(\bar{y}) d\mu_1(\bar{y}) e_n(\bar{y}) \right) f_m^*(\bar{x}) d\mu_2(\bar{x}) f_m(\bar{x}) \right|^q \right)^{\frac{1}{q}} \\ &= \sum_{m=1}^{\infty} \left(\left| \int_{E_2} \int_{E_1} h(\bar{x}, \bar{y}) e_n^*(\bar{y}) f_m^*(\bar{x}) d\mu_1(\bar{y}) d\mu_2(\bar{x}) e_n(\bar{y}) f_m(\bar{x}) \right|^q \right)^{\frac{1}{q}} \\ &= \sum_{m=1}^{\infty} \left(\left| \int_{E_2 \times E_1} h(\bar{x}, \bar{y}) g_{mn}^*(\bar{x}, \bar{y}) d\mu_1(\bar{y}) d\mu_2(\bar{x}) g_{mn}(\bar{x}, \bar{y}) \right|^q \right)^{\frac{1}{q}} \\ &= \sum_{m=1}^{\infty} \|\langle h, g_{mn}^* \rangle g_{mn}\| \end{aligned}$$

$$= 0$$

for every $n \in \mathbb{N}$. By summing up in (1.2), it follows that $\|h\| = 0$ or $h = \theta$. ■

Now we are ready to provide necessary and sufficient condition of the operator K to be an SM – operator, as stated in the following theorem.

Theorem 6. Let $(E_1, \Sigma_{1,E_1}, \mu_1)$ and $(E_2, \Sigma_{2,E_2}, \mu_2)$ be complete σ -finite measure spaces, respectively. The operator $K \in L_c(L_p(E_1, \Sigma_{1,E_1}, \mu_1), L_q(E_2, \Sigma_{2,E_2}, \mu_2))$ is an SM – operator if and only if there exists a kernel

$$k \in L_q(E_2 \times E_1, \Sigma_{2,E_2} \times \Sigma_{1,E_1}, \mu_2 \times \mu_1)$$

and

$$\sum_{n=1}^{\infty} \sum_{m=1}^{\infty} |kg_{mn}^*| \in L_1(E_2 \times E_1, \Sigma_{2,E_2} \times \Sigma_{1,E_1}, \mu_2 \times \mu_1)$$

such that

$$(Kf)(\bar{x}) = \int_{E_1} k(\bar{x}, \bar{y}) f(\bar{y}) d\mu_1(\bar{y})$$

for every $f \in L_p(E_1, \Sigma_{1,E_1}, \mu_1)$, $\bar{x} \in E_2$.

Proof: We have $K \in L_c(L_p(E_1, \Sigma_{1,E_1}, \mu_1), (E_2, \Sigma_{2,E_2}, \mu_2))$ by the previous discussion. If $\{e_n\}$ and $\{f_m\}$ are Schauder bases of $L_p(E_1, \Sigma_{1,E_1}, \mu_1)$ and $L_q(E_2, \Sigma_{2,E_2}, \mu_2)$, respectively, then by Theorem 5, we have $\{g_{mn}^*\}$, where $g_{mn}^*(\bar{x}, \bar{y}) = f_m^*(\bar{x})e_n(\bar{y})$, is a Schauder bases of $L_p(E_2 \times E_1, \Sigma_{2,E_2} \times \Sigma_{1,E_1}, \mu_2 \times \mu_1)$. Hence, for $M = \sup_m \{\|f_m(\bar{x})\|, \bar{x} \in E_2\}$, we have

$$\begin{aligned} & \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} \left\| \langle Ke_n(\bar{y}), f_m^*(\bar{x}) \rangle f_m(\bar{x}) \right\| \\ &= \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} \left\| \int_{E_2} \int_{E_1} (k(\bar{x}, \bar{y})e_n(\bar{y})) d\mu_1(\bar{y}) f_m^*(\bar{x}) d\mu_2(\bar{x}) \right\| \|f_m(\bar{x})\| \\ &\leq \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} \int_{E_2} \int_{E_1} |k(\bar{x}, \bar{y})e_n(\bar{y}) f_m^*(\bar{x})| d\mu_1(\bar{y}) d\mu_2(\bar{x}) \|f_m(\bar{x})\| \\ &\leq M \int_{E_2} \int_{E_1} \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} |k(\bar{x}, \bar{y})e_n(\bar{y}) f_m^*(\bar{x})| d\mu_1(\bar{y}) d\mu_2(\bar{x}) < \infty \end{aligned}$$

if and only if $k \in L_q(E_2 \times E_1, \Sigma_{2,E_2} \times \Sigma_{1,E_1}, \mu_2 \times \mu_1)$ and

$$\sum_{n=1}^{\infty} \sum_{m=1}^{\infty} |kg_{mn}^*| \in L_1(E_2 \times E_1, \Sigma_{2,E_2} \times \Sigma_{1,E_1}, \mu_2 \times \mu_1).$$

i.e., K is an SM – operator. Conversely, let K be an SM – operator, and let e_n , f_m , and g_{mn} be defined as above. Let us defined c_{nm} by

$$c_{nm} = \langle Ke_n, f_m^* \rangle$$

for $(m, n) \in \mathbb{N} \times \mathbb{N}$. Then

$$\sum_{n=1}^{\infty} \sum_{m=1}^{\infty} |c_{nm}| = \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} \left\| \langle Ke_n, f_m^* \rangle \right\| < \infty.$$

So there exists a $k \in L_q(E_2 \times E_1, \Sigma_{2,E_2} \times \Sigma_{1,E_1}, \mu_2 \times \mu_1)$ such that

$$\langle k, g_{mn}^* \rangle = c_{mn}$$

for all $(m, n) \in \mathbb{N} \times \mathbb{N}$. If K_0 denotes the operator induced by k , then by defining $h(\bar{x}, \bar{y}) = g^*(\bar{x})f(\bar{y})$, we have for all $f \in L_p(E_1, \Sigma_{1,E_1}, \mu_1)$, $g \in L_q(E_2, \Sigma_{2,E_2}, \mu_2)$ that

$$\begin{aligned} \langle K_0 f, g^* \rangle &= \int_{E_2} \left(\int_{E_1} k(\bar{x}, \bar{y}) f(\bar{y}) d\mu_1(\bar{y}) \right) g^*(\bar{x}) d\mu_2(\bar{x}) \\ &= \langle k, h \rangle \\ &= \left\langle k, \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \langle h, g_{mn} \rangle g_{mn}^* \right\rangle \\ &= \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \langle k, g_{mn}^* \rangle \langle h, g_{mn} \rangle \\ &= \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} c_{mn} \langle h, g_{mn} \rangle \\ &= \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \langle Ke_n, f_m^* \rangle \langle g^* f, f_m e_n^* \rangle \\ &= \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \langle Ke_n, f_m^* \rangle \langle g^*, f_m \rangle \langle f, e_n^* \rangle \\ &= \sum_{m=1}^{\infty} \left\langle \sum_{n=1}^{\infty} \langle f, e_n^* \rangle Ke_n, f_m^* \right\rangle \langle g^*, f_m \rangle \\ &= \left\langle \sum_{n=1}^{\infty} \langle f, e_n^* \rangle Ke_n, \sum_{m=1}^{\infty} \langle g^*, f_m \rangle f_m^* \right\rangle \\ &= \langle Kf, g^* \rangle. \end{aligned}$$

Hence, $K = K_0$. ■

The compactness of the operator K , can be describe as follows: Let us write again

$$f = \sum_{n=1}^{\infty} \langle f, e_n^* \rangle e_n$$

for every $f \in L_p(E_1, \Sigma_{1,E_1}, \mu_1)$ and a Schauder bases $\{e_n\}$ of $L_p(E_1, \Sigma_{1,E_1}, \mu_1)$ such that

$$Kf = \sum_{n=1}^{\infty} \langle f, e_n^* \rangle Ke_n$$

where

$$Ke_n = \sum_{m=1}^{\infty} \langle Ke_n, f_m^* \rangle f_m$$

for every Schauder bases $\{f_m\}$ of $L_q(E_2, \Sigma_{2,E_2}, \mu_2)$. Infact, since

$$\sum_{n=1}^p \langle f, e_n^* \rangle e_n$$

converge to f as $p \rightarrow \infty$, and K is continuous, then

$$K_p f = \sum_{n=1}^p \langle f, e_n^* \rangle Ke_n$$

must converge to Kf . Let K_p be the operator of finite rank defined as above, then

.

$$\begin{aligned}\|Kf - K_p f\| &\leq \left(\sum_{n=p+1}^{\infty} \left| \langle f, e_n^* \rangle \right| \|Ke_n\| \right) \leq \sum_{n=p+1}^{\infty} \left| \langle f, e_n^* \rangle \right| \sum_{m=p+1}^{\infty} \left\| \langle Ke_n, f_m^* \rangle f_m \right\| \\ &\leq \|f\| \sum_{n=p+1}^{\infty} \sum_{m=p+1}^{\infty} \left\| \langle Ke_n, f_m^* \rangle f_m \right\|\end{aligned}$$

Since $\|K\| < \infty$, then for any real number $\varepsilon > 0$, there is $p \in \mathbb{N}$, such that

$$\sum_{n=p+1}^{\infty} \sum_{m=p+1}^{\infty} \left\| \langle Ke_n, f_m^* \rangle f_m \right\| < \frac{\varepsilon}{2}$$

and,

$$\|K - K_p\| \leq \|K - K_p\| = \sum_{n=p+1}^{\infty} \sum_{m=p+1}^{\infty} \left\| \langle Ke_n, f_m^* \rangle f_m \right\| < \frac{\varepsilon}{2}$$

This shows that K is the limit in norm of a sequence of operators of finite rank. Hence, K is compact. ■

4. CONCLUDING REMARKS

The SM -operators $K : X \rightarrow Y$ where $X = L_p(E_1)$ $Y = L_q(E_2)$ which leave aside orthogonality (orthonormality) of their Schauder bases does not guarantee that $\|K\| = \|K^*\|$, but due to biorthogonality of these bases, will automatically turns this condition to be true especially when $p = 2, q = 2$. For further works, we have been doing on classical sequence spaces $\ell_p, 1 < p < \infty$.

Acknowledgments. The first author is partially supported by BPPS DIKTI. We thank all anonymous referee who carefully read this paper.

REFERENCES

1. Conway, J.B., 1990. *A Course in Functional Analysis*, Springer Verlag, New York.
2. Dapa P.S., 2000. *On strong M-bases in Banach spaces with PRI*, Collect. Math. 51, 3, 277-284.
3. Darmawijaya, S., Ansori, M., Supama, 2006. *On The SM-operators*, Submitted to Berkala Ilmiah MIPA UGM, Yogyakarta.
4. Johnson W.B., Rosenthal, H.P., Zippin, M., 1971. *On bases, finite-dimensional decompositions and weaker structures in Banach spaces*, Israel J. Math. 9, 488-506.
5. Kuo, H.H., 1975. *Gaussian Measures in Banach spaces*, Springer Verlag, Berlin.
6. Lindenstrauss J and Tzafriri L, 1996. *Classical Banach Spaces I and II*, Springer Verlag, New York.
7. Morrison, T.J., 2000. *Functional Analysis: An Introduction to Banach Space Theory*, John Wiley and Sons Inc., New York.
8. Royden, H.L., 1988. *Real Analysis*, Macmillan Publishing Company, New York.
9. Weidmann, J., 1980. *Linear Operators in Hilbert Spaces*, Springer Verlag. New York.
10. Zippin, M., 1968. *A remark on bases and reflexivity in Banach spaces*, Israel J. Math. 6, 74-79

Muslim Ansori: Mathematics Departement, Universitas Lampung, Bandar Lampung, Indonesia, Ph.D Mathematics Student UGM.
Email: ansomath@yahoo.com.

Soeparna Darmawijaya and Supama: Analysis & Application Research Group, Mathematics Departement, Universitas Gadjah Mada, Yogyakarta, Indonesia.
Email: maspomo@yahoo.com

A REAL MEASURE VERSUS A VECTOR MEASURE (ON THE SAME MEASURABLE SPACE)

SOEPARNA DARMAWIJAYA

Abstract. Let $(\Omega, \mathcal{A}, \mu)$ be completely σ -finite measure space, $A \in \mathcal{A}$, and $\partial : \mathcal{A} \rightarrow X$, with X a Banach space, be additive. In this paper, we prove that

1. $\partial \in \mathcal{V}^\infty(\mu, X) \Leftrightarrow \|\partial(A)\| \ll K\mu(A)$ for some $K > 0$.
2. $1 < p, q < \infty$ and $\frac{1}{p} + \frac{1}{q} = 1$, $\partial \in \mathcal{V}^q(\mu, X) \Leftrightarrow$ for every $\alpha \in \mathcal{R}$, $\|\alpha\partial(A)\| \ll K|\alpha|^p\mu(A)$ for some $K > 0$.

We shall also show that the Banach space $V^q(\mu, X)$ and $L_c(\mathcal{L}_p(E, \mu), X)$ are isomorphic and isometric.

Key words and Phrases: Real measure, vector measure.

1. INTRODUCTION

For simplicity we always assume that $(\Omega, \mathcal{A}, \mu)$ is a completely σ -finite measure space, X a Banach space.

An additive set function $\partial : \mathcal{A} \rightarrow X$ is said to be ***Lipchitzean with respect to μ*** if there is a constant $K > 0$ such that

$$\|\partial(A)\| \ll K\mu(A)$$

for every $A \in \mathcal{A}$.

First of all we have that ∂ is a vector measure on the measurable space $(\Omega, \mathcal{A})$,

2000 Mathematics Subject Classification: 28B05

since for every sequence of disjoint measurable sets $\{A_k\}$ with $\bigcup_{k=1}^n A_k$ we have

$$\left\| \partial(A) - \sum_{k=1}^n \partial(A_k) \right\| = \left\| \partial(A - \bigcup_{k=1}^n A_k) \right\| \leq K\mu(A - \bigcup_{k=1}^n A_k) \rightarrow 0$$

as $n \rightarrow \infty$.

2. MAIN RESULTS

We begin this section by proving the following theorem.

Theorem 2.1. *Let $\mathcal{V}^\infty(\mu, X)$ be the collection of additive set functions $\partial : \mathcal{A} \rightarrow X$ which are Lipchitzian with respect to μ . Then, $\mathcal{V}^\infty(\mu, X)$ is a Banach space with respect to the norm:*

$$\|\partial\|_\infty = \inf\{K > 0 : \|\partial(A)\| \leq K\mu(A) \text{ for } A \in \mathcal{A}\}$$

for every $\partial \in \mathcal{V}^\infty(\mu, X)$.

Proof. It is easy to prove that $\mathcal{V}^\infty(\mu, X)$ is a linear space and $(\mathcal{V}^\infty(\mu, X), \|\cdot\|_\infty)$ is a norm space.

Let $\{\partial_n\}$ be a Cauchy sequence in the norm space $(\mathcal{V}^\infty(\mu, X), \|\cdot\|_\infty)$. That is, for any number $\varepsilon > 0$ and $A \in \mathcal{A}$ there is a positive integer n_0 such that for every two positive integers $m, n \geq n_0$ we have

$$\|\partial_m - \partial_n\|_\infty < \frac{\varepsilon}{\mu(A) + 1}$$

It implies that

$$\|\partial_m(A) - \partial_n(A)\| = \|(\partial_m - \partial_n)(A)\| \leq \|(\partial_m - \partial_n)\|_\infty \mu(A) < \varepsilon$$

that is, for every $A \in \mathcal{A}$ we have $\{\partial_n(A)\}$ is a Cauchy sequence in a Banach space X . Hence, there is a vector $\partial(A) \in X$ such that

$$\lim_{n \rightarrow \infty} \partial_n(A) = \partial(A).$$

Further, it is easy to see that $\partial : \mathcal{A} \rightarrow X$ is additive and Lipchitzian with respect to μ , i.e., $\partial \in \mathcal{V}^\infty(\mu, X)$.

Let $E \in \mathcal{A}$ and $\mathcal{M}(E, \mu)$ the collection of measurable functions on E . If $\mathcal{S}$ denotes the collection of simple functions, it is well known that $\mathcal{S}$ is dense in $\mathcal{M}(E, \mu)$, that is, for any $f \in \mathcal{M}(E, \mu)$ there is sequence $\{s_n\} \subset \mathcal{S}$ which is convergent to f a.e. on E .

Now, let $s \in \mathcal{S}$ and $s \in \mathcal{L}_1(E, \mu)$:

$$s = \sum_{i=1}^m \alpha_i \chi_{E_i}$$

and $\partial \in \mathcal{V}^\infty(\mu, X)$. The ∂ -integral on E of s is the vector

$$\int_E s \, d\partial = \sum_{i=1}^m \alpha_i \cdot \partial(E_i \cap E)$$

It implies that

$$\begin{aligned} \left\| \int_E s \, d\partial \right\| &= \left\| \sum_{i=1}^m \alpha_i \cdot \partial(E_i \cap E) \right\| \ll \|\partial\|_\infty \sum_{i=1}^m |\alpha_i| \mu(E_i \cap E) \\ &= \|\partial\|_\infty \int_E |s| \, d\mu = \|\partial\|_\infty \cdot \|s\|_1 < \infty \end{aligned}$$

Further, if $f \in \mathcal{L}_1(E, \mu)$ there is a sequence $\{s_n\} \subset \mathcal{S}$ such that $\lim_{n \rightarrow \infty} s_n = f$ a. e. on E . So, we have $\partial \in \mathcal{V}^\infty(\mu, X) \rightarrow u \int_E \cdot \, d\partial \in L_c(\mathcal{L}_1(E, \mu), X)$,

$$\begin{aligned} \left\| \int_E f \, d\partial \right\| &= \lim_{n \rightarrow \infty} \left\| \int_E s_n \, d\partial \right\| \ll \|\partial\|_\infty \cdot \|f\|_1 \\ &= \|\partial\|_\infty \int_E |f| \, d\mu = \|\partial\|_\infty \cdot \|f\|_1 \end{aligned}$$

and

$$\|u\| \ll \|\partial\|_\infty$$

We conclude the last result into the following theorem.

Theorem 2.2 *The function u :*

$$\partial \in \mathcal{V}^\infty(\mu, X) \rightarrow u \int_E \cdot \, d\partial \in L_c(\mathcal{L}_1(E, \mu), X)$$

is linearity continuous and

$$\|u\| \ll \|\partial\|_\infty$$

Theorem 2.3 $\mathcal{V}^\infty(\mu, X)$ and $L_c(\mathcal{L}_1(E, \mu), X)$ are isomorphic and isometric.

Proof. (i) From Theorem 1.3 we have had

$$\partial \in \mathcal{V}^\infty(\mu, X) \rightarrow u \int_E \cdot d\partial \in L_c(\mathcal{L}_1(E, \mu), X)$$

is linearity continuous and

$$\|u\| \ll \|\partial\|_\infty \quad (a)$$

Let θ be the null vector in $L_c(\mathcal{L}_1(E, \mu), X)$. Then, for every $f \in \mathcal{L}_1(E, \mu)$ we have

$$\int_E f d\partial = \theta,$$

particullary for $f = \mathcal{X}_A$ and $A \in \mathcal{A}$, we have

$$\int_E \mathcal{X}_A d\partial = \partial(A) = \theta$$

If means that u is injective.

(ii) Conversely, let $d \in L_c(\mathcal{L}_1(E, \mu), X)$ arbitrarily. We define $\partial : \mathcal{A} \rightarrow \mathcal{X}$ with $\partial(A) = d(\mathcal{X}_A)$ for every $A \in \mathcal{A}$. Hence, for every $A, B \in \mathcal{A}$ with $A \cap B = \emptyset$ we have

$$\begin{aligned} \partial(A \cup B) &= d(\mathcal{X}_{A \cup B}) = d(\mathcal{X}_A + \mathcal{X}_B) = d(\mathcal{X}_A) + d(\mathcal{X}_B) \\ &= \partial(A) + \partial(B) \end{aligned}$$

and

$$\begin{aligned} \|\partial(A)\| &= \left\| \int_E \mathcal{X}_A d\partial \right\| \\ &\ll \|d(\mathcal{X}_A)\| \ll \|d\| \|\mathcal{X}_A\|, \\ &= \|d\| \int_E \mathcal{X}_A d\mu \\ &= \|d\| \mu(A). \end{aligned}$$

We conclude that $\partial \in \mathcal{V}^\infty(\mu, X)$ and u is bijective and

$$\|\partial\|_\infty \ll \|u\| \quad (b)$$

Therefore, from the results of (i) and (ii) we may conclude that $\mathcal{V}^\infty(\mu, X)$ and $L_c(\mathcal{L}_1(E, \mu), X)$ are isomorphic and isometric.

Corollary 2.4 $\mathcal{V}^\infty(\mu, \mathcal{R})$ and $\mathcal{L}_\infty(E, \mu) = \mathcal{L}_1(E, \mu)^* = L_c(\mathcal{L}_1(E, \mu), \mathcal{R})$ are isomorphic and isometric.

We observe that if $\partial \in \mathcal{V}^\infty(\mu, \mathcal{R})$ and $\partial \ll \mu$, that is ∂ is absolutely continuous with respect to μ , then there is $f \in \mathcal{L}_1(E, \mu)$ such that $\partial(A) = \int_A f d\mu$ for every $A \in \mathcal{A}$. Because $\partial(A) \geq 0$ it implies that f is a non negative measurable function. Hence, we have **Radon Nikodym Theorem**. Further, we also have $\left| \int_A f d\mu \right| \ll \|\partial\|_\infty \mu(A) = \int_A \|\partial\|_\infty d\mu$ for every $A \in \mathcal{A}$. It follows that $|f| \ll \|\partial\|_\infty$ a.e on E , $f \in \mathcal{L}_1(E, \mu)$, and $\|f\|_1 \leq \|\partial\|_\infty$.

Again, we assume that $(\Omega, \mathcal{A}, \mu)$ is a completely σ -finite measure space and X is a Banach space. We shall generalize the results before.

Definition 2.5 Let $1 < p, q < \infty$ and $\frac{1}{p} + \frac{1}{q} = 1$. An additive set function $\partial : \mathcal{A} \rightarrow \mathcal{R}$ is said to be **p th-Lipschitzian with respect to the measure μ** if there is a constant $K > 0$ such that for every $\alpha \in \mathcal{R}$ and $A \in \mathcal{A}$ we have

$$\|\alpha \partial(A)\| \ll K |\alpha|^p \mu(A).$$

Since for every sequence of disjoint measurable sets $\{A_k\} \subset \mathcal{A}$ with $A = \bigcup_{k=1}^{\infty} A_k$ we have

$$\left\| \alpha \partial(A) - \sum_{k=1}^n \alpha \partial(A_k) \right\| = \left\| \alpha \partial\left(A - \bigcup_{k=1}^n A_k\right) \right\| \ll K |\alpha|^p \mu\left(A - \bigcup_{k=1}^n A_k\right) \rightarrow 0 \text{ as } n \rightarrow \infty,$$

we conclude that ∂ is countable additive on $\mathcal{A}$, hence ∂ is a vector measure on the measurable space $(\Omega, \mathcal{A})$. Let $\mathcal{V}^q(\mu, X)$, $1 < p, q < \infty$ and $\frac{1}{p} + \frac{1}{q} = 1$, be the collection of additive set functions $\partial : \mathcal{A} \rightarrow \mathcal{R}$ which are p^{th} -Lipschitzian with respect to the measure μ .

Theorem 2.6 $\mathcal{V}^q(\mu, X)$ is a Banach space with respect to the norm $\|\cdot\|_q$:

$$\|\partial\|_q = \inf\{K > 0 : \|\alpha \partial(A)\| \ll K |\alpha|^p \mu(A), \alpha \in \mathcal{R} \text{ \& } A \in \mathcal{A}\}$$

Proof. (i) $\mathcal{V}^q(\mu, X)$ is a linear space, since for any $k \in \mathcal{R}$, $\partial, \partial_1, \partial_2 \in \mathcal{V}^q(\mu, X)$ we have

$$\|\alpha k \partial(A)\| = |k| \cdot \|\alpha \partial(A)\| \ll |k| \cdot K |\alpha|^p \mu(A) = K_1 |\alpha|^p \mu(A)$$

or $k \partial \in \mathcal{V}^q(\mu, X)$, where $K_1 = |k|K$, i.e., $k \partial \in \mathcal{V}^q(\mu, X)$ and

$$\begin{aligned} \|\alpha(\partial_1 + \partial_2)(A)\| &= \|\alpha \partial_1(A) + \alpha \partial_2(A)\| \ll \|\alpha \partial_1(A)\| + \|\alpha \partial_2(A)\| \\ &\ll K_1 |\alpha|^p \mu(A) + K_2 |\alpha|^p \mu(A) \ll K |\alpha|^p \mu(A) \end{aligned}$$

for some $K_1 > 0$ and $K_2 > 0$, where $K = \max\{K_1, K_2\}$ i.e. $\partial_1 + \partial_2 \in \mathcal{V}^q(\mu, X)$

(ii) $\|\cdot\|_q$ is a norm on $\mathcal{V}^q(\mu, X)$, since

(N1) for any $\partial \in \mathcal{V}^q(\mu, X)$ we have

$$\begin{aligned} \|\partial\|_q &= \inf\{K > 0 : \|\alpha\partial(A)\| \ll K|\alpha|^p\mu(A), \alpha \in \mathcal{R} \text{ \& } A \in \mathcal{A}\} \\ &\geq 0 \end{aligned}$$

and $\|\partial\|_q = 0$ if and only if $\partial = \theta$, θ is the null vector of $\mathcal{V}^q(\mu, X)$.

(N2) Since for every $\alpha \in \mathcal{R}$ and $A \in \mathcal{A}$ we have $\|\alpha\partial\|_q |\alpha|^p\mu(A)$, then for any $k \in \mathcal{R}$ we have $k\partial \in \mathcal{V}^q(\mu, X)$ and

$$\|\alpha k\partial(A)\| = |k| \cdot \|\alpha\partial(A)\| \ll |k| \cdot \|\partial\|_q |\alpha|^p\mu(A)$$

which implies $\|k\partial\|_q = |k| \cdot \|\partial\|_q$.

(N3) For every $\partial_1, \partial_2 \in \mathcal{V}^q(\mu, X)$ and, hence, for every $\alpha \in \mathcal{R}$ and $A \in \mathcal{A}$ we have

$$\begin{aligned} \|\alpha(\partial_1 + \partial_2)(A)\| &= \|\alpha\partial_1(A) + \alpha\partial_2(A)\| \ll \|\alpha\partial_1(A)\| + \|\alpha\partial_2(A)\| \\ &\ll \|\partial_1\|_q |\alpha|^p\mu(A) + \|\partial_2\|_q |\alpha|^p\mu(A) \\ &= \left(\|\partial_1\|_q + \|\partial_2\|_q\right) |\alpha|^p\mu(A) \end{aligned}$$

which implies $\|\partial_1 + \partial_2\|_q \ll \|\partial_1\|_q + \|\partial_2\|_q$.

(iii) $\mathcal{V}^q(\mu, X)$ is complete a.l. $\mathcal{V}^q(\mu, X)$ is a Banach space with respect to the norm $\|\cdot\|_q$: Let $\{\partial_n\}$ be an arbitrary Cauchy sequence in $\mathcal{V}^q(\mu, X)$. Hence, for every $\alpha \in \mathcal{R}$, $A \in \mathcal{A}$ and number $\varepsilon > 0$ there is a positive integer n_0 such that for any two positive integers $m, n \geq n_0$ we have

$$\|\partial_m - \partial_n\|_q < \frac{\varepsilon}{|\alpha|^{p-1}\mu(A)}$$

which implies

$$\|\partial_m(A) - \partial_n(A)\| = \frac{1}{|\alpha|} \|\alpha(\partial_m - \partial_n)(A)\| \ll \frac{1}{|\alpha|} \|\partial_m - \partial_n\|_q |\alpha|^p\mu(A) < \varepsilon,$$

that is, for every $A \in \mathcal{A}$ we have $\{\partial_n(A)\}$ is a Cauchy sequence on the Banach space. Hence, there is $\partial(A) \in X$ such that $\partial(A) = \lim_{n \rightarrow \infty} \partial_n(A)$. Further, $\partial \in \mathcal{V}^q(\mu, X)$ since for every $A, B \in \mathcal{A}$ with $A \cap B = \emptyset$ and $\alpha \in \mathcal{R}$ we have

$$\partial(A \cup B) = \lim_{n \rightarrow \infty} \partial(A \cup B) = \lim_{n \rightarrow \infty} (\partial_n(A) + \partial_n(B)) = \partial(A) + \partial(B)$$

and

$$\begin{aligned} \|\alpha\partial(A)\| &= \left\| \alpha \lim_{n \rightarrow \infty} \partial_n(A) \right\| \\ &= \left\| \lim_{n \rightarrow \infty} \alpha\partial_n(A) \right\| \ll \lim_{n \rightarrow \infty} \|\partial_n\|_q |\alpha|^p\mu(A) \ll K|\alpha|^p\mu(A), \end{aligned}$$

where $K = \sup_{n \geq 1} \|\partial_n\|_q$.

Let $\partial \in \mathcal{V}^q(\mu, X)$ and $E \in \mathcal{A}$. We define an operator integral $\int_E \cdot d\partial$ such that for every simple function $s = \sum_{k=1}^n \alpha_k \chi_{E_k} \in \mathcal{S}$,

$$\int_E s d\partial = \sum_{k=1}^n \alpha_k \partial(E_k \cap E)$$

It implies that

$$\begin{aligned} \left\| \int_E s d\partial \right\| &= \left\| \sum_{k=1}^n \alpha_k \partial(E_k \cap E) \right\| \ll \sum_{k=1}^n \|\alpha_k \partial(E_k \cap E)\| \\ &\ll \|\partial\|_q \cdot \sum_{k=1}^n |\alpha_k|^p \mu(E_k \cap E) = \|\partial\|_q \int_E |s|^p d\mu < \infty \end{aligned}$$

whenever $s \in \mathcal{L}_p(E, \mu)$. Therefore, for every $f \in \mathcal{L}_p(E, \mu)$ we have

$$\left\| \int_E f d\partial \right\| \ll \|\partial\|_q \int_E |f|^p d\mu < \infty.$$

Hence, we observe and conclude that $\int_E \cdot d\partial \in L_c(\mathcal{L}_p(E, \mu), X)$ for every $\partial \in \mathcal{V}^q(\mu, X)$ and the function u :

$$\partial \in \mathcal{V}^q(\mu, X) \rightarrow u \int_E \cdot d\partial \in L_c(\mathcal{L}_p(E, \mu), X)$$

is linearly continuous with

$$\|u\| \ll \|\partial\|_q.$$

Theorem 2.7 $\mathcal{V}^q(\mu, X)$ and $L_c(\mathcal{L}_p(E, \mu), X)$ are isometric and isomorphism.

Proof. We define a function u :

$$\partial \in \mathcal{V}^q(\mu, X) \rightarrow u \int_E \cdot d\partial \in L_c(\mathcal{L}_p(E, \mu), X)$$

It is similar to the last discussion above. The rest of the proof is similar to that of Theorem 2.3.

Corollary 2.8 If $1 < p, q < \infty$ and $\frac{1}{p} + \frac{1}{q} = 1$ the Banach space $\mathcal{V}^q(\mu, X)$ and $\mathcal{L}_q(E, \mu) = (\mathcal{L}_p(E, \mu))^*$ are isometric and isomorphic.

REFERENCES

1. CONWAY, J.B., *A Course in Functional Analysis*, Springer Verlag, New York, 1985.
2. DUNFORD N. AND J. SCHWARTZ, *Linear Operator I*, Interscience, New York, 1968.
3. LINDENSTRAUSS J AND TZAFRIRI L., *Classical Banach Spaces I and II*, Springer Verlag, New York, 1996.
4. MORISSON, T.J., *Functional Analysis: An Introduction to Banach Space Theory*. John Wiley & Sons, New York-Brisbane-Singapore-Toronto, 2001.
5. WEIDMANN, J., *Linear Operators in Hilbert Spaces*, Springer Verlag, New York, Berlin, 1980.

SOEPARNA DARMAWIJAYA: Department of Mathematics, Gadjah Mada University,
Yogyakarta, Indonesia.

Email: Soeparna_D@ugm.ac.id

SOME CONVERGENCE THEOREMS FOR THE HENSTOCK-STIELTJES INTEGRAL

UMI MAHNUNA HANUNG

Abstract. This paper is to discuss the Monotone Convergence Theorem, Fatou's Lemma, and the Dominated Convergence Theorem for the Henstock-Stieltjes integral.

Key words and Phrases: monotonically increasing function, continuous function, convergence theorem, and Henstock-Stieltjes integral.

1. INTRODUCTION

Given closed interval $[a, b] \subset \mathcal{R}$, $f : [a, b] \rightarrow \mathcal{R}$, $\varphi : [a, b] \rightarrow \mathcal{R}$ a bounded function and $\delta : [a, b] \rightarrow \mathcal{R}$ a positive function. If $D = \{([u_i, v_i], x_i)\}_{i=1}^n = \{([u, v], x)\}$ is a Perron δ -fine partition on $[a, b]$ we write

$$S(f, \varphi, D) = (D) \sum_{i=1}^n f(x_i) \Delta_i \varphi = (D) \sum f(x) \Delta \varphi$$

with $\Delta_i \varphi = \varphi(v_i) - \varphi(u_i)$ (or $\Delta \varphi = \varphi(v) - \varphi(u)$) for every $1 \leq i \leq n$.

A function f defined on $[a, b]$ is said to be Henstock-Stieltjes integrable with respect to φ on $[a, b]$ or, in short $f \in HS(\varphi, [a, b])$ if there is a number A such that for every $\varepsilon > 0$ there exists a positive function δ defined on $[a, b]$ such that for any Perron δ -fine partition $D = \{([u_i, v_i], x_i)\}_{i=1}^n$ on $[a, b]$ implies

$$|S(f, \varphi, D) - A| < \varepsilon .$$

The number A is called *HS-integral with respect to φ* on $[a, b]$ and, written by

$$A = (HS) \int_a^b f \, d\varphi.$$

Let $\mathfrak{S}([a, b])$ be a collection of all closed subinterval in $[a, b]$. A Henstock-Stieltjes-primitive function of f on $[a, b]$ as follows

$$F(x) = F([a, x]) = (HS) \int_a^x f d\varphi, \text{ with } F(a) = 0.$$

Hence, $f \in HS(\varphi, [a, b])$ if and only if there exists a real function F defined on $[a, b]$ such that for every $\varepsilon > 0$ there exists a positive function δ defined on $[a, b]$ such that

$$|(D) \sum [f(x_i) \Delta_i \varphi - F([u_i, v_i])]| < \varepsilon$$

for every Perron δ -fine partition $D = \{([u_i, v_i], x_i)\}_{i=1}^n$ on $[a, b]$. Where, if $\varphi : [a, b] \rightarrow \mathcal{R}$ be monotone and continuous function then we have that F is continuous function on $[a, b]$ (See Theorem 2.1).

2. MAIN RESULTS

As a starting point, we will first give a condition such that the primitive function of f , which is Henstock-Stieltjes integrable, is continuous.

Theorem 2.1. *Let f is Henstock-Stieltjes integrable function with respect to φ on $[a, b]$ with primitive F . If f be zero function on $[a, b]$ or φ be monotone and continuous function on $[a, b]$ then F be continuous on $[a, b]$.*

Proof. The proof for f be zero function is easy. Now we prove for φ be monotone and continuous on $[a, b]$. Given $y \in [a, b]$ and $\varepsilon > 0$. Since F be primitive function of f , (See [6]) there exists positive function δ on $[a, b]$ such that for any $D = \{([u_i, v_i], x_i)\}$ be Perron δ -fine partition on $[a, b]$ we have

$$(D) \sum |f(x) \Delta \varphi - F([u, v])| < \frac{\varepsilon}{4}.$$

By recalling that f is Henstock-Stieltjes integrable function, for each Perron δ -fine partition D on $[a, b]$ we could find the real number K_D with $K_D = \max\{|f(x)| : x \in [a, b] \text{ and } ([u, v], x) \in D\}$. And then, we take

$$K = \sup\{K_D : D \text{ be Perron } \delta\text{-fine partition on } [a, b]\}.$$

Since φ be continuous function on $[a, b]$, then for $y \in [a, b]$ and $\frac{\varepsilon}{2(K+1)} > 0$ there is real number $\eta = \eta(y, \frac{\varepsilon}{2(K+1)}) > 0$ such that for any $z \in [a, b]$ and $|y - z| < \eta$ implies

$$|\varphi(y) - \varphi(z)| < \frac{\varepsilon}{2(K+1)}.$$

Without loss of generality if assumed $y \leq z$, we have $[y, z] \subseteq [a, b]$. It is easy to be understood that δ is a positive function on $[y, z]$ and $cl([a, b] \setminus [y, z])$ too. Therefore,

if D_1 and D_2 are Perron δ -fine partitions on $[y, z]$ and $cl([a, b] \setminus [y, z])$, respectively, then $D_1 \cup D_2 = D^*$ is a Perron δ -fine partition on $[a, b]$ such that implies

$$(D^*) \sum_{i=1}^n |f(x) \Delta \varphi - F([u, v])| < \frac{\varepsilon}{4},$$

and based on Henstock's Lemma (See [6]) we obtain

$$(D_1) \sum |f(x) \Delta \varphi - F([u, v])| < \frac{\varepsilon}{2}.$$

Recall that φ be monotone function, then

$$(D_1) \sum |\Delta \varphi| = (D_1) \left| \sum \Delta \varphi \right| = |\varphi(z) - \varphi(y)| \geq 0.$$

Hence, we have

$$\begin{aligned} |F(y) - F(z)| &\leq (D_1) \sum |F([u, v])| \\ &= (D_1) \sum |F([u, v]) - f(x) \Delta \varphi + f(x) \Delta \varphi| \\ &\leq (D_1) \sum (|F([u, v]) - f(x) \Delta \varphi| + |f(x) \Delta \varphi|) \\ &= (D_1) \sum |F([u, v]) - f(x) \Delta \varphi| + (D_1) \sum |f(x) \Delta \varphi| \\ &\leq (D_1) \sum |F([u, v]) - f(x) \Delta \varphi| + (D_1) \sum K_{D^*} |\Delta \varphi| \\ &\leq (D_1) \sum |F([u, v]) - f(x) \Delta \varphi| + (D_1) \sum K |\Delta \varphi| \\ &= (D_1) \sum |F([u, v]) - f(x) \Delta \varphi| + K |\varphi(z) - \varphi(y)| \\ &< \frac{\varepsilon}{2} + K \cdot \frac{\varepsilon}{2(K+1)} < \varepsilon. \end{aligned}$$

This shows that F be continuous on $[a, b]$. This completes the proof.

Let $\{f_n\}$ be a sequence of Henstock-Stieltjes integrable functions with respect to φ on $[a, b]$ and suppose that $\{f_n\}$ converges to f almost everywhere on $[a, b]$. Is the function f Henstock-Stieltjes integrable with respect to φ on $[a, b]$ and does $(HS) \int_a^b f d\varphi = \lim_{n \rightarrow \infty} (HS) \int_a^b f_n d\varphi$? Any theorem that provides conditions for this problem to have an affirmative answer is called as a convergence theorem.

Theorem 2.2. (Monotone Convergence Theorem) *Let φ be monotonically increasing and continuous function defined on $[a, b]$, f_n be Henstock-Stieltjes integrable functions with respect to φ on $[a, b]$ for each n , and f be real function defined on $[a, b]$. If*

- (i) $\{f_n\}$ converges almost everywhere to f on $[a, b]$,
- (ii) $\{f_n\}$ be a monotone sequence on $[a, b]$,

(iii) $\{F_n([a, b])\} = \{\lim_{n \rightarrow \infty} \int_a^b f_n d\varphi\}$ converges to A ,

then f is Henstock-Stieltjes integrable with respect to φ on $[a, b]$ and

$$(HS) \int_a^b f d\varphi = A.$$

Proof. Without loss of generality, we may assume that the sequence $\{f_n\}$ is convergent on $[a, b]$. Suppose that the sequence $\{f_n\}$ is increasing. Let $\varepsilon > 0$. Since $f_n \in HS(\varphi, [a, b])$ for all n , there exists a positive function δ_n defined on $[a, b]$ such that for any Perron δ_n -fine partition $D_n = \{([u_i, v_i], x_i)\}_{i=1}^n$ on $[a, b]$ implies

$$(D_n) \sum |f_n(x_i) \Delta_i \varphi - F_n([u_i, v_i])| < \frac{\varepsilon}{3}.$$

Since $\{f_n\}$ be an monotonically increasing sequence of Henstock-Stieltjes integrable functions defined on $[a, b]$ and φ be monotonically increasing function, then for each $[u, v] \subseteq [a, b]$ we get

$$F_n([u, v]) \leq F_{n+1}([u, v]) \leq M,$$

for a real number M . Hence, the sequence $\{F_n([u, v])\}$ converges to

$$F([u, v]) = \lim_{n \rightarrow \infty} F_n([u, v]).$$

In other hand, sequence $\{F_n([a, b])\}$ converges to A . It means for $\varepsilon > 0$ above there exists a natural number n_0 such that for each $n \geq n_0$ implies

$$|F_n([a, b]) - A| < \frac{\varepsilon}{3}.$$

Since $\{f_n\}$ converges to f on $[a, b]$, then for each $x \in [a, b]$ and $\varepsilon > 0$ above we could choose the natural number $N_{(\varepsilon, x)} \geq n_0$ such that for $n \geq N_{(\varepsilon, x)}$ implies

$$|f_n(x) - f(x)| < \frac{\varepsilon}{3 \cdot 2^{N_{(\varepsilon, x)}}}.$$

We construct a positive function δ defined on $[a, b]$ by formula $\delta(x) = \delta_{N_{(\varepsilon, x)}}(x)$. For any Perron δ -fine partition $D = \{([u_i, v_i], x_i)\}_{i=1}^n$ on $[a, b]$ take a natural number $p = \min\{N_{(\varepsilon, x_i)} : ([u_i, v_i], x_i) \in D\}$, we get

$$\begin{aligned} F_p([a, b]) &= (D) \sum F_p([u_i, v_i]) \\ &\leq (D) \sum F_{N_{(\varepsilon, x)}}([u_i, v_i]) \\ &\leq (D) \sum F([u_i, v_i]) \\ &= F([a, b]) = A. \end{aligned}$$

Therefore,

$$\begin{aligned}
|(D) \sum f(x_i) \Delta_i \varphi - A| &\leq (D) \sum |f(x_i) - f_{N(\varepsilon, x_i)}(x_i)| \Delta_i \varphi \\
&\quad + (D) \sum |f_{N(\varepsilon, x_i)} \Delta_i \varphi - F_{N(\varepsilon, x_i)}([u_i, v_i])| \\
&\quad + |(D) \sum F_{N(\varepsilon, x_i)}([u_i, v_i]) - A| \\
&\leq (D) \sum |f(x_i) - f_{N(\varepsilon, x_i)}(x_i)| \Delta_i \varphi \\
&\quad + (D) \sum |f_{N(\varepsilon, x_i)} \Delta_i \varphi - F_{N(\varepsilon, x_i)}([u_i, v_i])| \\
&\quad + |(D) \sum F_p([u_i, v_i]) - A| \\
&< \varepsilon.
\end{aligned}$$

This shows that f is Henstock-Stieltjes integrable with respect to φ on $[a, b]$ and $(HS) \int_a^b f d\varphi = A$. This completes the proof.

Theorem 2.3. *Let f_1, f_2, g , and h be real functions defined on $[a, b]$ and φ be monotonically increasing and continuous function on $[a, b]$. If $g \leq f_1, f_2 \leq h$ with $f_1, f_2, g, h \in HS(\varphi, [a, b])$ then $\max\{f_1, f_2\}, \min\{f_1, f_2\} \in HS(\varphi, [a, b])$.*

Proof. Assume that $g = \ominus$ is zero function, that is $g(x) = 0$ for every $x \in [a, b]$. Let F_1 and F_2 be Henstock-Stieltjes primitive functions of f_1 and f_2 on $[a, b]$, respectively. Since $f_1, f_2 \in HS(\varphi, [a, b])$ then $f_1, f_2 \in HS(\varphi, [u, v])$ for any $[u, v] \subseteq [a, b]$. Moreover, defined

$$F([u, v]) = \max\{F_1([u, v]), F_2([u, v])\}$$

for every $[u, v] \subseteq [a, b]$. Clear that F is sub-additive function, i.e.,

$$F([u, v]) \leq F([u_1, v_1]) + F([u_2, v_2])$$

for every $[u, v] = [u_1, v_1] \cup [u_2, v_2]$ with $\text{int}([u_1, v_1]) \cap \text{int}([u_2, v_2]) = \emptyset$. Since φ be monotonically increasing function, for any partition $D_0 = \{([u, v])_{i=1}^n$ on $[a, b]$ we get

$$0 \leq (D_0) \sum F([u_i, v_i]) \leq F_h\{[a, b]\}$$

where F_h be Henstock-Stieltjes primitive function of h . Given any real number $\varepsilon > 0$. By hypothesis, there exists $\delta > 0$ such that

$$(D) \sum |f_j(x_i) \Delta_i \varphi - F_j([u_i, v_i])| < \frac{\varepsilon}{3}, \quad (j = 1, 2)$$

whenever $D = \{([u_i, v_i], x_i)\}_{i=1}^n$ is a Perron δ -fine partition on $[a, b]$. Taking the supremum over all such Perron δ -fine partition (D) on $[a, b]$, that is

$$A = \sup\{(D) \sum F([u_i, v_i]) : D \text{ is Perron } \delta - \text{fine partition on } [a, b]\}.$$

We will show that A is value of Henstock-Stieltjes integral of $\max\{f_1, f_2\}$ on $[a, b]$. Now let D be any Perron δ -fine partition on $[a, b]$ and $([u_i, v_i], x_i) \in D$. For each $j = 1, 2$ define

$$\varpi_j([u_i, v_i]) = \sup\{(D_*) \sum |f_j(x_{i_k}) \Delta_{i_k} \varphi - F_j([u_{i_k}, v_{i_k}])| : D_* \text{ is Perron } \delta\text{-fine partition on } [u_i, v_i]\}.$$

For each $([u_i, v_i], x_i)$ in D and D_* be any Perron δ -fine partition on $[u_i, v_i]$ we have

$$|f_j(x_i) \Delta_i \varphi - F_j([u_i, v_i])| \leq (D_*) \sum |f_j(x_{i_k}) \Delta_{i_k} \varphi - F_j([u_{i_k}, v_{i_k}])| < \varepsilon$$

with $j = 1, 2$. Hence, for $j = 1, 2$

$$|f_j(x_i) \Delta_i \varphi - F_j([u_i, v_i])| \leq \varpi_j([u_i, v_i])$$

and it follows that

$$|f_j(x_i) \Delta_i \varphi - F_j([u_i, v_i])| \leq \varpi_1([u_i, v_i]) + \varpi_2([u_i, v_i]),$$

we have

$$\begin{aligned} (D) \sum |f_j(x_i) \Delta_i \varphi - F_j([u_i, v_i])| &\leq (D) \sum \varpi_1([u_i, v_i]) + (D) \sum \varpi_2([u_i, v_i]) \\ &\leq \varpi_1([a, b]) + \varpi_2([a, b]) \\ &\leq 2\varepsilon. \end{aligned}$$

By definition of real number A , for $\varepsilon > 0$ above there exist Perron δ -fine partition D^* on $[a, b]$ such that $A - \varepsilon < (D^*) \sum F([u_i, v_i])$. Take D^{**} be any Perron δ -fine partition on $[a, b]$ which is smoother than D^* . We then have

$$0 \leq A - (D^{**}) \sum F([u_i, v_i]) \leq A - (D^*) \sum F([u_i, v_i]).$$

Combining the above inequalities yields

$$\begin{aligned} |(D^{**}) \sum f(x_i) \Delta_i \varphi - A| &\leq |(D^{**}) \sum f(x_i) \Delta_i \varphi - F([u_i, v_i])| \\ &\quad + |(D^{**}) \sum F([u_i, v_i]) - A| \\ &< 3\varepsilon. \end{aligned}$$

This completes the proof for the case in which f is zero function. Now suppose that g is not zero function. We find that $f_1 - g, f_2 - g, h - g \in HS(\varphi, [a, b])$ and $\ominus \leq f_1 - g, f_2 - g \leq h - g$. Hence, $\max\{f_1 - g, f_2 - g\} = \max\{f_1, f_2\} - g$ is Henstock-Stieltjes integral respect to φ on $[a, b]$. This implies, $\max\{f_1, f_2\} = \max\{f_1 - g, f_2 - g\} + g$ is Henstock-Stieltjes integral respect to φ on $[a, b]$. Next, consider that $\min\{f_1, f_2\} = -\max\{-f_1, -f_2\}$. It is easy to check that $\min\{f_1, f_2\}$ is Henstock-Stieltjes integrable with respect to φ on $[a, b]$. This completes the proof.

Corollary 2.4. *Let f_1 and f_2 be real functions defined on $[a, b]$ and φ be monotonically increasing and continuous function defined on $[a, b]$. If $f_1, f_2, \max\{f_1, f_2\} \in HS(\varphi, [a, b])$ then $\min\{f_1, f_2\} \in HS(\varphi, [a, b])$.*

Proof. It is clear that $-\max\{f_1, f_2\}$ is Henstock-Stieltjes integrable with respect to φ on $[a, b]$ and $-\max\{f_1, f_2\} \leq f_1, f_2 \leq \max\{f_1, f_2\}$. And then by Theorem 2.3 we obtain $\min\{f_1, f_2\}$ is Henstock-Stieltjes integrable with respect to φ on $[a, b]$. This completes the proof.

Theorem 2.5. (Fatou's Lemma) *Let φ be increasing monotone and continuous function defined on $[a, b]$, f_n be nonnegative real functions and $f_n \in HS(\varphi, [a, b])$ for each n , and f be real function defined on $[a, b]$. If*

(i) $\{f_n\}$ converges almost everywhere to f on $[a, b]$,

(ii) $\{(HS) \int_a^b f_n d\varphi\}$ be a bounded sequence,

then

$$f \in HS(\varphi, [a, b]) \text{ and } (HS) \int_a^b f d\varphi \leq \liminf (HS) \int_a^b f_n d\varphi.$$

Proof. For each n define real function on $[a, b]$ by $f_n^* = \min\{f_i, f_{i+1}, \dots, f_{i+n}\}$. By Theorem 2.3 we have $f_n^* \in HS(\varphi, [a, b])$. Moreover, $\{-f_n^*\}$ be increasing monotone and bounded sequence. Therefore, $\{-f_n^*\}$ converges to $\sup\{-f_n^*\} = \sup_{n \geq i} \{-f_n\} = -\inf_{n \geq i}(f_n)$. Furthermore, remind that φ be increasing monotone function then

$$(HS) \int_a^b -f_n^* d\varphi \leq (HS) \int_a^b -f_{n+1}^* d\varphi \leq M$$

with M is a nonnegative real number. Hence, the sequence $\{(HS) \int_a^b -f_n^* d\varphi\}$ is increasing monotone and bounded. Therefore, the sequence $\{(HS) \int_a^b -f_n^* d\varphi\}$ converges. By the Monotone Convergence Theorem, the function $-\inf_{n \geq i}(f_n)$ is Henstock-Stieltjes integrable with respect to φ on $[a, b]$ and

$$\begin{aligned} & (HS) \int_a^b \left(-\inf_{n \geq i}(f_n) \right) d\varphi \\ &= \lim_{n \rightarrow \infty} (HS) \int_a^b (-f_n^*) d\varphi \\ &= -\lim_{n \rightarrow \infty} (HS) \int_a^b f_n^* d\varphi \\ &= -\lim_{n \rightarrow \infty} \min \left\{ (HS) \int_a^b f_i d\varphi, (HS) \int_a^b f_{i+1} d\varphi, \dots, (HS) \int_a^b f_{i+n} d\varphi \right\} \\ &\geq -\inf_{n \geq i} (HS) \int_a^b f_n d\varphi. \end{aligned}$$

We then have the sequence $\{\inf_{n \geq i}(f_n)\}$ be monotonically increasing and bounded, such that $\{\inf_{n \geq i}(f_n)\}$ converges, i.e, $\lim_{i \rightarrow \infty} \inf_{n \geq i}(f_n) = f$. Since, φ be monotonically increasing function, the sequence $\{(HS) \int_a^b \inf_{n \geq i}(f_n) d\varphi\}$ be monotonically

increasing and bounded, such that $\{(HS) \int_a^b \inf_{n \geq i} (f_n) d\varphi\}$ converges. Again, by the Monotone Convergence Theorem we have $f \in HS(\varphi, [a, b])$ and

$$(HS) \int_a^b f d\varphi = \lim_{i \rightarrow \infty} (HS) \int_a^b \inf_{n \geq i} (f_n) d\varphi \leq \liminf (HS) \int_a^b f_n d\varphi.$$

This completes the proof.

Theorem 2.6. (Dominated Convergence Theorem) *Let f_n , f , g , and h be real functions defined on $[a, b]$ for each n and φ be monotonically increasing and continuous function defined on $[a, b]$. If*

- (i) $\{f_n\}$ converges almost everywhere to f on $[a, b]$, and $f_n \in HS(\varphi, [a, b])$ for each n ,
- (ii) $g \leq f_n \leq h$ almost everywhere on $[a, b]$ for each n , with $g, h \in HS(\varphi, [a, b])$,

then

$$f \in HS(\varphi, [a, b]) \text{ and } (HS) \int_a^b f d\varphi = \lim_{n \rightarrow \infty} (HS) \int_a^b f_n d\varphi.$$

Proof. Without loss of generality, we may assume that $g \leq f_n \leq h$ on $[a, b]$ for each n . Consider that $\Theta \leq f_n - g \leq h - g$ on $[a, b]$ and $h - g, f_n - g \in HS(\varphi, [a, b])$ for each n . Clear that the sequence $\{f_n - g\}$ converges almost everywhere to $f - g$ on $[a, b]$, and for each n we have

$$(HS) \int_a^b (f_n - g) d\varphi \leq M$$

for a real number $M \geq 0$. Thus, $\{(HS) \int_a^b (f_n - g) d\varphi\}$ be bounded real numbers sequence. Recall that φ be monotonically increasing function, then by the Fatou's Lemma we obtain the function $f - g \in HS(\varphi, [a, b])$ that implies $f \in HS(\varphi, [a, b])$, and

$$\begin{aligned} & (HS) \int_a^b f d\varphi - (HS) \int_a^b g d\varphi \\ & \leq \liminf (HS) \int_a^b (f_n - g) d\varphi \\ & = \liminf (HS) \int_a^b f_n d\varphi - \limsup (HS) \int_a^b g d\varphi \\ & = \liminf (HS) \int_a^b f_n d\varphi - (HS) \int_a^b g d\varphi. \end{aligned}$$

Thus, we have

$$(HS) \int_a^b f d\varphi \leq \liminf (HS) \int_a^b f_n d\varphi. \quad (1)$$

Next, consider that $\Theta \leq h - f_n \leq h - g$ on $[a, b]$ and $h - g, h - f_n \in HS(\varphi, [a, b])$ for each n . Similarly, we will get that

$$(HS) \int_a^b f \, d\varphi \geq \limsup (HS) \int_a^b f_n \, d\varphi . \quad (2)$$

From (1) and (2) inequalities, we then have

$$\limsup (HS) \int_a^b f_n \, d\varphi \leq \liminf (HS) \int_a^b f_n \, d\varphi .$$

Now, we use the fact

$$\liminf (HS) \int_a^b f_n \, d\varphi \leq \limsup (HS) \int_a^b f_n \, d\varphi$$

to compute

$$(HS) \int_a^b f \, d\varphi = \lim_{n \rightarrow \infty} (HS) \int_a^b f_n \, d\varphi .$$

This completes the proof.

Corollary 2.7. *Let f_n and f be real functions defined on $[a, b]$ for each n and φ be monotonically increasing and continuous function defined on $[a, b]$. If*

- (i) $\{f_n\}$ converges almost everywhere to f on $[a, b]$, and $f_n \in HS(\varphi, [a, b])$ for each n ,
- (ii) $|f_n| \leq M$ almost everywhere on $[a, b]$ for each n ,

then

$$f \in HS(\varphi, [a, b]) \text{ and } (HS) \int_a^b f \, d\varphi = \lim_{n \rightarrow \infty} (HS) \int_a^b f_n \, d\varphi .$$

Proof. Using the Dominated Convergence Theorem with consider that $g(x) = -M$ and $h(x) = M$. This completes the proof.

3. CONCLUDING REMARKS

In the theory of integral, the convergence theorems give some sufficient conditions for a sequence of Henstock-Stieltjes integrable functions $\{f_n\}$ that converges to f such that f is Henstock-Stieltjes integrable and

$$(HS) \int_a^b f \, d\varphi = \lim_{n \rightarrow \infty} (HS) \int_a^b f_n \, d\varphi .$$

Some convergence theorems such as the Monotone Convergence Theorem, Fatou's Lemma, and the Dominated Convergence Theorem have succeeded in being generalized for the Henstock-Stieltjes integral.

Acknowledgement. I would like to thank to Prof. Dr. Soeparna Darmawijaya for the suggestion and encouragement so that inspiring the author in research.

REFERENCES

1. B. G. GORDON, *The Integral of Lebesgue, Denjoy, Perron and Henstock*, American Mathematical Society, USA, 1994.
2. CH. R. INDRATI, Integral Henstock-Kurzweil di Dalam Ruang Euclide Berdimensi-n, *Disertasi, Universitas Gadjah Mada, Yogyakarta, Indonesia*, 2002.
3. J. S. LIM., ET.AL., On Henstock Stieltjes Integral, *Kangweon-Kyungki. Math. Monthly*, **6**, No. 1, 87-96, 1998.
4. P. Y. LEE, *Lanzhou Lectures on Henstock Integration*, World Scientific, 1989.
5. P. Y. LEE AND R. VYBORNÝ, *Integral: An Easy Approach after Kurzweil and Henstock*, Cambridge University Press, 2000.
6. U. M. HANUNG AND S. DARMAWIJAYA, Integral Henstock-Stieltjes, *Prosiding Seminar Nasional, FMIPA Universitas Gadjah Mada, Yogyakarta, Indonesia*, 2005.
7. Z. T. GONG AND P. Y. LEE, Another Convergence Theorem for the Henstock Integral, *Sains Malaysiana*, **21** (4), 41-46, 1992.

UMI MAHNUNA HANUNG: Department of Mathematics, Gadjah Mada University, Yogyakarta, Indonesia.

E-mail: hanung_ugm@yahoo.com.

ON THE YOUNG INEQUALITY AND LOCALLY BOUNDEDNESS OF SUPERPOSITION OPERATORS ON MODULARED SPACES

SUPAMA

Abstract. In this paper, we concern about one of the very useful non-linear operators that is it so-called the superposition operator. The first main goal is to construct the Young inequality in a modulated space. Further, by using the inequality we formulate the sufficient and necessary conditions for the locally boundedness of the superposition operator on the modulated space.

Key words and Phrases: F -norm, FK -spaces, AK -spaces, Banach lattices.

1. INTRODUCTION

Function spaces and operator theory have very wide applications, from economics through engineering. Because of the reason, they recently are the central focus of the research in pure and applied analysis, and receive a lot of attention from mathematicians in the areas of modern analysis and applied mathematics.

As we have known that for any linear operator on a Banach space, the continuity and the boundedness are equivalent. It may fail for non-linear operators. Following the facts, we intend to discuss some recent achievements in the operator theory. We focus the discussion on the very useful non-linear operator that is it so-called the superposition operator.

The key references are [1],[4],[5], and [6]. Some basic and very useful concepts on linear functionals, which inspire us more, are discussed in [3]. The reference [2] gives us an idea in constructing the modulated space $W_\phi[a, \infty)$.

As usual, the real and natural numbers systems will be denoted by $\mathfrak{R}$ and $\mathcal{N}$, respectively.

Let X be a Banach lattice and (E, Σ, μ) , where $E \subset \mathfrak{R}$, a measure space. A function $s : E \rightarrow X$ is called a simple function if there exist a collection of disjoint measurable sets $E_1, E_2, \dots, E_n \subset E$ with $\cup_{k=1}^n E_k = E$ and a sequence of $a_1, a_2, \dots, a_n$ in X such that

$$s(x) = \sum_{k=1}^n a_k \chi_{A_k}(x), \quad x \in E \quad (1)$$

If s is a simple function on E , then its representation is not unique. If each a_k 's in the expression (1) are distinct and nonzero then (1) is called a canonical representation of s .

A function $f : E \rightarrow X$ is said to be measurable if there exists a sequence of simple functions $\{s_n\}$ from E into X such that

$$\lim_{n \rightarrow \infty} \|f(x) - s_n(x)\| = 0$$

almost everywhere on E .

Let $s : E \rightarrow X$ be a simple function with the canonical representation is given as (1). For any measurable set $A \subset E$, then the vector

$$\int_A s(x) d\mu(x) = \sum_{k=1}^n a_k \mu(A \cap A_k),$$

if it exists, is uniquely determined. The vector $\int_A s(x) d\mu(x)$ is then called the integral of s on A .

The measurable function $f : E \rightarrow X$ is said to be integrable on a measurable set $A \subset E$ if there exists a sequence of simple functions $\{s_n\}$ from E into X such that $\lim_{n \rightarrow \infty} \|f(x) - s_n(x)\| = 0$ almost everywhere on E and

$$\lim_{n \rightarrow \infty} \int_A \|f(x) - s_n(x)\| d\mu(x) = 0$$

Further, the integral of f on A is given by

$$\int_A f(x) d\mu(x) = \lim_{n \rightarrow \infty} \int_A s_n(x) d\mu(x) \quad (2)$$

Let (E, Σ, μ) , where $E \subset \mathfrak{R}$, be a measure space, $A \in \Sigma$, and X a Banach lattice. The collection of all measurable X -valued functions on A will be denoted by $\mathcal{M}_X(A)$. We then define

$$\begin{aligned} L_1(A) &= \{f \in \mathcal{M}_X(A) : \int_A |f(x)| d\mu < \infty\}, \text{ and} \\ \mathcal{L}_1(A) &= \{f \in \mathcal{M}_X(A) : \int_A |f(x)| d\mu = \alpha, \text{ for some } \alpha \in X\}. \end{aligned}$$

As usual, for any Banach lattice X , X^+ denotes the collection of all positive vectors in X , namely $X^+ = \{x \in X; x \geq 0\}$. If X and Y are Banach lattices, respectively, then a function $\phi : X \rightarrow Y$ is called an N -function if it holds the followings

- i. $\phi(x) = 0 \iff x = 0$,
- ii. ϕ is increasing on X^+ ,
- iii. ϕ is continuous on X , and
- iv. $\phi(-x) = \phi(x)$ for every $x \in X$.

The N -function ϕ is said to satisfy the Δ_2 -condition if there exists a real number $M > 0$ such that $\phi(2x) \leq M\phi(x)$ for every $x \in X^+$.

Let X be a Banach lattice and ϕ N -function that satisfies the Δ_2 -condition. For any real number $a > 1$, we define

$$W_\phi[a, \infty) = \{f \in \mathcal{M}_X[a, \infty) : \lim_{x \rightarrow \infty} \rho_x(f) = 0\},$$

where

$$\rho_x(f) = \frac{1}{x} \left\| \int_a^x \phi(f(t)) d\mu \right\|.$$

We can show that the functions ρ and $\|\cdot\|$,

$$\begin{aligned} \rho(f) &= \sup\{\rho_x(f) : x \in [a, \infty)\}, \\ \|f\| &= \inf\{\epsilon > 0; \rho\left(\frac{f}{\epsilon}\right) \leq \epsilon\}, \end{aligned}$$

are a modular and an F -norm on $W_\phi[a, \infty)$, respectively. Further, $W_\phi[a, \infty)$ is a solid, FK - and AK -space.

2. SUPERPOSITION OPERATORS

Let X and Y be Archimedean, σ -order complete Banach lattices, respectively and $g(\cdot, \cdot) : A \times X \rightarrow Y$ a function such that $g(\cdot, t)$ is measurable for any $t \in X$ and $g(x, 0) = 0$ for each $x \in A$. The operator $P_g : \mathcal{M}_X(A) \rightarrow \mathcal{M}_Y(A)$,

$$P_g(f)(x) = g(x, f(x)), \quad f \in \mathcal{M}_X(A), \quad x \in A$$

is called the superposition operator.

Through out this paper we always assume that the Banach lattices X and Y are Archimedean and σ -order complete, unless otherwise stated. We observe the following lemma.

Lemma 2.1. *Let $A \in \Sigma$, $\mathcal{D} \subset \mathcal{M}_X(A)$ be an FK - and AK -space and the function $g(\cdot, \cdot) : A \times X \rightarrow Y$ satisfies $g(\cdot, t)$ is measurable on A for every $t \in X$, $g(x, 0) = 0$,*

and $g(x, \cdot)$ is continuous on X for every $x \in A$. If there exist $\alpha \in Y^+$ and $\beta \in Y^+ - \{0\}$ such that the condition $\frac{1}{\mu(B)} \int_B \phi(f(x)) d\mu \leq \beta$ implies $\int_B |g(x, f(x))| d\mu \leq \alpha$ for every $f \in \mathcal{D}$ and measurable set $B \subset A$ with $\mu(B) \neq 0$, then for every measurable set $B \subset A$ there exists a non-negative function $h \in \mathcal{L}_1(B) \subset \mathcal{M}_X(B)$ with $\int_B h(x) d\mu \leq \alpha$ such that for any $x \in B$,

$$|g(x, t)| \leq h(x) + 2\alpha\beta^{-1}\mu(B)^{-1}\phi(t),$$

as $\frac{\phi(t)}{\mu(B)} \leq \beta$.

Proof. Let $B \subset A$ be a measurable set and $t \in X$. We define the function k on $A \times X$ and the function h on B as follow

$$k(x, t) = \begin{cases} |g(x, t)| - 2\alpha\beta^{-1}\mu(B)^{-1}\phi(t) & , \text{ whenever } |g(x, t)| \geq 2\alpha\beta^{-1}\mu(B)^{-1}\phi(t) \\ 0 & , \text{ otherwise} \end{cases}$$

$$h(x) = \sup\{k(x, t) : \frac{\phi(t)}{\mu(B)} \leq \beta\} = k(x, u(x))$$

It is clear that the both functions k and h depend on B . Since ϕ is continuous, then the $h(x)$ is uniquely determined. From the definition above, it is clear that the function h is non-negative on B . Further, since $g(x, \cdot)$ and ϕ are both continuous on X then $k(x, \cdot)$ is continuous on X for every $x \in B$. The continuity of ϕ then implies the set $\{t \in X : \frac{\phi(t)}{\mu(B)} \leq \beta\}$ is closed and bounded, and hence h is bounded on B . What remain to show is $h \in \mathcal{L}_1(B)$.

Let $B \subset A$ be any measurable set, then $\int_B \phi(u(x)) d\mu$ can be decomposed as

$$\int_B \phi(u(x)) d\mu = \int_{A_1} \phi(u(x)) d\mu + \int_{A_2} \phi(u(x)) d\mu + \dots + \int_{A_n} \phi(u(x)) d\mu,$$

where $A_1, A_2, \dots, A_n \subset B$ are measurable, $\cup_{i=1}^n A_i = B$, $\mu(A_i \cap A_j) = 0$, for $i \neq j$, $i, j = 1, 2, \dots, n$, and $\frac{\beta}{2} \leq \frac{1}{\mu(B)} \int_{A_i} \phi(u(x)) d\mu \leq \beta$, for $i = 1, 2, \dots, n-1$, and $0 \leq \frac{1}{\mu(B)} \int_{A_n} \phi(u(x)) d\mu \leq \beta$. Hence,

$$\begin{aligned} \int_B h(x) d\mu &= \int_B k(x, u(x)) d\mu \leq \int_B |g(x, u(x))| d\mu - 2\alpha\beta^{-1}\mu(B)^{-1} \int_B \phi(u(x)) d\mu \\ &= \sum_{i=1}^n \int_{A_i} |g(x, u(x))| d\mu - 2\alpha\beta^{-1}\mu(B)^{-1} \sum_{i=1}^n \int_{A_i} \phi(u(x)) d\mu \\ &\leq n\alpha - 2\alpha\beta^{-1}(n-1)\left(\frac{\beta}{2}\right) = \alpha. \end{aligned}$$

Thus, $h \in \mathcal{L}_1(B)$. Further, by the definition of h and k , the assertion follows.

Notice that the real numbers system $\mathfrak{R}$ is a Banach lattice. Therefore, following the Lemma 2.1 we have the corollary below.

Corollary 2.2 Let $A \subset \mathbb{R}$ be a measurable set, $\mathcal{D}$ an FK- and AK-space, and $g(.,.) : A \times X \rightarrow \mathbb{R}$ a function such that $g(.,t)$ measurable for every $t \in X$, $g(x,0) = 0$ and $g(x,.)$ is continuous on X for every $x \in A$. If there are real numbers $\alpha, \beta > 0$ such that for any $f \in \mathcal{D}$ and a measurable set $B \subset A$ with $\mu(B) \neq 0$, the condition $\frac{1}{\mu(B)} \int_B \phi(f(x)) d\mu \leq \beta$ implies $\int_B |g(x, f(x))| d\mu \leq \alpha$, then for any measurable set $B \subset A$ there exists a non negative function $h \in L_1(B)$ with $\int_B h(x) d\mu \leq \alpha$ such that for every $x \in V$,

$$|g(x, t)| \leq h(x) + 2\alpha\beta^{-1}\mu(B)^{-1}\phi(t),$$

whenever $\frac{\phi(t)}{\mu(B)} \leq \beta$.

If g satisfies a certain condition, then we can show that $P_g : W_\phi[a, \infty) \rightarrow \mathcal{L}_1[a, \infty) \subset \mathcal{M}_Y[a, \infty)$.

Theorem 2.3 Let $g(.,.) : [a, \infty) \times X \rightarrow Y$ be a function such that $g(.,t)$ is measurable for any $t \in X$, $g(x,0) = 0$, dan $g(x,.)$ is continuous on X for every $x \in [a, \infty)$. If there exist $\alpha \in Y^+$, $\beta \in Y^+ - \{0\}$, and for any $s \in [a, \infty)$ there exists a non-negative function $h \in \mathcal{L}_1[a, s]$ with $\int_a^s h(x) d\mu \leq \alpha$ such that for any $x \in [a, s]$,

$$|g(x, t)| \leq h(x) + 2\alpha\beta^{-1}s^{-1}\phi(t),$$

as $\frac{\phi(t)}{s} \leq \beta$, then the superposition operator P_g maps $W_\phi[a, \infty)$ into $\mathcal{L}_1[a, \infty)$.

Proof. Let $f \in W_\phi[a, \infty)$, then there exists a real number $M > 0$ such that for every $s \geq M$,

$$\frac{1}{s} \int_a^s \phi(f(x)) d\mu < \beta.$$

Following the hypothesis, for every $s \geq M$ we have

$$\begin{aligned} \int_a^s |g(x, f(x))| d\mu &\leq \int_a^s h(x) d\mu + 2\alpha\beta^{-1}s^{-1} \int_a^s \phi(f(x)) d\mu \\ &\leq \alpha + 2\alpha\beta^{-1}\beta = 3\alpha. \end{aligned}$$

Since it holds for every $s \geq M$ then $\int_a^\infty |g(x, f(x))| d\mu$ exists, or $g(., f(.)) \in \mathcal{L}_1[a, \infty)$. Thus, $P_g : W_\phi[a, \infty) \rightarrow \mathcal{L}_1[a, \infty) \subset \mathcal{M}_Y[a, \infty)$.

The Theorem 2.3 states the sufficient condition that the superposition operator P_g maps the space $W_\phi[a, \infty)$ onto $\mathcal{L}_1[a, \infty)$. It is very difficult to prove the converse of the theorem. However, if the function g is real valued, i.e. $g(.,.) : [a, \infty) \times L \rightarrow \mathbb{R}$, we can construct the sufficient and necessary conditions that P_g maps $W_{\phi, \mathbb{R}}[a, \infty)$ onto $L_1[a, \infty)$, where

$$W_{\phi, \mathbb{R}}[a, \infty) = \{f \in \mathcal{M}_{\mathbb{R}}[a, \infty) : \lim_{x \rightarrow \infty} \rho_x(f) = 0\}.$$

Since ϕ is an even function, then we can write

$$\rho_x(f) = \frac{1}{x} \int_a^x \phi(f(t)) d\mu$$

Theorem 2.4 *Let $g(.,.) : [a, \infty) \times X \rightarrow \mathfrak{R}$ be a function such that $g(., t)$ is measurable for every $t \in X$, $g(x, 0) = 0$, and $g(x, .)$ is continuous on X for every $x \in [a, \infty)$. The superposition operator P_g maps the space $W_{\phi, \mathfrak{R}}[a, \infty)$ onto $L_1[a, \infty)$ if and only if there exists real numbers $\alpha, \beta > 0$ and for each $s \in [a, \infty)$ there exists a non negative function $h \in L_1[a, s]$ satisfies $\int_a^s h(x) d\mu \leq \alpha$ such that for every $x \in [a, s]$,*

$$|g(x, t)| \leq h(x) + 2\alpha\beta^{-1}s^{-1}\phi(t),$$

whenever $\frac{\phi(t)}{s} \leq \beta$.

Proof. The sufficiency follows from the Theorem 2.3 by taking $Y = \mathfrak{R}$.

For the necessity, the functional $\mathcal{F} : W_{\phi, \mathfrak{R}}[a, \infty) \rightarrow \mathfrak{R}$,

$$\mathcal{F}(f) = \int_a^\infty |g(x, f(x))| d\mu,$$

is continuous and orthogonally additive (See the reference [6]). So, there exists a real number $\eta > 0$ such that for every $f \in W_{\phi, \mathfrak{R}}[a, \infty)$ with $\|f\| \leq \eta$ we have

$$\int_a^\infty |g(x, f(x))| d\mu < 1.$$

It is easy to prove that there exists a real number $\beta > 0$ such that the condition $\rho(f) < \beta$ implies $\|f\| < \eta$. Thus, there exists $\beta > 0$ such that for every $f \in W_{\phi, \mathfrak{R}}[a, \infty)$,

$$\int_a^\infty |g(x, f(x))| d\mu < 1,$$

whenever $\rho(f) < \beta$. These implies for every $s \in [a, \infty)$,

$$\int_a^s |g(x, f(x))| d\mu < 1,$$

whenever $\frac{1}{s} \int_a^s \phi(f(x)) d\mu \leq \beta$.

Further, if we take $\alpha = \sup\{\int_a^s |g(x, f(x))| d\mu : \frac{1}{s} \int_a^s \phi(f(x)) d\mu \leq \beta\}$, then Corollary 2.2 implies the existence of the non negative function $h \in L_1[a, s]$ satisfies $\int_a^s h(x) d\mu \leq \alpha$, such that for any $x \in [a, s]$,

$$|g(x, t)| \leq h(x) + 2\alpha\beta^{-1}s^{-1}\phi(t),$$

as $\frac{\phi(t)}{s} \leq \beta$.

Following the Theorem 2.4, for the sequel we always assume that the generator of the superposition operator P_g is the function $g(.,.) : [a, \infty) \times X \rightarrow \mathfrak{R}$, unless otherwise stated.

3. BOUNDEDNESS OF SUPERPOSITION OPERATORS ON

$$W_{\phi, \mathfrak{R}}[a, \infty)$$

Let $E \subset \mathcal{M}_X[a, \infty)$ be a modular spaces with the modular ρ and $g(., .) : [a, \infty) \times X \rightarrow \mathfrak{R}$ a function such that $g(., t)$ is measurable for every $t \in X$ and $g(x, 0) = 0$ for every $x \in [a, \infty)$. The superposition operator $P_g : E \rightarrow L_1[a, \infty)$ is said to be *locally modular bounded* at $f \in E$ if there exist real numbers $\alpha, \beta, \gamma > 0$ such that for every $h \in E$ with $\rho(\gamma(f - h)) \leq \alpha$ we have

$$\|P_g(h)\| \leq \beta.$$

Operator P_g is said to be *locally bounded* at $f \in E$ if there exist real numbers $\alpha, \beta > 0$ such that for any $h \in E$ with $\|f - h\| \leq \alpha$ we have

$$\|P_g(h)\| \leq \beta.$$

Further, operator P_g is said to be *locally modular bounded* (*locally bounded*) if P_g is locally modular bounded (locally bounded) at every $f \in E$. It is easy to prove that the superposition operator $P_g : E \rightarrow L_1[a, \infty)$ is locally modular bounded iff it is locally bounded.

Theorem 3.1 *Let $g(., .) : [a, \infty) \times X \rightarrow \mathfrak{R}$ be a function such that $g(., t)$ is measurable for every $t \in X$, $g(x, 0) = 0$ for every $x \in [a, \infty)$, and for every $s \in [a, \infty)$ and $n \in \mathcal{N}$ there exist non negative functions $f_1, h \in L_1[a, s]$ and there is a $\delta > 0$ such that for any $x \in [a, s]$,*

$$|g(x, t)| \leq h(x) + x^{-1} f_1(x) \phi\left(\frac{1}{n}t\right),$$

as $\frac{1}{x} \phi\left(\frac{1}{n}t\right) \leq \delta$. The superposition operator $P_g : W_{\phi, \mathfrak{R}}[a, \infty) \rightarrow L_1[a, \infty)$ is locally modular bounded iff for every $x \in [a, \infty)$, the function $g(x, .)$ is bounded on any closed and bounded subset in X .

Proof. For the sufficient, take any $x \in [a, \infty)$ and $t \in L$. If $E \subset [a, \infty)$ is any measurable set with $\mu(E) < \infty$ then $t\chi_E \in W_{\phi, \mathfrak{R}}[a, \infty)$. By the hypothesis there exists real numbers $\alpha_1, \beta, \gamma > 0$ such that for every $k \in L$, the condition $\rho(\gamma(t - k)\chi_E) \leq \alpha_1$ implies

$$\|P_g(k\chi_E)\| \leq \beta.$$

Further, since ϕ is continuous on X , then there exists a real number $\alpha > 0$ such that

$$\rho(\gamma(t - k)\chi_E) \leq \alpha_1,$$

whenever $\|t - k\| \leq \alpha$. Thus, there exists a real number $\alpha, \beta > 0$ such that for any $x \in [a, \infty)$,

$$|g(x, k)| \leq \|P_g(k\chi_E)\| \leq \beta,$$

as $\|t - k\| \leq \alpha$. It means $g(x, .)$ is locally bounded at t . Further, if $A \subset X$ is closed and bounded then by the Heine-Borel Theorem, $g(x, .)$ is closed on A .

For the necessity, let $s \in [a, \infty)$ and $n \in \mathcal{N}$ be arbitrarily. By the hypothesis there exists a non negative functions $f_1, h \in L_1[a, s]$ and a real number $\delta > 0$ such that for any $x \in [a, s]$,

$$|g(x, t)| \leq h(x) + x^{-1}f_1(x)\phi\left(\frac{1}{n}t\right),$$

whenever $\frac{1}{x}\phi\left(\frac{1}{n}t\right) \leq \delta$. Further, suppose $f \in W_{\phi, \mathfrak{R}}[a, \infty)$, then there is a real number $M > a$ such that for every $x \geq M$ we have

$$\rho_x\left(\frac{1}{n}f\right) = \frac{1}{x} \int_a^x \phi(f(s))d\mu \leq \frac{\delta}{4}.$$

However, from the definition of the function ϕ , we can choose $k \in W_{\phi, \mathfrak{R}}[a, \infty)$ such that

$$\frac{1}{x}\phi\left(\frac{1}{n_0}k\right) < \delta \quad \text{and} \quad \rho\left(\frac{1}{n_0}(f - k)\right) < \frac{\delta}{4},$$

for some $n_0 \in \mathcal{N}$. For any $x < M$, let

$$m(x) = \sup\{|g(x, t)| : \phi\left(\frac{1}{n}(t - f(x))\right) \leq \frac{\delta}{4}\}.$$

Since ϕ is continuous on X then the set

$$A = \{t : \phi\left(\frac{1}{n}(t - f(x))\right) \leq \frac{\delta}{4}\}$$

is closed and bounded in X . Therefore, by the hypothesis $m(x) < \infty$ for any $x < M$. Thus, for every $s > M$ and $k \in W_{\phi, \mathfrak{R}}[a, \infty)$ with $\rho\left(\frac{1}{n_0}(f - k)\right) < \frac{\delta}{4}$ we have

$$\begin{aligned} \int_a^s |g(x, k(x))|d\mu &= \int_a^M |g(x, k(x))|d\mu + \int_M^s |g(x, k(x))|d\mu \\ &\leq \int_a^M m(x)d\mu + \int_M^s \{h(x) + x^{-1}f_1(x)\phi\left(\frac{1}{n_0}k(x)\right)\}d\mu \\ &\leq K + \|h\| + \frac{\delta}{2}\|f_1\|, \quad \text{for some } K > 0. \end{aligned}$$

If we take $s \rightarrow \infty$ then

$$\|P_g(k)\| \leq K + \|h\| + \frac{\delta}{2}\|f_1\|.$$

Finally, by letting $\alpha = \frac{\delta}{4}$, $\beta = K + \|h\| + \frac{\delta}{2}\|f_1\|$, and $\gamma = \frac{1}{n_0}$ the assertion follows.

Corollary 3.2 *Let $g(., .) : [a, \infty) \times X \rightarrow \mathfrak{R}$ be a function as in Theorem . The superposition operator $P_g : W_{\phi, \mathfrak{R}}[a, \infty) \rightarrow L_1[a, \infty)$ is locally bounded iff for every $x \in [a, \infty)$, the function $g(x, .)$ is bounded on any closed and bounded subset in X .*

Let $E \subset m_X[a, \infty)$ be a modular space with the modular ρ and $g(., .) : [a, \infty) \times X \rightarrow \mathfrak{R}$ a function such that $g(., t)$ is measurable for every $t \in X$ and $g(x, 0) = 0$ for every $x \in [a, \infty)$. The superposition operator $P_g : E \rightarrow L_1[a, \infty)$ is said to be **modular bounded** on E if for every $f \in E$ and for any real constants $\alpha, \beta > 0$ there exists $\delta > 0$ such that for every $h \in E$,

$$\|P_g(h)\| \leq \delta,$$

whenever $\rho(\alpha(f - h)) \leq \beta$. The operator P_g is said to be (*bounded*) on X if for every real number $\alpha > 0$ there exists a $\delta > 0$ such that for every $f \in X$ with $\|f\| \leq \alpha$, we have

$$\|P_g(f)\| \leq \delta$$

We can easily prove the following theorems.

Theorem 3.3 *Let $f \in W_{\phi, \mathfrak{R}}[a, \infty)$ and $\alpha, \beta > 0$, then there exists real numbers $\alpha_1, \beta_1 > 0$ such that for every $h \in W_{\phi, \mathfrak{R}}[a, \infty)$ with $\rho(\alpha(f - h)) \leq \beta$ we have $\rho(\alpha_1 h) \leq \beta_1$.*

Theorem 3.4 *The superposition operator $P_g : W_{\phi, \mathfrak{R}}[a, \infty) \rightarrow L_1[a, \infty)$ is modular bounded iff for every real numbers $\alpha, \gamma > 0$ there exists $\beta > 0$ such that for any $h \in W_{\phi, \mathfrak{R}}[a, \infty)$ with $\rho(\gamma h) \leq \alpha$ we have $\|P_g(h)\| \leq \beta$.*

Theorem 3.5 *The superposition operator $P_g : W_{\phi, \mathfrak{R}}[a, \infty) \rightarrow L_1[a, \infty)$ is modular bounded iff it is bounded.*

Theorem 3.6 *Let the function $g(., .) : [a, \infty) \times X \rightarrow \mathfrak{R}$ satisfies the following conditions: $g(., t)$ is measurable for every $t \in X$, $g(x, 0) = 0$ for every $x \in [a, \infty)$, and for every real number $\beta > 0$ and $s \in [a, \infty)$ there exist non negative functions $f, h \in L_1[a, s]$ such that for every $x \in [a, s]$*

$$|g(x, t)| \leq h(x) + \frac{1}{x}f(x)\phi(t),$$

whenever $\frac{\phi(t)}{x} \leq \beta$. If $g(x, .)$ is bounded on X then $P_g : W_{\phi, \mathfrak{R}}[a, \infty) \rightarrow L_1[a, \infty)$ is modular bounded.

Proof. Take any $s \in [a, \infty)$ and any real number $\beta > 0$. By the hypothesis there exist non negative functions $f, h \in L_1[a, s]$ such that for every $x \in [a, s]$,

$$|g(x, t)| \leq h(x) + \frac{1}{x}f(x)\phi(t),$$

whenever $\frac{\phi(t)}{x} \leq \beta$.

Let $k \in W_{\phi, \mathfrak{R}}[a, \infty)$ be an arbitrarily, then there exists an $M > a$ such that for every $x \geq M$

$$\rho_x(f) = \frac{1}{x} \int_a^x \phi(k(\zeta))d\mu \leq \beta.$$

For any $x < M$, we define

$$m(x) = \sup\{|g(x, t)| : t \in X\}.$$

Since $g(x, \cdot)$ is bounded on X , then $m(x) < \infty$. Further, if $\gamma > 0$ then the condition $\rho(\gamma k) \leq \beta$ implies

$$\begin{aligned} \int_a^s |g(\zeta, \gamma k(\zeta))| d\mu &= \int_a^M |g(\zeta, \gamma k(\zeta))| d\mu + \int_M^s |g(\zeta, \gamma k(\zeta))| d\mu \\ &\leq \int_a^M m(\zeta) d\mu + \int_M^s \left\{ h(\zeta) + \frac{1}{\zeta} f(\zeta) \phi(\gamma k(\zeta)) \right\} d\mu \\ &\leq K + \|h\| + \beta \|f\|, \quad \text{for some } K > 0. \end{aligned}$$

Since, it is hold for all $s \in [a, \infty)$ then

$$\int_a^\infty s |g(\zeta, \gamma k(\zeta))| d\mu \leq K + \|h\| + \beta \|f\|, \quad \text{untuk suatu } K > 0.$$

Further, if we take $\delta = K + \|h\| + \beta \|f\|$, then

$$\|P_g(\gamma k)\| \leq \delta.$$

By Theorem 3.4, P_g is modular bounded.

Corollary 3.7 *Let $g(\cdot, \cdot) : [a, \infty) \times X \rightarrow \mathfrak{R}$ be a function as in the Theorem. If $g(x, \cdot)$ is bounded on X then $P_g : W_{\phi, \mathfrak{R}}[a, \infty) \rightarrow L_1[a, \infty)$ is bounded.*

3. CONCLUDING REMARKS

We have formulated and proved the Young inequality. Based on the inequality, we constructed a sufficient and necessary conditions so that the superposition operator P_g maps $W_{\phi, \mathfrak{R}}[a, \infty)$ into $L_1[a, \infty)$. We also proved that the superposition operator $P_g : W_{\phi, \mathfrak{R}}[a, \infty) \rightarrow L_1[a, \infty)$ is locally bounded iff for every $x \in [a, \infty)$, the function $g(x, \cdot)$ is bounded on any closed and bounded subset in X .

REFERENCES

1. PLUCIENNIK, R., Boundedness of Superposition Operators on w_0 and W_0 , *Institute of Mathematics Technical University, Poznan, Poland*, 1991.
2. MEYER, P. AND NIEBERG, *Banach Lattice*, Springer-Verlag, New-York, 1991.

3. ORLICZ, W, *Linear Functional Analysis*, World Scientific, 1992.
4. SRI DARU UNONINGSIH, RYSZARD PLUCIENNIK AND LEE PENG YEE, Boundedness of Superposition Operators on Sequence Spaces, *Commentationes Mathematicae*, 209 - 216, 1995.
5. SUPAMA AND SOEPARNA DARMAWIJAYA, Superposition Operators on $W_0(\phi)$, *Bulletin of Malaysian Mathematical Society*, 2004.
6. SUPAMA AND SOEPARNA DARMAWIJAYA, A Representation Theorem for Orthogonally Additive Operators on Modular Spaces of Banach Lattice Valued Functions, *Journal of Indonesian Mathematical Society*, 2004.

SUPAMA: Department of Mathematics, Gadjah Mada University, Yogyakarta, Indonesia.
E-mail: supama@ugm.ac.id.

PART III

Applied Mathematics

A MATHEMATICAL SCENARIO OF AN OPTIMUM AND SUSTAINABLE HARVESTING OF TEAK WOOD

ARMAN AND EDI CAHYONO

Abstract. Harvesting of wood from the forest without carefully planned, more often than not, results in to over harvesting. Tropical forest in the world, mainly in the developing countries, is shrinking due to over harvesting. On the other hand, the wood as a natural resource of the forest is needed to support the economy, especially for those developing countries. This also happens for Indonesia. Hence, an optimum and sustainable harvesting should be developed. Such harvesting could produce optimal wood but it still maintains a desirable condition of the forest. In this paper we discuss a scenario of an optimum and sustainable harvesting of teak wood from mathematical point of view. Let a homogeneous forest be harvested once in an interval of time Δt , e.g. five years. The idea of an optimum and sustainable harvesting is that the forest condition after being harvested at the time t should be similar to the condition at the previous time $t - \Delta t$. This initial condition should be defined by considering several aspects, mainly from environmental point of view. We also present an optimum and sustainable harvesting for the case of teak wood (*Tectona grandis*) which grow in South East Sulawesi (Sultra).

Key words and Phrases : optimum, sustainable, teak wood

1. INTRODUCTION

Teak wood represent one of forestry commodity having high price. This teak wood serve the purpose of industrial raw material. system of management of production forest have to get the governmental attention as party in charge. This matter is intended in order that existing production forest can yield the optimal wood, but forest continuity and its function remain to be awaked. One of internal issue of management of teak wood is how model of harvesting to be getting optimum wood production and have continuation, (A.Saputra and Arman, 2005).

In order that forest observation can be done in control and fulfill the everlasting forest management principle. Hence we require to do the forest management giving a high economic value through understanding to its growth model, (Rosmarlinasiah, 2005).

We shall introduce a mathematical scenario for the sustainable harvesting of teak wood whose trees are classified by height. The height of a tree is assumed to determine its economic value when it is cut down and sold. Initially, there is a distribution of trees of various heights. The forest is then allow to grow for a certain period of time, after which some of the trees of various heights are harvested. But the trees left un-harvested are to be of the same height configurations as the original forest, so that the harvest is sustainable. We want to find one for which the total economic value of all the trees removed is as large

as possible. This determines the optimal sustainable yield of the forest, and is the largest yield which can be attained continually without depleting the forest.

2. MATHEMATICAL MODELLING

We shall see the harvesting of teak wood from mathematical point of view. We know that, the trees of different heights have different economic values. Suppose there are n different price groups corresponding to certain height intervals as follows in Table 1.

Table 1. Economic value of trees based on height intervals

Group	Economic Value	Height Interval
1 (Seedlings)	none	$[0, h_1)$
2	v_2	$[h_1, h_2)$
3	v_3	$[h_2, h_3)$
$\vdots$	$\vdots$	$\vdots$
$m-1$	v_{m-1}	$[h_{m-2}, h_{m-1})$
M	v_m	$[h_{m-1}, h_\infty)$

The first group consists of seedlings with heights in the interval $[0, h_1)$, and these seedlings are of no economic value. The second group consists of trees with height in the interval $[h_1, h_2)$. The m -th group consists of trees with height greater than or equal to h_{m-1} .

Let x_i ($i = 1, 2, \dots, m$) be the number of trees within the i -th group which remain after each harvest. We form a column vector with these numbers

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \quad (1)$$

And call it the *non harvest vector*. For a sustainable harvesting policy, the forest is to be returned after each harvest to the fixed configuration given by the non harvest vector x .

Since the total number of trees in the forest is fixed, let us set

$$x_1 + x_2 + \dots + x_n = s \quad (2)$$

where s is predetermined by the amount of land available and the amount of space each tree requires. Referring to Figure 1, we have the following situation: The forest configuration is given by the vector x after each harvest. Between harvests, the trees grow and produce the new forest configuration before each harvest. A certain number of trees are removed from each group at the harvest. Finally, a seedling is planted in place of each tree removed to return the forest again to the configuration x .

Let us first consider the growth of the forest between harvests. During this period, a tree in the i -th group may grow and move up to a higher height group. Or its growth may be retarded for some reason, and it will remain in the same group. We consequently define the following growth parameters, w_i for $i = 1, 2, \dots, m-1$.

w_i = the fraction of trees in the i -th group that grow into the $(i+1)$ -st group during a growth period.

For simplicity, we shall assume that a tree can move at most one height group upward in one growth period. With this assumption we have

$1 - w_i$ = the fraction of trees in the i -th group that remain in the i -th group during a growth period.

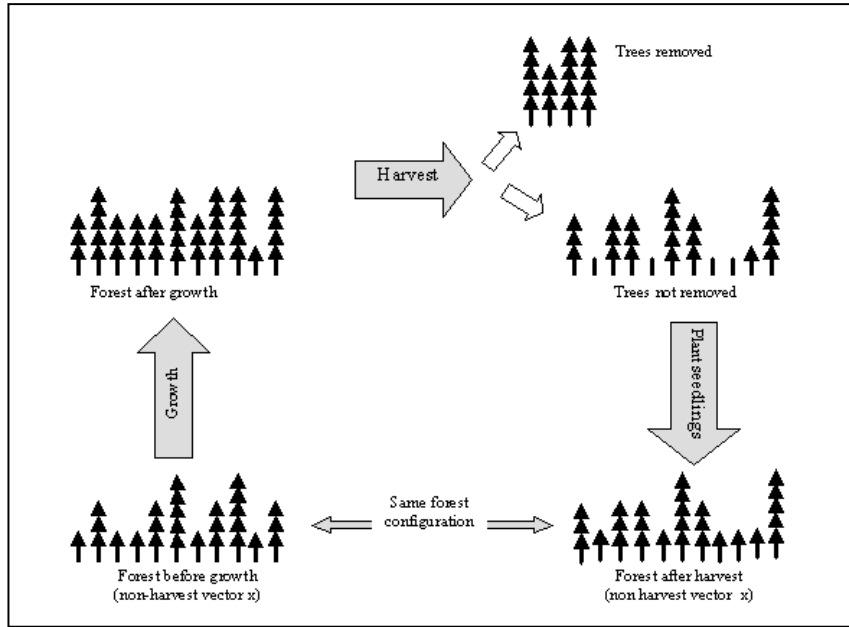


Figure 1. Scenario of harvesting of teak wood

With these $m-1$ growth parameters, we form the following $m \times m$ growth matrix:

$$G = \begin{bmatrix} 1-w_1 & 0 & 0 & \cdots & 0 \\ w_1 & 1-w_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1-w_{m-1} & 0 \\ 0 & 0 & 0 & \cdots & w_{m-1} & 1 \end{bmatrix} \quad (3)$$

Since the entries of the vector x are the numbers of trees in the m groups before the growth period.

Suppose during the harvest we remove y_i ($i = 1, 2, \dots, m$) trees from the i -th group. We shall call the column vector y is the *harvest vector*.

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

Thus, a total of $y_1 + y_2 + \dots + y_m$ trees are removed at each harvest. This is also the total number of trees added to the first group (the new seedlings) after each harvest. If we define the following $m \times m$ replacement matrix

$$R = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad (4)$$

then the column vector

$$Ry = \begin{bmatrix} y_1 + y_2 + \dots + y_m \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (5)$$

specifies the configuration of trees planted after each harvest.

At this point, we are ready to write the following equation which characterizes a sustainable harvesting policy:

$$\begin{bmatrix} \text{configuration} \\ \text{at end of} \\ \text{growth period} \end{bmatrix} - [\text{harvest}] + \begin{bmatrix} \text{new seedling} \\ \text{replacement} \end{bmatrix} = \begin{bmatrix} \text{configuration} \\ \text{at beginning of} \\ \text{growth period} \end{bmatrix}$$

or mathematically

$$Gx - y + Ry = x$$

This equation can be rewritten as

$$(I - R)y = (G - I)x \quad (6)$$

We shall refer to Equation (6) as the sustainable harvesting condition. Any vector x and y with nonnegative entries, and such that $x_1 + x_2 + \dots + x_m = s$, which satisfy this matrix equation determine a sustainable harvesting policy for the forest. Let us note that if $y_1 > 0$, then the harvester is removing seedlings, of no economic value, and then replacing them with new seedlings. Since there is no point in doing this, we shall assume

$$y_1 = 0 \quad (7)$$

With this assumption, it may be verified that (6) is the matrix form of the following set of equations:

$$\begin{aligned}
 y_2 + y_3 + \cdots + y_{m-1} + y_m &= w_1 x_1 \geq 0 \\
 y_2 &= w_1 x_1 - w_2 x_2 \geq 0 \\
 y_3 &= w_2 x_2 - w_3 x_3 \geq 0 \\
 &\vdots \\
 y_{m-1} &= w_{m-2} x_{m-2} - w_{m-1} x_{m-1} \geq 0 \\
 y_m &= w_{m-1} x_{m-1} \geq 0
 \end{aligned} \tag{8}$$

Notice that the first of Equation (8) is the sum of the remaining $m-1$ equations.

Since we must have $y_i \geq 0$ for $i = 2, 3, \dots, m$, Equation 8 require that

$$w_1 x_1 \geq w_2 x_2 \geq \cdots \geq w_{m-1} x_{m-1} \geq 0 \tag{9}$$

Conversely, if x is a column vector with nonnegative entries which satisfy Equation (9), then Equation (7) and (8) define a column vector y with nonnegative entries. Furthermore, x and y then satisfy the sustainable harvesting condition (6). In other words, a necessary and sufficient condition that a nonnegative column vector x determine a forest configuration which is capable of sustainable harvesting is that its entries satisfy (9).

4. OPTIMUM AND SUSTAINABLE HARVESTING

In this section, we shall find the optimal and sustainable harvesting of teak wood. Since we remove y_i trees from the i -th group ($i = 1, 2, \dots, m$) and since each tree in the i -th group has an economic value of v_i , the total yield of the harvest, H , is given by

$$H = v_2 y_2 + v_3 y_3 + \cdots + v_m y_m \tag{10}$$

Using (8), we may substitute for the y_i 's in (10) to obtain

$$H = v_2 (w_1 x_1 - w_2 x_2) + v_3 (w_2 x_2 - w_3 x_3) + \cdots + v_m (w_{m-1} x_{m-1}) \tag{11}$$

Combining Equation (11), (1), and (9), we may now state the problem of maximizing the yields of the forest over all possible sustainable harvesting policies.

Let us first set, H_k is yield obtained by harvesting all of the k -th group and none of the other groups.

The largest value of H_k for $k = 2, 3, \dots, m$ will then be the optimal sustainable yield, and the corresponding value of k will be the group which should be completely harvested to attain the optimal sustainable yield. Since no group but the k -th is harvested, we have

$$y_1 = y_2 = \cdots = y_{k-1} = y_{k+1} = \cdots = y_m = 0 \tag{12}$$

In addition, since all of the k -th group is harvested, no trees are left un-harvested in the k -th group, and no trees are ever present in the height classes above the k -th group. Thus,

$$x_k = x_{k+1} = \cdots = x_m = 0 \tag{13}$$

Substitution of (12) and (13) into the sustainable harvesting condition (8) then gives

$$\begin{aligned}
 y_k &= w_1 x_1 \\
 0 &= w_1 x_1 - w_2 x_2 \\
 0 &= w_2 x_2 - w_3 x_3 \\
 &\vdots \\
 0 &= w_{k-2} x_{k-2} - w_{k-1} x_{k-1} \\
 y_k &= w_{k-1} x_{k-1}
 \end{aligned} \tag{14}$$

Equation (14) may also be written as

$$y_k = w_1 x_1 = w_2 x_2 = \dots = w_{k-1} x_{k-1} \quad (15)$$

from which it follows that

$$\begin{aligned} x_2 &= w_1 x_1 / w_2 \\ x_3 &= w_1 x_1 / w_3 \\ &\vdots \\ x_{k-1} &= w_1 x_1 / w_{k-1} \end{aligned} \quad (16)$$

If we substitute equation (13) and (16) into (1),

$$x_1 + x_2 + \dots + x_m = s$$

we may solve for x_1 and obtain

$$x_1 = \frac{s}{1 + \frac{w_1}{w_2} + \dots + \frac{w_1}{w_{k-1}}} \quad (17)$$

For the yield H_k , we combine (10), (12), (15), and (17) to obtain

$$H_k = v_2 y_2 + v_3 y_3 + \dots + v_{k-1} y_{k-1} + v_k y_k + v_{k+1} y_{k+1} + \dots + v_m y_m$$

$$H_k = v_k y_k$$

$$H_k = v_k w_1 x_1$$

$$H_k = v_k w_1 \frac{s}{1 + \frac{w_1}{w_2} + \dots + \frac{w_1}{w_{k-1}}}$$

$$H_k = w_1 \frac{v_k s}{w_1 \left(\frac{1}{w_1} + \frac{1}{w_2} + \dots + \frac{1}{w_{k-1}} \right)}$$

$$H_k = \frac{v_k s}{\left(\frac{1}{w_1} + \frac{1}{w_2} + \dots + \frac{1}{w_{k-1}} \right)} \quad (18)$$

Equation (18) determines H_k in terms of the known growth and economic parameters for any $k = 2, 3, \dots, m$. Thus, the optimal sustainable yield is found as follows:

Theorem 4.1. (C. Rorres, 1984). *The optimal and sustainable yield is the largest value of*

$$H_k = \frac{v_k s}{\left(\frac{1}{w_1} + \frac{1}{w_2} + \dots + \frac{1}{w_{k-1}} \right)}$$

for $k = 2, 3, \dots, m$. The corresponding value of k is the number of the group which is completely harvested.

Based on Theorem 4.1, we can find non harvest vector x corresponding

to the optimal and sustainable harvesting policy.

From the non harvest vector x

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_{k-1} \\ x_k \\ x_{k+1} \\ \vdots \\ x_m \end{bmatrix}$$

By using Equation (13), we obtain

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_{k-1} \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (19)$$

If we substitute Equation (16) into (19), we obtain

$$x = \begin{bmatrix} x_1 \\ w_1 x_1 / w_2 \\ \vdots \\ w_1 x_1 / w_{k-1} \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = w_1 x_1 \begin{bmatrix} 1/w_1 \\ 1/w_2 \\ \vdots \\ 1/w_{k-1} \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (20)$$

We can substitute Equation (17) into (20),

$$x = \frac{s}{\left(\frac{1}{w_1} + \frac{1}{w_2} + \dots + \frac{1}{w_{k-1}} \right)} \begin{bmatrix} 1/w_1 \\ 1/w_2 \\ \vdots \\ 1/w_{k-1} \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (21)$$

5. THE CASE OF TEAK WOOD (*Tectona grandis*) WHICH GROW IN SOUTH EAST SULAWESI

From the result of the research of Rosmarlinasiah (1996) conducted in HTI-Swakelola area, Konawe Selatan Regency, South east Sulawesi with the perception amount as much 30 trees. The age group from that tree is 2 until 6 year.

Every tree highly and different diameter can own the different economic value. Hence, the teak wood can be classified become some group of height. With the assumption that only tree with the height is greater than or equal to 6 meter having economic value to be harvested. We can be classified by 5 height group with the economics value owned by each its group, as Table 2.

Table 2. Classification of teak wood based on height interval

Group	Economic Value	Height Interval	Number of Tree
1	None	[0,6)	9
2	v_2	[6,8)	7
3	v_3	[8,10)	6
4	v_4	[10,12)	6
5	v_5	[12, ∞)	2

The growth parameters w_i ($i = 1,2,3,4$) can be calculated by

$$w_i = \frac{a_{i+1}}{a_i}, \quad 0 \leq w_i \leq 1 \quad (22)$$

where a_i = number of trees in the i -th group ($i = 1,2,3,4$).

Hence, we obtain the growth matrix

$$G = \begin{bmatrix} 0,23 & 0 & 0 & 0 & 0 \\ 0,77 & 0,15 & 0 & 0 & 0 \\ 0 & 0,85 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0,67 & 0 \\ 0 & 0 & 0 & 0,33 & 1 \end{bmatrix}$$

Using Equation (18), we can calculate the optimum and sustainable harvesting of teak wood as follows:

$$H_2 = \frac{v_2 s}{\frac{1}{g_1} + \frac{1}{0,77}} = \frac{30}{\frac{1}{0,77}} v_2 = 23,09 v_2$$

$$H_3 = \frac{v_3 s}{\frac{1}{g_1} + \frac{1}{g_2}} = \frac{30}{\frac{1}{0,77} + \frac{1}{0,85}} v_3 = 12,12 v_3$$

$$H_4 = \frac{v_4 s}{\frac{1}{g_1} + \frac{1}{g_2} + \frac{1}{g_3}} = \frac{30}{\frac{1}{0,77} + \frac{1}{0,85} + \frac{1}{1}} v_4 = 8,63 v_4$$

$$H_5 = \frac{v_5 s}{\frac{1}{g_1} + \frac{1}{g_2} + \frac{1}{g_3} + \frac{1}{g_4}} = \frac{30}{\frac{1}{0,77} + \frac{1}{0,85} + \frac{1}{1} + \frac{1}{0,33}} v_5 = 4,61 v_5$$

where v_i is the economic value of teak wood in the i -th group ($i = 2, 3, 4, 5$).

6. CONCLUDING REMARKS

The optimal and sustainable yields of a periodic harvest is calculated when the trees of different height classes can have different economic values.

The trees left un-harvested are to be of the same height configuration as the original forest, so that the harvest is sustainable.

The total economic value of all trees removed is as largest possible. This determines the optimal sustainable yield of the forest, and is the largest yield which can be attained continually without depleting the forest.

REFERENCES

1. A. Saputra, A and Arman,. *Aplikasi Model Matematika dalam Pengelolaan Hutan untuk Mencapai Hasil Optimal pada Setiap Panen*, Jurusan Matematika FMIPA Universitas Haluoleo, 2005.
2. C. Jeffries, *Mathematical Modeling in Ecology*, a workbook for students, Birkhauser, Boston-Basel, Berlin, 1989.
3. C. Rorres, *Application of Linear Algebra*, John Wiley & Sons, Inc, 1984.
4. Rosmarlinasiah, Prosiding Semiloka Kehutanan, *Forum kerjasama Delapan Perguruan Tinggi, Dirjen Dikti, Universitas Haluoleo*, 2005.
5. Rosmarlinasiah, Studi Biometrik Hutan Tanaman Jati (*Tectona grandis L.f*), *Laporan Hasil Penelitian, Proyek Pengembangan Sebelas Lembaga Pendidikan Tinggi (ADB-Loan No. 1253-INO), Lembaga Penelitian Universitas Haluoleo*, 1996.

ARMAN and EDI CAHYONO: Department of Mathematics, FMIPA-Universitas Haluoleo.
Email: mukhsar@math-unhalu.org

INVERSE PROBLEM of LAPLACE EQUATION in PLANE DOMAIN CONTAINING a CRACK

WONO SETYA BUDHI, MARWAN WIRIANTO

Abstract. In the last our paper [2], we model the direct problem of Laplace equation and propose the strategy for solving in the inverse case. Here, we describe the detail of our strategy for solving the inverse problem. The Boundary Integral Equation still will be the our main tool in this paper. We give two examples to illustrate the behaviour of the proposed method.

Key words and Phrases: Boundary Integral Equation, Inverse Problem, Laplace Equation

1. INTRODUCTION

A practically and mathematically very interesting inverse problem concerns the determination of an unknown internal structure body from boundary measurements of voltage potentials and corresponding current fluxes. In this paper we shall discuss an application of boundary integral equation to solve such problem.

The inverse problem that we consider he is a problem of determining the shape, the size and the location of a crack by boundary measurements. Suppose that a given body contains an interior crack, the shape and the location of which are unknown. On the boundary of this body we set the value of the field and then we are measuring the flow data. Mathematically, this can be interpreted as prescribing Dirichlet data and measuring the corresponding Neumann data. Using two Dirichlet data which are different[3], and the corresponding Neumann data, and the fact the homogeneous Neumann condition is satisfied on the crack, we can reconstruct the position and the shape of the crack. The condition on the crack means that the fracture is perfectly insulating in electrostatics and in thermostatics. We proposed to solve this by minimizing the error of a certain boundary integral equation (BIE).

2000 *Mathematics Subject Classification:* 45Q05

This paper can be considered as an extension work of paper by Nishimura and Kobayashi [1]. Nishimura and Kobayashi used only a straight line crack (2D) as their validation. Using this assumption, some equations can be reduced into the simple forms because of the end of the crack has zero gap of the potential value. On the other hand, if we consider the non-self-intersecting curve, we are not only increasing the number of variables but also working with a hypersingular integral equation. For numerical point of view, the shape curve-dependent has been a big problem for inverse problems. Here we propose to approximate the crack with piecewise linear and transform the hypersingular integral using the integral parsial method.

2. THE DIRECT and THE INVERSE PROBLEM

As an introduction for inverse problem of Laplace equation, we shall begin this paper with the statement of its direct problem. Let Ω be a open connected bounded domain in $\mathbb{R}^2$ which has a smooth boundary $\partial\Omega$. Also let C be a smooth non-self-intersecting curve in Ω .

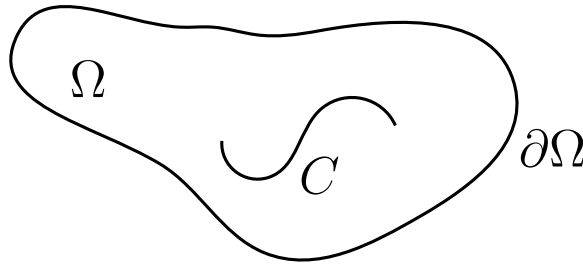


Figure 1: Domain of consideration

The direct problem for Laplace's equation is formulated into the following boundary value problem: Find a function u in $\Omega \setminus C$ which satisfies

$$\Delta u = 0 \quad \text{in } \Omega \setminus C \quad (1)$$

$$u = f \quad \text{on } \partial\Omega \quad (2)$$

$$\left. \frac{\partial u}{\partial n} \right|^+ = \left. \frac{\partial u}{\partial n} \right|^- = 0 \quad \text{on } C \quad (3)$$

$$\lim_{x(\in C) \rightarrow x^0(\in \partial C)} \varphi(x) = 0 \quad \varphi := u^+ - u^- \quad (4)$$

where the superscript $+$ ($-$) indicates the limit of u on C from the positive (negative) side and φ is the gap of u across C . The positive (negative) side in this statement

indicates the side of C into which the normal vector n ($-n$). For convenience, we call the curve C that satisfies (1)-(4) as a crack.

The solution to this problem has an integral representation, given by (see Brebbia [4])

$$u(x) = \int_{\partial\Omega} \left[u^G(y; x) \frac{\partial u}{\partial n_y}(y) - u(y) \frac{\partial u^G}{\partial n_y}(y; x) \right] ds_y - \int_C \frac{\partial u^G}{\partial n_y}(y; x) \varphi(y) ds_y \quad (5)$$

for $x \in \Omega \setminus C$, where $u^G(y; x)$ is the fundamental solution for Laplace's equation in R^2 , $u^G(y; x) = -\frac{1}{2\pi} \log |x - y|$ so that

$$\Delta_y u^G(y; x) = -\delta(y - x)$$

where $\delta(y - x)$ is a Dirac delta function concentrate in $x \in R^2$.

The unknown parts of u and $\partial u / \partial n$ on $\partial\Omega$ and φ on C are determined from the boundary integral equations given by

$$\frac{u(x)}{2} = \int_{\partial\Omega} \left[u^G(y; x) \frac{\partial u}{\partial n_y}(y) - u(y) \frac{\partial u^G}{\partial n_y}(y; x) \right] ds_y - \int_C \frac{\partial u^G}{\partial n_y}(y; x) \varphi(y) ds_y \quad (6)$$

where $x \in \partial\Omega$. Meanwhile, the homogeneous condition on the crack can be written as

$$0 = \frac{\partial}{\partial n_x} \left\{ \int_{\partial\Omega} \left[u^G(y; x) \frac{\partial u}{\partial n_y}(y) - u(y) \frac{\partial u^G}{\partial n_y}(y; x) \right] ds_y \right\} - \int_C \frac{\partial^2 u^G}{\partial n_x \partial n_y}(y; x) \varphi(y) ds_y \quad (7)$$

where $x \in C$.

In the applications one usually discretizes these equations with the help of certain shape functions and collocation method, we transform the integral equation into linear system equations. Finally, we solve the resulting algebraic equations numerically.

From a mathematical point of view, we can summarize the direct problem of Laplace equation as a problem of finding the unknown parts of u , $\partial u / \partial n$ on $\partial\Omega$ and φ on C with the shape and the position of curve C are exactly known. This problem have discussed with many authors, see for instance [2]. Now, in this paper, we consider the inverse problem of Laplace equation.

Suppose that we somehow know that Ω contains one single crack C , whose shape and location are unknown. We now prescribe a series of Dirichlet data $\tilde{u}^I$ ($I = 1, \dots, N$) on $\partial\Omega$ and measure the corresponding Neumann data $\partial \tilde{u}^I / \partial n$, or vice versa, where $N \geq 2$. The inverse problem of Laplace equation can be described as finding the most plausible shape and location of the crack C from these data. The plausible crack C in a way that there exists a function $u^I(x)$ ($x \in \Omega \setminus C$) which satisfies (1)-(4) and the boundary conditions

$$u^I = \tilde{u}^I \quad \text{and} \quad \frac{\partial u^I}{\partial n} = \frac{\partial \tilde{u}^I}{\partial n}, \quad \text{on } \partial\Omega$$

As regards the uniqueness of the solution when there exists at least one solution C , Friedman and Vogelius (see Friedman [3] showed that 2 experiment ($I = 2$) implies the uniqueness of the position, the size and the shape of C .

3. THE METHOD

In general, our inverse problem may not have a solution. Guided by (6) and (7), however, we consider the following minimization problem:

$$\min_{C \subset \Omega} J(C) \quad (8)$$

where the cost function $J(C)$ is defined as follows: Given a curve C arbitrarily, solve (see Equation (7))

$$\frac{\partial g^I}{\partial n_x}(x) + \int_C \frac{\partial^2 u^G}{\partial n_x \partial n_y}(y; x) \varphi^I(y) dS_y = 0 \quad (9)$$

for the unknown φ^I on the given C where

$$g^I(x) = \int_{\partial\Omega} \left[u^I(y) \frac{\partial u^G}{\partial n_y}(y; x) - u^G(y; x) \frac{\partial u^I}{\partial n_y}(y) \right] dS_y \quad (10)$$

We can do that because we have already known the values u^I and $\partial u^I / \partial n$ on $\partial\Omega$ for each experiment I .

Later, we define the function J as

$$J(C) = \frac{1}{2} \sum_{I=1}^N \int_{\partial\Omega} \left(f^I(x) + \int_C \frac{\partial u^G}{\partial n_y}(y; x) \varphi^I(y) dS_y \right)^2 dS_x \quad (11)$$

where

$$f^I(x) = \frac{1}{2} u^I(x) + \int_{\partial\Omega} \left[u^I(y) \frac{\partial u^G}{\partial n_y}(y; x) - u^G(y; x) \frac{\partial u^I}{\partial n_y}(y) \right] dS_y \quad (12)$$

for $x \in \partial\Omega$

Indeed, the definition of $J(C)$ is sum of errors Equation (6) along $\partial\Omega$. So, the most plausible crack C is then obtained as the minimiser of $J(C)$ (see Equation (8)).

Using this method, we solve the problem by converting it into a constrained minimization problem. As we now, there are many tools in non-linear programming as a way of solving this minimization problem numerically. Our choice in this paper is conjugate gradient method, which has been used widely in sciences and engineering applications.

In this method of non-linear programming, we needs not only the cost function but the grad of cost function also. In the numerical experiment, the computation

of grad function would be done numerically with finite difference method. This method is the simplest one but the effectiveness of this approach would be dubious from point of view of both time consuming and accuracy. In future paper, we therefore need to derive the exact formulae of this grad function. This computation is facilitated as one introduces an immobile ‘reference crack’. For more detail, see for instance [1]

4. THE NUMERICAL RESULTS

In this section we present two numerical examples to test our method. Since our interest is to see the numerical performance of the proposed method, we simulated the boundary measurement numerically. Namely, we employed BIEM to solve direct problem with the given true crack geometry and Dirichlet data (u^I) on $\partial\Omega$, and then used both (u^I) and $\partial u^I/\partial n$ thus obtained as the input to the inverse solver.

4.1. A STRAIGHT LINE

The first example is a straight line crack. We consider for Ω a circular domain having a radius $r = 3$. On the exterior boundary $\partial\Omega$ we give 2 Dirichlet data given by $u^1 = x_1, u^2 = x_2$. The exterior boundary $\partial\Omega$ is modeled by 24 piecewise linear boundary elements, and 5 degree of freedom spline elements are used for the crack. In the first example, the initial guess $((1, -2)$ and $(2, -2))$ and the true crack $((-0.3, -0.23)$ and $(0.23, 0.3))$ are straight line shown in Figure 1.

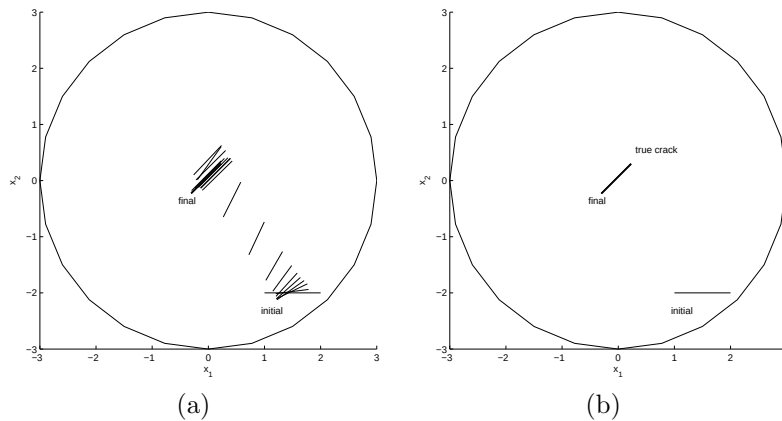


Figure 2: Mode of convergence (a), final result (b)

4. A NON-SELF INTERSECTING CURVE

As second example, we want to see how our approach can be used in non-self-intersecting curve. We decide that the crack would be a parabolic curve $((x_1, x_2) = (t, t^2)$ where $-1 \leq t \leq 1$). Like in first example, We consider for Ω a circular domain having a radius $r = 3$. On the exterior boundary $\partial\Omega$ we give 2 Dirichlet data given by $u^1 = x_1, u^2 = x_2$. The exterior boundary $\partial\Omega$ is modeled by 24 piecewise linear boundary elements, and 5 DOF spline elements are used for the crack.

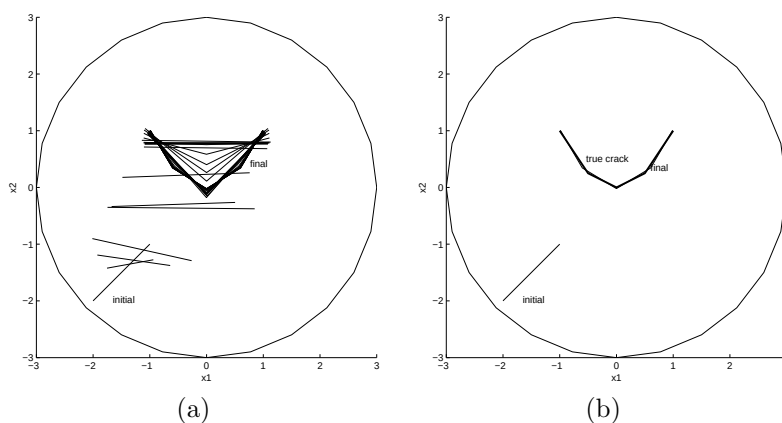


Figure 3: Mode of convergence (a), final result (b)

REFERENCES

1. N. NISHIMURA AND S. KOBAYASHI, *A Boundary Integral Equation Method For An Inverse Problem Related to Crack Detection*, International Journal for Numerical Methods in Engineering, Kyoto, 1991
2. M. WIRIANTO AND W.S. BUDHI, *Forward Problem of Laplace Equation in Plane Domain Containing a Crack*, Prosiding of National Conference in Mathematics, Semarang-Indonesia, 2006
3. A. FRIEDMAN AND M. VOGELIUS, *Determining Cracks by Boundary Measurements*, IMA preprint series #476, 1989
4. C.A. BREBBIA AND J. DOMINGUEZ, *Boundary Elements An Introductory Course*, Southampton, 1992
5. M. WIRIANTO, *Persamaan Integral Batas untuk Masalah Retakan*, Master Thesis, Institut Teknologi Bandung, 2006
6. K.E. ATKINSON *An Introduction Numerical Analysis*, Canada, 1989

WONO SETYA BUDHI: Analysis and Geometry Group, FMIPA Institut Teknologi Bandung, Indonesia
Email : wono@dns.math.itb.ac.id

MARWAN WIRIANTO: Department of Geotechnology, Delft University of Technology, The Netherlands
Email : marwan_wirianto@yahoo.com, m.liam@tudelft.nl

OPTIMAL DOWNLINK RATE ALLOCATION IN MULTICELL CDMA NETWORKS

A.I. ENDRAYANTO, A.F. GABOR, R.J. BOUCHERIE

Abstract. We study downlink rate allocation for a three cells CDMA system. Based on the discretized cell model, the rate optimization problem that maximizes the total downlink rate allocation is formulated. We propose an approximation procedure for obtaining a rate allocation in three cells case. Via numerical examples, we show that this procedure gives a good approximation of the optimal downlink rate allocation.

Key words and Phrases: Rate Allocation Optimization, Discretized Cell Model, Approximation Algorithm, CDMA Wireless Networks Model

1. INTRODUCTION

One of the most important features of future wireless communication systems is their support of different user data rates. As a major complicating factor, due to their scarcity, the radio resources have to be used very efficiently. In Code Division Multiple Access (CDMA) systems, transmissions of different terminals are separated using (pseudo) orthogonal codes. The impact of multiple simultaneous calls is an increase in the interference level, that limits the capacity of the system. The assignment of transmission powers to calls is an important problem for network operation, since the interference caused by a call is directly related to the power. In the CDMA downlink, the transmission power is related to the downlink rates. Hence, for an efficient system utilization, it is necessary to adopt a rate allocation scheme in the transmission powers assignment.

The downlink rate assignment problem has been extensively studied in the literature [2, 5, 11, 13, 15]. In [5], Duan et al. present a procedure for finding the

2000 Mathematics Subject Classification: 65K10, 90B80

power and rate allocations that minimizes the total transmit power in one cell. In [11], Javidi analyzes several rate assignments in the context of the trade-off between fairness and overall throughput. The rates are supposed to be continuous and the algorithms proposed for the rate allocation are based on solving the Lagrangean dual. Another approach for joint optimal rates and powers allocation, based on Perron-Frobenius theory, is proposed by Berggren [2] and by O'Neill et al. [13]. Berggren [2] describes a distributed algorithm for assigning base station transmitter (BTSs) powers such that the common rate of the users is maximized, while in [13] multiple rates are considered. Again, both algorithms assume continuous rates. In [7], Endrayanto et al. present a model for characterizing downlink and uplink power assignment feasibility, for a single data rate. Boucherie et al. [3] extended this model to two cells. They propose a downlink rate allocation scheme which approximates very close the maximum of a generally chosen utility function.

This paper is an extension of the results in [4]. We consider three cells of a CDMA system in the plane. Our goal is to assign rates to users in each cell such that the total utility of the system is maximized. We propose a discretized model for the three cells case by projecting the location of the users into the closest line between two BTSs. Moreover, we develop an heuristic rate allocation procedure for the three cells system. The heuristic procedure is based on solving a set of multiple choice knapsack problems.

The remainder of this paper is organized as follows. In Section 2 we present the model of three cells. In Section 3 we formulate the rate allocation optimization problems and present an approximation procedure for finding a near optimal solution. Some numerical examples will be presented in section 4. We conclude our work and present ideas for further research in Section 5.

2. MODEL

Consider a CDMA system with three cells X , Y and Z with no soft-handoff area as in Fig. 1. For the moment, let assume that users are only located in area inside the triangle. Let $[X, Y]$, $[Y, Z]$, $[X, Z]$ denote the line between BTS X and BTS Y , BTS Y and BTS Z , BTS X and BTS Z . Let B_{X_1} , B_{Y_2} be the points where the borders of cells X and Y intersect $[X, Y]$; B_{X_2} , B_{Z_1} be the points where the borders of cells X and Z intersect $[X, Z]$; B_{Y_1} , B_{Z_2} be the points where the borders of cells Y and Z intersect $[Y, Z]$. Suppose each line is divided in small segments, numbered as follows: the segments between (X, B_{X_1}) are numbered $1, 2, \dots, I_1$ and the segments between (X, B_{X_2}) are numbered $I_1 + 1, I_1 + 2, \dots, I_1 + I_2$; the segments between (Y, B_{Y_1}) are numbered $1, 2, \dots, J_1$ and the segments between (Y, B_{Y_2}) are numbered $J_1 + 1, J_1 + 2, \dots, J_1 + J_2$; and the segments between (Z, B_{Z_1}) are numbered $1, 2, \dots, K_1$ and the segments on (Z, B_{Z_2}) are numbered $K_1 + 1, K_1 + 2, \dots, K_1 + K_2$.

Let I be the incenter of the triangle $\triangle XYZ$. We approximate the location of a users in the interior of $\triangle XYZ$ as follows: the users in $\triangle XIY$, $\triangle XIZ$, respectively

ΔZYI are projected on the segments $[X, Y]$, $[X, Z]$, respectively $[Y, Z]$. Consider e.g., a segment i on $[X, Y]$ (see Fig. 1). All the users situated in the polygon bounded by segment i , the perpendiculars on $[X, Y]$ at the end points of segment i , and $[X, I]$ will be projected on segment i . We approximate the initial system with the one in which all users are projected on the lines of ΔXYZ as described above (see Fig. 2). Note that it is not necessary to know exactly the position of each user, it suffices to know approximately in which segment its location will be projected. This approximation is good enough under the deterministic path loss model (see [1, 8, 9]), i.e., the received power of a user located in a segment i at distance d_i from a BTS is equal to $P_i l_{i,X}$, where P_i is the transmission power towards segment i and $l_{i,X} = d_{i,X}^{-\gamma}$, $\gamma \geq 0$ is the path loss exponent.

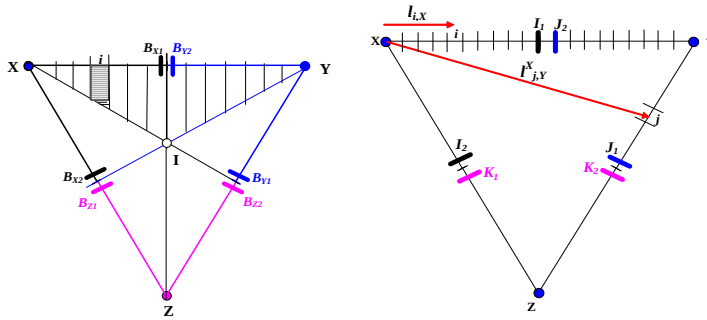


Fig. 1. Three cells cell model.

Fig. 2. Discretized Cell Model

In the approximate system, denote by m_i be the number of users in segment i of cell X , denote by n_j be the number of users in segment j of cell Y and denote by q_k the number of users in segment k of cell Z . We assume that in each segment the users are located in the middle of the segments. We suppose that the users in each segment receive the same rate and power. We measure the satisfaction of a user in a segment i by means of a positive utility function $u_i(R_i)$, where R_i is the rate allocated in segment i . For a presentation of the utility functions commonly used in the literature see [16]. Our goal is to allocate rates from a discrete and finite set $\{R_1, \dots, R_L\}$ to users such that the total utility, i.e., the sum of the utilities of all users, is maximized under the condition that the prescribed quality of service is met for all users and that a feasible power assignment exists.

2.1. Downlink Interference Model

Consider the discretized cell model for three cells in Fig. 2. For a user in segment i of cell X , the energy per bit to interference ratio is given by

$$\left(\frac{E_b}{I_0} \right)_i = \frac{W}{r_i} \frac{X_i l_{i,X}}{I_{\text{intra}} + I_{\text{inter}} + N_0},$$

where W is the system chip rate, r_i is the downlink data rate, X_i is the transmit power of BTS X to the user located in a segment i , $l_{i,X}$ is the path loss from BTS X to the user in segment i and N_0 is the thermal noise. The interferences from users in its own cell, $I_{\text{intra}} = \alpha l_{i,X} \left(\sum_{l=1}^{I_1+I_2} m_l X_l - X_i \right)$ and the interference from users in other cells, $I_{\text{inter}} = l_{i,X}^Y \left(\sum_{j=1}^{J_1+J_2} n_j Y_j \right) + l_{i,X}^Z \left(\sum_{k=1}^{K_1+K_2} q_k Z_k \right)$, where α is a non-orthogonality factor, $0 \leq \alpha \leq 1$; $l_{i,X}^Y$, respectively $l_{i,X}^Z$ are the path loss from BTS Y , respectively BTS Z , to a user in segments i of cell X ; Y_j , respectively Z_k are the transmit power of BTS Y , respectively BTS Z , to a user in the segment j of cell Y , respectively segment k of cell Z . Under the perfect power control, the energy per bit to interference ratio of the user in segment i should satisfy the power control equation $\left(\frac{E_b}{I_0} \right)_i = \epsilon^*$, where ϵ^* is the required energy per bit to interference ratio. In practice, the required energy per bit to interference ratio depends on the downlink data rate allocation [10, 12]. Without loss of generality, we assume that ϵ^* is the same for all downlink rate r_i . Then, the required transmit power of BTS X to the user in segment i is

$$X_i = V(r_i) \left(\alpha P_X + g_{i,X}^Y P_Y + g_{i,X}^Z P_Z + \frac{1}{l_{i,X}} N_0 \right), \quad (1)$$

where $V(r_i) = \frac{\epsilon^* r_i}{W + \alpha \epsilon^* r_i}$, $r_i \in \{r_1, r_2, \dots, r_L\}$, $g_{i,X}^Y = \frac{l_{i,X}^Y}{l_{i,X}}$ and $g_{i,X}^Z = \frac{l_{i,X}^Z}{l_{i,X}}$ is the fraction of path loss to a user in segment i from BTS Y , and respectively BTS Z , to path loss from own BTS, and $P_X = \left(\sum_{i=1}^{I_1+I_2} m_i X_i \right)$, $P_Y = \left(\sum_{j=1}^{J_1+J_2} n_j Y_j \right)$ and $P_Z = \left(\sum_{k=1}^{K_1+K_2} q_k Z_k \right)$ is the total transmit power of BTS X , BTS Y and BTS Z , respectively. Then, the total transmit power of BTS X can be expressed by multiplying Eq. (1) with the number of terminals in segment i , n_i , and then sum it up for all segments. We obtain the following relation of the total transmit power of BTS X , BTS Y and BTS Z for a given users distribution in the cells and for given rate allocations to users in all segments.

$$\begin{aligned} & \left(1 - \sum_{i=1}^{I_1+I_2} \alpha m_i V(r_i) \right) P_X - \left(\sum_{i=1}^{I_1+I_2} m_i V(r_i) g_{i,X}^Y \right) P_Y \\ & - \left(\sum_{i=1}^{I_1+I_2} m_i V(r_i) g_{i,X}^Z \right) P_Z = \left(\sum_{i=1}^{I_1+I_2} \frac{m_i V(r_i)}{l_{i,X}} \right) N_0, \end{aligned} \quad (2)$$

Similarly, we can express the total transmit power for BTS Y and Z . Hence, we can write the transmit power equations in the following matrix form

$$(\mathbf{I} - \mathbf{T}) \mathbf{P} = \mathbf{c}, \quad (3)$$

where $\mathbf{I}$ is an identity matrix, $\mathbf{P} = (P_X \ P_Y \ P_Z)^T$ is a column vector representing the total transmit power of BTSs in the system;

$\mathbf{c} = N_0 \left(\sum_{i=1}^{I_1+I_2} \frac{m_i V(r_i)}{l_{i,X}} \quad \sum_{j=1}^{J_1+J_2} \frac{n_j V(r_j)}{l_{j,Y}} \quad \sum_{k=1}^{K_1+K_2} \frac{q_k V(r_k)}{l_{k,Z}} \right)^T$ is a column vector representing the right hand side of the system of equations and

$$\mathbf{T} = \begin{pmatrix} I_X & I_{XY} & I_{XZ} \\ I_{YX} & I_Y & I_{YZ} \\ I_{ZX} & I_{ZY} & I_Z \end{pmatrix}. \quad (4)$$

where

- $I_X = \alpha \sum_{i=1}^{I_1+I_2} m_i V(r_i)$, $I_Y = \alpha \sum_{j=1}^{J_1+J_2} n_j V(r_j)$ and $I_Z = \alpha \sum_{k=1}^{K_1+K_2} q_k V(r_k)$ are related to the intracell interference of cell X , Y and Z ;
- $I_{XY} = \sum_{i=1}^{I_1+I_2} m_i V(r_i) g_{i,X}^Y$ and $I_{XZ} = \sum_{i=1}^{I_1+I_2} m_i V(r_i) g_{i,X}^Z$ are related to the intercell interferences from cell X to cell Y , and respectively cell Z ;
- $I_{YX} = \sum_{j=1}^{J_1+J_2} n_j V(r_j) g_{j,Y}^X$ and $I_{YZ} = \sum_{j=1}^{J_1+J_2} n_j V(r_j) g_{j,Y}^Z$ are related to the intercell interferences from cell Y to cell X , and respectively cell Z ;
- $I_{ZX} = \sum_{k=1}^{K_1+K_2} q_k V(r_k) g_{k,Z}^X$ and $I_{ZY} = \sum_{k=1}^{K_1+K_2} q_k V(r_k) g_{k,Z}^Y$ are related to the intercell interferences from cell Z to cell X , and respectively cell Y .

2.2. Downlink Transmit Power Feasibility

In this paper, we are interested in finding a feasible solution for system (3). The feasible transmit powers of all BTSs are guaranteed when the Perron-Frobenius eigenvalue of matrix $\mathbf{T}$ is smaller than *one* [14], i.e.,

$$\mathbf{P} \geq \mathbf{0} \text{ and } \mathbf{P} = (\mathbf{I} - \mathbf{T})^{-1} \mathbf{c} \Leftrightarrow \lambda(\mathbf{T}) < 1. \quad (5)$$

This provides a clear motivation for discretizing the cells into segments, since it facilitates obtaining an analytical model for characterizing the transmit power feasibility for a certain rate allocation and a certain user distribution. In the next section, based on $\lambda(\mathbf{T})$, we propose a downlink rate allocation scheme that approximates the maximum of the total utility.

3. THE RATE ALLOCATION PROBLEM

Let $R = \{R_1, R_2, \dots, R_L\}$ be the set of admissible rates, where $R_1 < R_2 < \dots < R_L$. The decision of dropping the users of a segment is equivalent with assigning zero rate to the respective segment, case in which $R_1 = 0$. The problem of allocating rates from the set R to users such that the total utility of the users is

maximized, under the condition of ensuring the required energy per bit to interference ration and a feasible power assignment, can be formulated as follows:

$$\begin{aligned}
 & \max \sum_{i=1}^{I_1+I_2} u_i(r_i) + \sum_{j=1}^{J_1+J_2} u_j(r_j) + \sum_{k=1}^{K_1+K_2} u_k(r_k) \\
 (P) : \quad & \text{s.t.} \quad \left(\frac{E_b}{I_0} \right)_i (r, p) = \epsilon^*, \text{ for each user } i \\
 & r_i \in \{R_1, \dots, R_L\}, p_i \geq 0, \\
 & \text{for each user in segment } i \text{ of cell } X, Y \text{ and } Z,
 \end{aligned}$$

where r_i , respectively p_i represent the rate, respectively the power allocated to segment i and ϵ^* is the threshold for the the energy per bit to interference ratio. For the case of two cells, using the Perron-Frobenius eigenvalue, the feasibility condition leads to a nice separation of the rate allocation optimization problem (see [3]). However, in the case of three cells we do not obtain a nice separation as in the two cells cases. Therefore we propose an approximation procedure based on the rate allocation problem for two cells. Given a rate allocation in a cell, we solve the rate allocation in the other two cells. The main idea of the heuristic procedure is as follows: we first consider cell X and cell Y. Once a rate allocation has been determined for cells X and Y, consider cell Y and cell Z and incorporate the interference from cell X as noise. Now consider cell X and cell Z and the interference from cell Y as noise, etc. This procedure may be followed until sufficient convergence is reached. We show in the next lemma that the iteration of the procedure leads to an increasing value of the total utility as defined as the objective function of optimization problem P .

Lemma 1 *The total utility of the $(N + 1)^{th}$ iteration is at least equal to the total utility of the N^{th} iteration.*

Proof. Let $\{r_i\}_X^N$ be the rate allocation of cell X that is being fixed in the N^{th} iteration. Let, $\{r_j\}_Y^N$ and $\{r_k\}_Z^N$ be the optimal rate allocations for cell Y and cell Z in the N^{th} iteration. Consider the next iteration with the rate allocation in cell Y being fixed $\{r_j\}_Y^{N+1} = \{r_j\}_Y^N$. It is clear that the sets $\{r_i\}_X^N$ and $\{r_k\}_Z^N$ are also feasible rate allocations in the $(N + 1)^{th}$ iteration. Hence, the total utility of the $(N + 1)^{th}$ iteration is at least equal to the total utility of the N^{th} iteration ■

The lemma above shows that we can guarantee that the approximation procedure leads to an increasing value of the total utility. The details of the approximation procedure are formulated in the following section.

3.1. Approximation for Three Cells Rate Allocation

Let $\{r_k\}_Z$ be the set of a feasible rate in Z, i.e., $r_Z = (r_1, r_2, \dots, r_{(K_1+K_2)})$. Given the rate allocation in cell Z, we are interested in finding the optimal rate

allocation of cell X and cell Y . Thus, the presence of cell Z as a noise power has to be incorporated in the formulation of the rate allocation problem. This is done as follows:

1. First, using (2), we express P_Z as a function of the other parameters:

$$P_Z = A_Z P_X + B_Z P_Y + K_Z, \quad (6)$$

where $A_Z = \frac{I_{ZX}}{(1-I_Z)}$, $B_Z = \frac{I_{ZY}}{(1-I_Z)}$ and $K_Z = \left(\sum_{k=1}^{K_1+K_2} \frac{q_k V(r_k)}{l_{k,Z}} \right) / (1 - I_Z)$.

2. By replacing P_Z with (6) in system (3), we obtain a new system, depending only on P_X and P_Y . The new system can be rewritten in the same form as (3) as follows:

$$(\mathbf{I} - \tilde{\mathbf{T}}) \tilde{\mathbf{P}} = \tilde{\mathbf{c}}, \quad (7)$$

where $\tilde{\mathbf{P}} = (P_X \ P_Y)^T$; $\tilde{\mathbf{c}} = \hat{\mathbf{c}} + K_Z \begin{pmatrix} I_{XZ} \\ I_{YZ} \end{pmatrix}$ with $\hat{\mathbf{c}}$ be the column vector obtained by deleting the third row of matrix $\mathbf{c}$ and

$$\tilde{\mathbf{T}} = \hat{\mathbf{T}} + \begin{pmatrix} I_{XZ} \\ I_{YZ} \end{pmatrix} \begin{pmatrix} A_Z & B_Z \end{pmatrix}. \quad (8)$$

with $\hat{\mathbf{T}}$ be the 2×2 submatrix of $\mathbf{T}$ obtained by deleting the entries of third row and the third column. From (8), we notice that the feasibility of the system with two cells under the presence of the third cell is smaller or equal to the the feasibility of the system of two cells only, i.e., $\lambda(\tilde{\mathbf{T}}) \leq \lambda(\hat{\mathbf{T}}) < 1$.

3. System (7) imply that the positive transmit power P_X and P_Y exist if and only if $\lambda(\tilde{\mathbf{T}}) < 1$, which is equivalent to the following three conditions (for details see [3]):

$$\begin{aligned} I_X + I_{XZ} A_Z &< 1, \\ I_Y + I_{YZ} B_Z &< 1, \end{aligned}$$

$$\frac{1 - (I_X + I_{XZ} A_Z)}{I_{XY} + I_{XZ} B_Z} > \frac{I_{YX} + I_{YZ} A_Z}{1 - (I_Y + I_{YZ} B_Z)}.$$

The first equation is related to the pole capacity constraint of cell X and the effect of the interference from cell Z . The second equation is related to the pole capacity constraint of cell Y and the effect of the interference from cell Z . Moreover, in the third inequality, the left part is related to cells X and Z only and the right part is related to cells Y and Z only.

Hence, given a rate allocation in cell Z , the rate allocation problem of cell X and cell Y can be formulated as follows

$$\begin{aligned}
P(Z) : \quad & \max \sum_{i=1}^{I_1+I_2} u_i(r_i) + \sum_{j=1}^{J_1+J_2} u_j(r_j) \\
& \text{s.t.} \sum_{i=1}^{I_1+I_2} \left[\alpha + g_{i,X}^Z A_Z \right] m_i V(r_i) < 1, \\
& \sum_{j=1}^{J_1+J_2} \left[\alpha + g_{j,Y}^Z B_Z \right] n_j V(r_j) < 1, \\
& 1 - \frac{\sum_{i=1}^{I_1+I_2} \left[\alpha + g_{i,X}^Z A_Z \right] m_i V(r_i)}{\sum_{i=1}^{I_1+I_2} \left[g_{i,X}^Y + g_{i,X}^Z B_Z \right] m_i V(r_i)} \\
& > \frac{\sum_{j=1}^{J_1+J_2} \left[g_{j,Y}^X + g_{j,Y}^Z A_Z \right] n_j V(r_j)}{1 - \sum_{j=1}^{J_1+J_2} \left[\alpha + g_{j,Y}^Z B_Z \right] n_j V(r_j)}, \\
& r_i, r_j \in \{R_1, \dots, R_L\}.
\end{aligned}$$

As in [3], one can prove that the optimum of problem $P(Z)$ is equal to

$\max_{t_Z \in [t_{\min}, t_{\max}]} [P_1(t_Z) + P_2(t_Z)]$, where $[t_{\min}, t_{\max}]$, $P_1(t_Z)$ and $P_2(t_Z)$ are defined as follows:

$$\begin{aligned}
P_1(t_Z) : \quad & \max \sum_{i=1}^{I_1+I_2} u_i(r_i) \\
& \text{s.t.} \sum_{i=1}^{I_1+I_2} \left[\alpha + t_Z g_{i,X}^Y + g_{i,X}^Z [A_Z + t_Z B_Z] \right] m_i V(r_i) < 1, \\
& \sum_{i=1}^I r_i > 0, \quad r_i \in \{R_1, \dots, R_L\},
\end{aligned}$$

and

$$\begin{aligned}
P_2(t_Z) : \quad & \max \sum_{j=1}^{J_1+J_2} u_j(r_j) \\
& \text{s.t.} \sum_{j=1}^{J_1+J_2} \left[\alpha + \frac{g_{j,Y}^X}{t_Z} + g_{j,Y}^Z \left[\frac{A_Z}{t_Z} + B_Z \right] \right] n_j V(r_j) < 1, \\
& r_i \in \{R_1, \dots, R_L\},
\end{aligned}$$

where

$$[t_{\min}, t_{\max}] = \left[\left[\min_{r_j \in R^L} t_{Z_1}(r_j) \right]^+, \left[\max_{r_i \in R^L, r_i \neq 0} t_{Z_2}(r_i) \right]^+ \right] \text{ with}$$

$$t_{Z_1}(r_j) = \frac{\sum_{j=1}^{J_1+J_2} \left[g_{j,Y}^X + g_{j,Y}^Z A_Z \right] n_j V(r_j)}{1 - \sum_{j=1}^{J_1+J_2} \left[\alpha + g_{j,Y}^Z B_Z \right] n_j V(r_j)}, \quad (9)$$

$$t_{Z_2}(r_i) = \frac{1 - \sum_{i=1}^{I_1+I_2} \left[\alpha + g_{i,X}^Z A_Z \right] m_i V(r_i)}{\sum_{i=1}^{I_1+I_2} \left[g_{i,X}^Y + g_{i,X}^Z B_Z \right] m_i V(r_i)}. \quad (10)$$

Note that $P_1(t_Z)$ and $P_2(t_Z)$ are multiple knapsack problems, so they can be easily solved (see [4, 6]). Thus, given the rate allocation in cell Z , we can find the optimal rate allocation in cell X and cell Y . Moreover, from the above formulations we observe that for a rate allocation in cell Y , $\{r_j\}_Y$, and given rate allocation in cell Z , we can calculate the value of $t_{Z_1}(r_j)$ using formula (9). Similarly, we can calculate the value of $t_{Z_2}(r_i)$ using formula (10) for a rate allocation in cell X and given rate allocation in cell Z . Thus, the number of t_Z values that we have to analyze is at most the number of rate allocations in cells X and Y .

Based on the above observations, we develop a heuristic procedure for finding the optimal rate allocations for three cells in the next section.

3.2. Heuristic Procedure

1. Initialize with zero the rate allocation in cell Z . Perform the following algorithm to find optimal rate allocations for cell X and cell Z .

Algorithm: *Finding optimal rate allocation for cell X and cell Y :*

- For any feasible rate allocation in cell Y , find set $\{t_{Z_1}\}$ from Eq. (9) and for any feasible rate allocation in cell X find set $\{t_{Z_2}\}$ from Eq. (10)
 - For all $t_Z \in \{t_{Z_1}\} \cup \{t_{Z_2}\}$, solve $P_1(t_Z)$ and $P_2(t_Z)$ with an algorithm for the multiple choice knapsack problem.
 - Choose the $t_Z \in \{t_{Z_1}\} \cup \{t_{Z_2}\}$ for which $\max_{t_Z} \{P_1(t_Z) + P_2(t_Z)\}$ is attained.
2. Given the rate allocation in cell X obtained in step 1, we determine the optimal rate allocation for cell Y and Z using the algorithm above where cell X and cell Z change roles.
 3. Given the rate allocation in cell Y obtained in step 2, we determine the optimal rate allocation for cell X and Z using the algorithm above where cell Y and cell Z change roles.
 4. Repeat step 1 to step 3 until the convergence is found $|U^{(N+1)} - U^{(N)}| \leq \delta$, for δ small, where U^N is the total utility of the system at N^{th} iteration.

We were not able to prove whether the value obtained by this procedure is a global optimum. Some numerical experiments using this algorithm are presented in the next section.

4. NUMERICAL EXAMPLES

Recall the discretized cell model in Fig. 2. Suppose that the distance between each two BTSs is 2 km and the cell radius of a cell is 1 km, i.e., the case of symmetric downlink cell borders. We divide the line between two BTSs into 8 equal-width segments of 250m each. We have chosen examples with small number of segments for practical reasons: for these examples is easy to calculate the global optimum through enumeration, so we could compare our results.

The system parameters for this example are given as follows: the system chip rate $W = 3.84$ MHz, thermal noise $N_0 = -169$ dBm/Hz, path loss exponent $\gamma = 4$, downlink non-orthogonality $\alpha = 0.3$, QoS required $\frac{E_b}{I_0}$, $\epsilon^* = 5$ dB and downlink transmission rate $r_i \in \{32, 64, 144\}$ kbps [10]. For this paper, we consider the utility of a cell as the total sum of the rates allocated in the segments of the cell, i.e., $u_i(r_i) = n_i r_i$.

In our numerical experiments, we have considered two types of load: homogeneous load and non-homogeneous load.

In the case of homogeneous load, we did 30 experiments where the number of users n in each segment of an instance is generated from uniform distribution on $[1, 50]$, i.e., $n \sim U(1, 50)$. We have observed that in all the examples studied, the solution obtained by the heuristic converged to the optimum (which we obtained by enumerating all possible solutions) in two or three iterations. Moreover, the total utility in each cell converged as well. We also observed that the total utility is almost equally distributed among cells.

In the case of non-homogeneous load, we have conducted 90 experiments, constructed as in Table 1, where the number of users in segment i is generated from a uniform distribution on $[a, b]$, i.e., $n_i \sim U(a, b)$. The algorithm converges very fast, attaining again the optimal solution. However, it can be noted that the time needed for converges increases proportional with the load of the system.

Case of Unifom Distribution	Convergence after			
	2 nd	3 rd	4 th	5 th
$X, Y, Z \sim U(0, 8)$	66.7%	33.3%	0%	0%
$X \sim U(0, 10), Y, Z \sim U(0, 8)$	73%	13%	14%	0%
$X \sim U(0, 10), Y, Z \sim U(0, 5)$	46.7%	46.7%	6%	0.6%
$X, Y, Z \sim U(0, 30)$	20%	60%	18%	2%
$X, Y \sim U(20, 30), Z \sim U(0, 5)$	65%	25%	10%	0%
$X, Y, Z \sim U(20, 30)$	70%	20%	5%	5%

Table 1. Numerical experiments with uniform distribution.

5. SUMMARY AND FURTHER RESEARCH

In this paper, we propose a discretized model for a three cells CDMA system. For this model, we present a heuristic procedure for finding a downlink rate allocation which maximizes the total utility of the system. Tested on small numerical experiments, the heuristic always converged to the optimal solution in few iterations.

It is among our aims for further research to develop a downlink rate algorithm that takes into account mobility of users and limited transmit powers of BTSs.

Acknowledgement. The research is partly supported by the Technology Foundation STW, Applied Science Division of NWO and the Technology Programme of the Ministry of Economic Affairs, The Netherlands.

REFERENCES

1. J. B. ANDERSEN, T.S. RAPPAPORT, S. YOSHIDA, Propagation measurements and models for wireless communications channels, *IEEE Commun. Mag.*, **33**, 42–49, 1995.
2. F. BERGGREN, Distributed power control for throughput balancing in CDMA systems, *Proceedings of IEEE PIMRC*, **1**, 24–28, 2001.
3. R.J. BOUCHERIE, A.F. BUMB, A.I. ENDRAYANTO, G.J. WOEGINER, A combinatorial approximation for CDMA downlink rate allocation, *The 7th INFORMS Telecommunication Conference*, 2004.
4. A.K. CHANDRA, D.S. HIRSCHBERG, C.K. WONG, Approximate algorithms for some generalized knapsack problems, *Theoretical Computer Science*, **3**, 293–304, 1976.
5. X. DUAN, Z. NIU, J. ZHENG, Downlink Transmit Power Minimization in Power-Controlled Multimedia CDMA Systems, *Proceedings of IEEE 13th Int. Symposium Personal, Indoor and Mobile Radio Communication*, 1976.
6. M.E. DEYER, W.O. RIHA, J. WALKER, A hybrid dynamic-programming/branch-and-bound algorithm for the multiple-choice knapsack problem, *Journal of Computational and Applied Mathematics*, **58**, 43–54, 1995.
7. A.I. ENDRAYANTO, J.L. VAN DEN BERG, R. J. BOUCHERIE, An analytical model for CDMA downlink rate optimization taking into account uplink coverage restrictions, *Journal of Performance Evaluation, Special issue: Performance Modelling and Evaluation of Heterogeneous Networks*, **59/2-3**, 225–246, 2005.
8. M. HATA, Empirical formula for propagation loss in land mobile radio services, *IEEE Transactions on Vehicular Technology*, **29**, 317–325, 1980.
9. S. HANLY, R. MATHAR, On the Optimal Base Station Density for CDMA Cellular Networks, On the optimal base-station density for CDMA cellular networks, *IEEE Transactions on Communications*, **50-8**, 1274–1281, 2002.
10. H. HOLMA AND A. TOSKALA, *WCDMA for UMTS*, John Wiley and Sons, 2000.
11. T. JAVIDI, Decentralized Rate Assignments in a Multi-Sector CDMA Network, *Proceedings of IEEE Globecom Conference*, 2003.
12. R. LITJENS, *Capacity allocation in wireless communication networks*, PhD. thesis, University of Twente, 2003.
13. D. O'NEILL, D. JULIAN AND D. BOYD, Seeking Foschini's Genie: Optimal Rates and Powers in Wireless Networks, *IEEE Transactions on Vehicular Technology*, 2003.
14. E. SENETA, *Non-Negative Matrices*, London, Allen and Unwin, 1973.

15. V.A. SIRIS, Cell Coverage based on Social Welfare Maximization, *Proceeding of IST Mobile and Wireless Telecommunications Summit Greece*, June, 2002.
16. C. TOUATI, E. ALTMAN, J. GALTIER, Fair power transmission rate control in wireless networks, *Proceedings of IEEE Globecom*, 2002.

A.I. ENDRAYANTO: Department of Mathematics, Faculty of Mathematics and Natural Sciences, Gadjah Mada University, Yogyakarta, Indonesia.
E-mail: endrayanto@ugm.ac.id

A.F. GABOR: Faculty of Mathematics and Computer Science and EURANDOM, Eindhoven University of Technology, 5600 MB, Eindhoven, The Netherlands.
E-mail: a.f.gabor@tue.nl

R.J. BOUCHERIE: Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, 7500 AE Enschede, The Netherlands, the Netherlands.
E-mail: r.j.boucherie@ewi.utwente.nl

BEGGS-BRILL CORRELATION DISCRETIZATION FOR PREDICTING PRESSURE DISTRIBUTION IN TWO PHASE FLOW PIPELINE NETWORK

A. I. HASAN, R. P. PAMUNGKAS, L. S. RIZA, W. U. BAROTO
K. A. SIDARTO, AND L. MUCHARAM

Abstract. Gas is a main energy resource, beside oil that took the strategic lead on national development country. Those resources as society live agitator and national economic development. As domestic energy resource, gas and oil has important effect for country income and feedstock for national industry. The optimization of gas production from gas condensate well is on the top of research list which has objectives for supporting gas production increasing program. The objective of this paper is to study about Beggs – Brill correlation discretization method for predicting pressure distribution of two- phase flow in pipeline network. Beggs-Brill correlation is to complicated to solve, so discretization method is needed. With this modified method we can calculate pressure distribution of two phase-flows in pipeline network easily. Network analysis model was adopted from Kirchhoff's Law. Some study case using field data has been successfully simulated and the results of calculating pressure distribution were compared with commercial software.

Key words and Phrases: Multiphase flow, Network, Pressure distribution.

1. INTRODUCTION

To delivered gas between two gathering station, petroleum engineer often using pipeline. During gas transmission along pipeline, with vary to pressure and temperature, condensate will be formed. In this case, *multiphase* term will refer to Gas-Condensate flow.

One problem that usually faced in surface facilities in petroleum industry is how to predict pressure distribution in pipeline network, since the transmission to delivered Gas was through pipeline network. One correlation that usually used for predicting pressure distribution along pipeline is Beggs-Brill correlation. But, since the Beggs-Brill correlation is in implicit form, it so hard to be applied. In this paper we present a method to predict pressure distribution along pipeline network with using discretization of Beggs-Brill correlation.

In order to derive the method, there are 3 basic assumptions that use in this work:

1. Steady state conditions,
2. Fluid composition and fluid temperature are constant across the system,
3. Effects of elbows, valves, fitting, or meters are ignored.

2. BEGGS-BRILL DISCRITIZATION

As mentioned before, Beggs and Brill Correlation is an equation that usually used to describes pressure distribution of two-phase flow in pipeline network. It will be assumed that two-phase flow model is expressed as:

$$\Delta P_i = \gamma_i Q_i^{\alpha_i} \quad (1)$$

Using algebraic, equation (1) can be written as follows:

$$\Delta P_i = \gamma_i Q_i^{(\alpha_i - 1)} Q_i \quad (2)$$

Defining:

$$\beta_i = \gamma_i Q_i^{(\alpha_i - 1)} \quad (3)$$

Substituting equation (3) into equation (2), the two-phase flow equation can be written as:

$$\Delta P_i = \beta_i Q_i \quad (4)$$

Instead of volumetric gas rate, Q_i , the rate can be converted into terms of mass flow rate, W_m to express the pressure drop in equation (4). This can be written as

$$(W_m)_i = \eta_i \Delta P_i \quad (5)$$

Where i indicates pipeline number.

Beggs and Brill's two-phase flow model is implemented to the system of equations of the network system. The two-phase flow equation derived by Beggs and Brill's can be written as:

$$\frac{\Delta P}{\Delta Z} = \frac{\left\{ \frac{g}{g_c} \bar{\rho}_m \sin \theta + \frac{f_m \bar{G}_m \bar{v}_m}{2 g_c d} \right\}}{\left\{ 1 - \frac{\bar{\rho}_m \bar{v}_m \bar{v}_{sg}}{g_c P} \right\}} \quad (6)$$

Again, with algebraic manipulations, equation (6) can be expressed in the similar form as equation (5) where η_i is defined as:

$$\eta_i = \frac{1 - C_3 W_m}{\Delta Z \left\{ \frac{C_1}{W_m} + C_2 W_m \right\}} \quad (7)$$

The derivation of the expressions for C_1 , C_2 , and C_3 follows.

Defining that:

$$W_{mi} = \bar{\rho}_m \bar{v}_m A_p \quad (8)$$

Using equation (8), the gradient pressure in equation (6) becomes:

$$\frac{\Delta P}{\Delta Z} = \frac{\left\{ \frac{g}{g_c} \bar{\rho}_m \sin \theta + \frac{g f_m W_m^2}{g_c \pi^2 d^5 \bar{\rho}_m} \right\}}{\left\{ 1 - \frac{4 \bar{v}_{sg} W_m}{\pi d^2 g_c P} \right\}} \quad (9)$$

$$\Delta P = \Delta Z \left\{ \frac{C_1 + C_2 W_m^2}{1 - C_3 W_m} \right\} \quad (10)$$

Where:

$$C_1 = \frac{g}{g_c} \bar{\rho}_m \sin \theta \quad (11)$$

$$C_2 = \frac{gf_m}{g_c \pi^2 d^5 \rho_m} \quad (12)$$

$$C_3 = \frac{4v_{sg}}{\pi d^2 g_c P} \quad (13)$$

Using discretization Method, the pressure gradient equation becomes:

$$\Delta P^{n+1} = \Delta Z \left\{ \frac{\frac{C_1}{W_m^n} + C_2 W_m^n}{1 - C_3 W_m^n} \right\} W_m^{n+1} \quad (14)$$

Or:

$$\Delta P^{n+1} = \eta' W_m^{n+1} \quad (15)$$

Or:

$$W_m^{n+1} = \eta \Delta P^{n+1} \quad (16)$$

For the i -th leg, the equation is written as:

$$W_m^{n+1} = \eta_i \Delta P^{n+1} \quad (17)$$

After getting Beggs-Brill discretization, there are two other importing step used in this paper, there are looped network model, phase behavior, and iteration procedure.

1. Looped Network Model

The looped network model is basically developed by adopting Kirchhoff's laws for electrical circuits. Further development for the looped network model for pipeline system is applying the continuity equations and energy equations at the node (junction) and the loop, respectively. The energy loss in each leg of the loop is required for the energy balance equation. This is supplied by a reformulated Beggs and Brill's two-phase flow equation.

➤ The Continuity Equations

If we have a closed circuit system, by adopting Kirchhoff's laws for any junction in it, the continuity equations can be applied for each node and expressed as:

$$\sum_{j=1}^{N_i} (\rho v A_p)_{ij} = 0 \quad ; \quad i = 1, \dots, N \quad (18)$$

Or in general, equation (18) can be written as:

$$\sum_{j=1}^{N_i} (W_m)_{ij} = 0 \quad ; \quad i = 1, \dots, N \quad (19)$$

In the presence of the mass flow rate coming from an external source or supply and leaving from the system or demand, the continuity equations can be written as:

$$\sum_{j=1}^{N_i} (W_m)_{ij} = D_i - S_i \quad ; \quad i = 1, \dots, N \quad (20)$$

Where:

$$\begin{aligned} i &= i\text{-th node within the network} \\ j &= j\text{-th leg connected to node } i \\ N_i &= \text{number of legs connected to node } i \end{aligned}$$

- N = total number of nodes within the network
 D_i = demand or flow leaving the system from the i -th node
 S_i = supply or flow coming into the system through node i

➤ The Energy Equations

By adopting Kirchhoff's laws for any closed circuit, the energy equations for the looped network system can be written as

$$\sum_{k=1}^{N_i} \Delta P_M = 0 \quad ; \quad k = 1, \dots, M \quad (21)$$

Where:

- kl = k -th leg in the l -th loop
 l = l -th loop within the network
 N_i = number of legs within the l -th loop
 M = total number of loops within the network

The pressure drop in each leg of the network can be solved using any matrix solution technique. In this study, we use Gaussian elimination technique. A problem evolves from the fact that some elements of the coefficient matrix are not constant but functions of W_m in each leg, which in turn is a function of the (ΔP_j) , the dependent variable vector. Because the mass flow rate, W_m in each legs is also an unknown variable, an iterative procedure is required in this calculation. The iteration equation is formulated as:

$$(W_m)_i^{(k+1)} = \eta_i^{(k)} \Delta P_i^{(k+1)} \quad (22)$$

2. Phase Behavior

In this study, phase behavior calculation is approach with correlation calculations from various sources. Fluid properties need to be calculated are the density and viscosity of gas, condensate, and water.

The first calculation is the calculation of gas compressibility factor. Pseudo critical properties of gas are defined as:

$$P_{pc} = 756.8 - 131\gamma_g - 3.6\gamma_g^2 \quad (23)$$

$$T_{pc} = 169.2 + 349.5\gamma_g - 74\gamma_g^2 \quad (24)$$

If there are some impurities within the gas, the pseudo critical properties need to be adjusted as follows:

$$\epsilon = 120(A^{0.9} - A^{1.6}) + 15(B^{0.5} - B^4) \quad (25)$$

Where:

- A = sum of mole fraction of CO_2 and H_2S
 B = mole fraction of N_2

$$T_{pc}' = T_{pc} - \epsilon$$

$$P_{pc}' = \frac{P_{pc} T_{pc}'}{T_{pc} + y_{\text{H}_2\text{S}}(1 - y_{\text{H}_2\text{S}}) \epsilon} \quad (26)$$

Pseudo reduced properties are calculated as:

$$T_{pr} = \frac{T}{T_{pc}} \quad (27)$$

$$P_{pr} = \frac{P}{P_{pc}} \quad (28)$$

The gas compressibility factor then can be written as follows:

$$z = 1 + A_{11}\rho_{pr} + A_{22}\rho_{pr}^2 - A_{33}\rho_{pr}^5 + 0.6134\left(1 + 0.721\rho_{pr}^2\right)\left(\frac{\rho_{pr}^2}{T_{pr}^2}\right)\exp(-0.721\rho_{pr}^2) \quad (29)$$

Where:

$$\begin{aligned} A_{11} &= 0.3265 - \left(\frac{1.07}{T_{pr}}\right) - \left(\frac{0.5339}{T_{pr}^3}\right) + \left(\frac{0.01569}{T_{pr}^4}\right) - \left(\frac{0.05165}{T_{pr}^5}\right) \\ A_{22} &= 0.5475 - \left(\frac{0.7361}{T_{pr}}\right) + \left(\frac{0.1844}{T_{pr}^2}\right) \\ A_{33} &= 0.1056 \left(\frac{-0.7361}{T_{pr}} + \frac{0.1844}{T_{pr}^2}\right) \\ \rho_{pr} &= 0.27 \left(\frac{P_{pr}}{z T_{pr}}\right) \end{aligned}$$

➤ Density of gas

Following are the equations used to calculate the density of gas.

$$\gamma_g = \frac{M_a}{29} \quad (30)$$

$$M_a = 29\gamma_g \quad (31)$$

$$\rho_g = \frac{PM}{10.732 z T} \quad (32)$$

➤ Viscosity of gas

The calculation of gas viscosity based on standing correlations considering the effect of non hydrocarbon molecules on viscosity at atmospheric pressure.

$$\mu_{g1} = \mu_{g1}^{uncorrected} + (\Delta\mu)_{H_2S} + (\Delta\mu)_{N_2} + (\Delta\mu)_{CO_2} \quad (33)$$

Where:

$$\begin{aligned} \mu_{g1}^{uncorrected} &= \left[1.79(10^{-5}) - 2.062(10^{-6})\gamma_g\right](T - 460) \\ &\quad + 8.188(10^{-3}) - 6.15(10^{-3})\log \gamma_g \\ (\Delta\mu)_{H_2S} &= y_{H_2S} \left[8.48(10^{-3})\log \gamma_g + 9.59(10^{-3})\right] \\ (\Delta\mu)_{N_2} &= y_{N_2} \left[9.08(10^{-3})\log \gamma_g + 6.24(10^{-3})\right] \\ (\Delta\mu)_{CO_2} &= y_{CO_2} \left[8.49(10^{-3})\log \gamma_g + 3.73(10^{-3})\right] \end{aligned}$$

Further development for gas viscosity calculation has been done by Dempsey to determine viscosity of gas at various pressures, as follow:

$$\mu_g = \mu_{g1} \frac{\exp(C_{11})}{T_{pr}} \quad (34)$$

Where:

$$\begin{aligned}
 C_{11} &= B_{11} + B_{22}T_{pr} + B_{33}T_{pr}^2 + B_{44}T_{pr}^3 \\
 B_{11} &= -2.46211820 + 2.97054714P_{pr} \\
 &\quad - 2.86264054 \times 10^{-1}P_{pr}^2 + 8.05420522 \times 10^{-3}P_{pr}^3 \\
 B_{22} &= 2.80860949 - 3.49803305P_{pr} \\
 &\quad + 3.60373020 \times 10^{-1}P_{pr}^2 - 1.044324 \times 10^{-2}P_{pr}^3 \\
 B_{33} &= -7.93385684 \times 10^{-1} + 1.39643306P_{pr} \\
 &\quad - 1.49144925 \times 10^{-1}P_{pr}^2 + 4.41015512 \times 10^{-3}P_{pr}^3 \\
 B_{44} &= 8.39387178 \times 10^{-2} - 1.86408848 \times 10^{-1}P_{pr} \\
 &\quad + 2.03367881 \times 10^{-2}P_{pr}^2 - 6.09579263 \times 10^{-4}P_{pr}^3
 \end{aligned}$$

➤ Density of Condensate

The calculation to determine condensate density is done with the following simple relationship.

$$\rho_o = \gamma_o \rho_w$$

➤ Viscosity of Condensate

For determining viscosity of condensate, Glaso's mathematical relationship is used.

$$\mu_o = \left[3.141(10^{10}) \right] (T - 460)^{-3.444} [\log(API)]^a \quad (35)$$

$$API = \left(\frac{141.5}{\gamma_o} \right) - 131.5 \quad (36)$$

$$a = 10.313 [\log(T - 460)] - 36.447 \quad (37)$$

➤ Density of Water

Water density is computed using the following formulation

$$\rho_w = 62.368 + 0.438603 S + 1.60074(10^{-3})S^2 \quad (38)$$

Where:

S = salinity in weight percent solids

➤ Viscosity of Water

Water viscosity obtained from the following formulation

$$\mu_{w1} = A(T - 460)^B \quad (39)$$

Where:

$$A = A_0 + A_1S + A_2S^2 + A_3S^3$$

$$A_0 = 109.574 \quad A_1 = -8.40564$$

$$A_2 = 0.313314 \quad A_3 = 8.72213(10^{-3})$$

And,

$$\begin{aligned}
B &= B_0 + B_1 S + B_2 S^2 + B_3 S^3 \\
B_0 &= -1.12166 & B_1 &= 2.63951(10^{-2}) \\
B_2 &= -6.79461(10^{-4}) & B_3 &= -5.47119(10^{-5})
\end{aligned}$$

Then, viscosity of water in a certain value of pressure can be obtained from the following relationship

$$\mu_w = \mu_{w1} \left[0.9994 + 4.0295(10^{-5})P + 3.1062(10^{-9})P^2 \right] \quad (40)$$

➤ Density and Viscosity of Liquid

After density and viscosity of condensate and water are obtained, next step is to combine them into liquid property using the following equations

$$\rho_l = \left(\frac{CGR}{CGR + WGR} \right) \rho_o + \left(\frac{WGR}{CGR + WGR} \right) \rho_w \quad (41)$$

$$\mu_l = \left(\frac{CGR}{CGR + WGR} \right) \mu_o + \left(\frac{WGR}{CGR + WGR} \right) \mu_w \quad (42)$$

3. Iteration Procedure

1. Set the values of mass flow rate, W_m in all legs with some reasonable value.
2. Estimate the pressure in every node.
3. Estimate the average pressure, $\bar{P}$, in each legs using this relationship:

$$\bar{P}_j = \frac{P(i, j) + P(i, j+1)}{2} \quad (43)$$

4. With the estimated value of $\bar{P}$ and $\bar{T}$, predict $\rho_g, \rho_L, \mu_g, \mu_L$ and λ_L using the phase behavior model explained above.
5. Using estimated flow rate, W_m and fluid properties from step (4), estimate the density of mixture, ρ_m , and the velocity of mixture, v_m using the following relationship:

$$\rho_m = \rho_g (1 - H_L) + \rho_L H_L \quad (44)$$

$$v_m = \frac{W_m}{\rho_m A_p} \quad (45)$$

6. Estimate the liquid holdup, H_L , for a horizontal configuration in each leg, using Beggs-Brill's empirical correlation equation.

$$H_L(0) = f_g(\lambda_L, N_{FR}) \quad (46)$$

Where:

$$N_{FR} = \frac{v_m^2}{g d} \quad (47)$$

$H_L(0)$ = Liquid holdup for horizontal pipeline

7. Estimate the superficial gas velocity, v_{sg} , using the following relationship:

$$\rho_m = \rho_g (1 - H_L) + \rho_L H_L \quad (48)$$

$$Q_T = \frac{W_m}{\rho_m} \quad (49)$$

$$Q_g = Q_T (1 - H_L) \quad (50)$$

$$v_{sg} = \frac{Q_g}{A_p} \quad (51)$$

Where:

- A_p = cross section area of pipeline
 W_m = total mass flow rate
 H_L = liquid holdup

8. Estimate the Reynolds Number N_{Re} using:

$$N_{Re} = \frac{(\rho_L H_L + \rho_g (1 - H_L)) v_m d}{\mu_L H_L + \mu_g (1 - H_L)} \quad (52)$$

9. Estimate the friction factor of mixture, f_m , using Chen's equation:

$$\frac{1}{\sqrt{(f_m)_D}} = -2 \log \left[\frac{\varepsilon}{3.7065 d_p} - \frac{5.0452}{N_{Re}} \log \left\{ \frac{1}{2.8257} \left(\frac{\varepsilon}{d_p} \right)^{1.1098} + \frac{5.8506}{N_{Re}^{0.8981}} \right\} \right]$$

10. Calculate the constants C_1 , C_2 , and C_3 which are defined in equations (11), (12), and (13).

11. For each legs, calculate the coefficient factor, η_i which is defined by:

$$\eta_i = \frac{1 - C_3 W_m}{\Delta Z \left\{ \frac{C_1}{W_m} + C_2 W_m \right\}} \quad (53)$$

12. Set up the matrix and solve for ΔP from the equation (9) using Gaussian elimination technique.

13. Find the pressure at each node in the system.

14. Calculate $(W_m)_i^{(k+1)}$ using:

$$(W_m)_i^{(k+1)} = \eta_i^{(k)} \Delta P_i^{(k+1)} \quad (54)$$

15. Calculate the Euclidean of the pressure which is defined as

$$\xi = \sqrt{\sum_{i=1}^N (P_i^m - P_i^{m+1})^2} \quad (55)$$

16. Check for convergence of pressure in each node and mass flow rate in each leg. If the value of ξ computed from step (16) is less than a specified tolerance, τ and the mass flow rates in the system also satisfy the continuity equations, then it is said to converge. If ξ is greater than τ , repeat the calculation starting from step (3) using updated values of the mass flow rate, $W_m^{(k+1)}$, obtained from step (14), and updated values of pressure which is obtained by

$$P_i = \frac{P_i^k + P_i^{k+1}}{2} \quad (56)$$

3. STUDY CASE

The pipeline model that used in this study case is a simple pipeline network with two inlets and one outlet. Look at picture below:

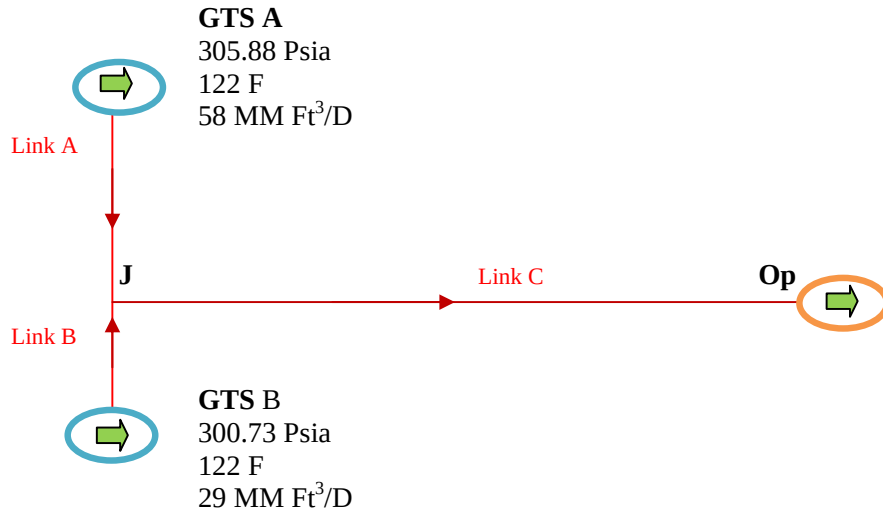


Figure 2 simple pipeline networks

The aims are to predict pressure at N1 and pressure at M1. This study case used real field data from one of branch of X network in Borneo Island. The data included source data at both of inlet and pipe geometry from each link.

4.1 Source Data

Table 14.2 Source Data at GTS A

Data		
Flow Rate	58	MM Ft ³ /D
Pressure	305.88	Psia
Temperature	122	F
CGR	19	Bbl/MMscf
WGR	25	Bbl/MMscf
SG gas	0.71	Sp gr
SG condensate	49.731	API
SG water	1	Sp gr
N2	0.06	Mole %
CO2	6	Mole %
H2S	0	Mole %

Table 14.3 Source Data at GTS B

Data		
Flow Rate	29	MM Ft ³ /D
Pressure	300.73	Psia
Temperature	122	F
CGR	19	Bbl/MMscf
WGR	25	Bbl/MMscf
SG gas	0.71	Sp gr
SG condensate	49.731	API
SG water	1	Sp gr

N2	0.06	Mole %
CO2	6	Mole %
H2S	0	Mole %

4.2 Pipe Geometry Data

Table 14.4 Pipe Geometry Data from GTS A to J (Link A)

Data		
Inside Diameter	18.5	In
Absolute Roughness	0.0018	Ft
Pipe Elevation	0	Degree
Pipe Length	4000	M

Table 14.5 Pipe Geometry Data from GTS B to J (Link B)

Data		
Inside Diameter	18.5	In
Absolute Roughness	0.0018	Ft
Pipe Elevation	0	Degree
Pipe Length	10	M

Table 14.6 Pipe Geometry Data from J to Op (Link C)

Data		
Inside Diameter	18.5	In
Absolute Roughness	0.0018	Ft
Pipe Elevation	0	Degree
Pipe Length	5800	M

The result from simulation using OPPINET Software which is compared with Data from X field can be found below:

Table 14.7 Result of Pressure Distribution

Result		OPPINET Software Prediction		X field Data		
		Pressure	Drop	Pressure	Drop	
Link A	GTS A	305.88		305.88		Psia
	J	300.70	5.18	300.73	5.15	Psia
Link B	GTS B	300.73		300.73		Psia
	J	300.70	0.03	300.73	0	Psia
Link C	J	300.70		300.73		Psia
	Op	285.39	15.31	284.69	16.04	Psia

From the result above, the difference between X field data with prediction is less than 1 % at M1. This result shows that the model developed in this paper can predict pressure distribution in pipeline network.

4. CONCLUDING REMARKS

Two-Phase flow model has been developed to predict pressure distribution in pipeline network system. The model is based on four main concepts; two-phase flow model, the looped network analysis model, the phase behavior model, and procedure of iteration. Data input handling is simple and the result are also displayed in a way that make it attractive to engineer involved in designing network system that may be subjected to two-phase flow. The model has been tested using field data from X network in Borneo Island. The test result founded that the pressure distribution which predicted using Beggs-Brill correlation discretization have closely value with X field network data.

Acknowledgement.

We would like to thank you to Prof. Septoratto Siregar and all researcher and research assistant at OPPINET which gave us many reference in order to doing our work.

REFERENCES

1. Ahmed, T, Hydrocarbons Phase Behavior, Gulf Publishing Company, Houston, Texas, 1989
2. Mucharam, L., Adewumi, M.A, A Compositional Two Phase Flow Model for Analyzing and Designing Complex Pipeline Network System, Pennsylvania State U, Pennsylvania, 1990
3. McCain, W.D, The Properties of Petroleum Fluids, PennWell Publishing Company, Tulsa, Oklahoma, 1990

A.I. Hasan, R.P.Pamungkas, L.S. Riza: Research Consortium OPPINET Institut Teknologi Bandung
e-mail: agus_ismail@yahoo.com, rela_pamungkas@yahoo.com.sg,
lala@dns.itb.ac.id

W.U. Baroto: Jurusan Prodi Teknik Perminyakan Institut Teknologi Bandung; Research Consortium OPPINET Institut Teknologi Bandung
e-mail: wu_baroto@yahoo.com

K.A. Sidarto : Jurusan Prodi Matematika Institut Teknologi Bandung; Research Consortium OPPINET Institut Teknologi Bandung
e-mail: sidarto@dns.itb.ac.id

HYPERTRANSCENDENTAL EQUATION

CHIN CHOU HSIN

Abstract. The hypertranscendental equation is constructed. The goal is to investigate the reality of hypertranscendency.

Key words and Phrases : hypertranscendency, hypertransor

1. INTRODUCTION

Physics is based on the measure of observation. In quantum theory the measurable states are forming an algebraic space. However there exist irrational measure and transcendental measure in mathematics [1,2,3]. The quantum groups and non-commutative geometry are the central theme to study whether the transcendental measure can be satisfied by Yang-Baxter equation [4,5]. The scaling symmetries are considered by q-differentials [6,7,8]. Therefore q-extensions of hypergeometric series, Bessel series and Legendre series are much studied currently. Actually, this technique is described as a method of constructing orthogonal basis in spaces of continuous functions, and then developing some type of infinite dimensional transcendental vector space. Natural questions to ask are that can we extend the analysis to hypertranscendental q-differentials [9]? and can we apply the realization to quantum field theory under hypertranscendental action [10]? In this short note, it is an attempt to construct some equations which satisfy a three-term recurrence relation formed by the exponents of q satisfying the tangent circle packing condition [11].

2. HYPERTRANSCENDENCY

Let logic C be the hypertranscendental symbolic formal language HTSFL which is defined by the grammar

$$G_{\tau} = (\{C\}, \{+, \cdot, \tilde{*}, *, \circ, (,), z_1, z_2, z_3, \dots, z_{\tau}, \dots\}).$$

Let $z_{\tau} \in \{\Delta, \Delta^{-1}, \tilde{\Delta}, \tilde{\Delta}^{-1}, Q, Q^{-1}, \tilde{Q}, \tilde{Q}^{-1}\}$, where the operators $\Delta, \Delta^{-1}, \tilde{\Delta}, \tilde{\Delta}^{-1}, Q, Q^{-1}, \tilde{Q}, \tilde{Q}^{-1}$, are defined as follows

$$\Delta f(z) := \Delta_{(\alpha, \beta)} f(z) = \frac{f(q^{\alpha} z) - f(q^{\beta} z)}{q^{\alpha} z - q^{\beta} z},$$

$$\Delta\Delta^{-1}f(z) = \Delta^{-1}\Delta f(z) = f(z),$$

$$\tilde{\Delta}f(z) := \tilde{\Delta}_{(\alpha,\beta)}f(z) = \frac{f(z^{q^\alpha}) - f(z^{q^\beta})}{z^{q^\alpha} - z^{q^\beta}},$$

$$\tilde{\Delta}\tilde{\Delta}^{-1}f(z) = \tilde{\Delta}^{-1}\tilde{\Delta}f(z) = f(z),$$

$$Qg(f(z)) := g(qf(z)),$$

$$QQ^{-1}f(z) = Q^{-1}Qf(z) = f(z),$$

$$\tilde{Q}g(f(z)) := g(f^q(z)),$$

$$\tilde{Q}\tilde{Q}^{-1}f(z) = \tilde{Q}^{-1}\tilde{Q}f(z) = f(z),$$

where α, β belong to a number field and $q \rightarrow 1$.

The operator $\Delta^n, \forall n \in \mathbb{Z}^+$, can be obtained recursively, i.e.
 $\Delta^n = \Delta^{n-1}Q + \sigma_{n-1}$, where

$$\sigma_{n-1} = \frac{[Q^\alpha - (q^{\alpha-\beta})^{n-1}Q^\beta - (q^\alpha)^n q^{-(n-1)}(q^\alpha - q^\beta)zQ] \prod_{i=1}^{n-2} [Q^\alpha - (q^{\alpha-\beta})^i Q^\beta]}{(q^\alpha)^{\frac{1}{2}n(n+1)}(q^\alpha - q^\beta)z^n}.$$

Let the symbols

$$\begin{aligned} [n]_{(q^\alpha, q^\beta)} &:= \frac{(q^\alpha)^n - (q^\beta)^n}{q^\alpha - q^\beta} \\ &= (q^\alpha)^0 (q^\beta)^{n-1} + (q^\alpha)^1 (q^\beta)^{n-2} + (q^\alpha)^2 (q^\beta)^{n-3} + \dots + (q^\alpha)^{n-1} (q^\beta)^0 \\ &= \sum_{i=1}^{n-1} (q^\alpha)^i (q^\beta)^{n-1-i}, \end{aligned}$$

and

$$\begin{bmatrix} n \\ k \end{bmatrix}_{(q^\alpha, q^\beta)} := \frac{[n]_{(q^\alpha, q^\beta)} [n-1]_{(q^\alpha, q^\beta)} [n-2]_{(q^\alpha, q^\beta)} \dots [n-k+1]_{(q^\alpha, q^\beta)}}{[1]_{(q^\alpha, q^\beta)} [2]_{(q^\alpha, q^\beta)} [3]_{(q^\alpha, q^\beta)} \dots [k]_{(q^\alpha, q^\beta)}},$$

then the Leibniz rule reads

$$\Delta^n_{(q^\alpha, q^\beta)}[g(z)f(z)] = \sum_{k=0}^n \begin{bmatrix} n \\ k \end{bmatrix}_{(q^\alpha, q^\beta)} [(Q^\alpha)^{n-k} \Delta^k g(z)][(Q^\beta)^k \Delta^{n-k} f(z)].$$

The rational hypertranscendental equations can be constructed if we consider α, β are rational numbers. For instance, the tangent circle packing condition is

$$(z\Delta_{(1/s, \beta_3)})^{rs} = (z\Delta_{(1/s_1, \beta_1)})^{r_1 s_1} (z\Delta_{(1/s_2, \beta_2)})^{r_2 s_2} \text{ or}$$

$$(z\Delta_{(\alpha_3, 1/s)})^{rs} = (z\Delta_{(\alpha_1, 1/s_1)})^{r_1 s_1} (z\Delta_{(\alpha_2, 1/s_2)})^{r_2 s_2}, \text{ where}$$

$$\alpha_1 \text{ or } \beta_1 = r_1 / s_1, \alpha_2 \text{ or } \beta_2 = r_2 / s_2, \alpha_3 \text{ or } \beta_3 = r / s,$$

$$r / s = (r_1 + r_2) / (s_1 + s_2), r_1, s_1, r_2, s_2, r, s \in \mathbb{Z}^+, r < s.$$

Similarly, for the operator $\tilde{\Delta}$, we have the fractional hypertranscendental equation

$$((z^{q^{1/s}} - z^{q^{\beta_3}})\tilde{\Delta}_{(1/s, \beta_3)})^{rs} = ((z^{q^{1/s_1}} - z^{q^{\beta_1}})\tilde{\Delta}_{(1/s_1, \beta_1)})^{r_1 s_1} ((z^{q^{1/s_2}} - z^{q^{\beta_2}})\tilde{\Delta}_{(1/s_2, \beta_2)})^{r_2 s_2},$$

$$\text{or } ((z^{q^{\alpha_3}} - z^{q^{1/s}})\tilde{\Delta}_{(\alpha_3, 1/s)})^{rs} = ((z^{q^{\alpha_1}} - z^{q^{1/s_1}})\tilde{\Delta}_{(\alpha_1, 1/s_1)})^{r_1 s_1} ((z^{q^{\alpha_2}} - z^{q^{1/s_2}})\tilde{\Delta}_{(\alpha_2, 1/s_2)})^{r_2 s_2}.$$

Let the hypertranscendental progression of derivative be

$$\text{HTD } f(z) := \frac{f({}^n z_{\tau_1}) - f({}^n z_{\tau_2})}{{}^n z_{\tau_1} - {}^n z_{\tau_2}},$$

where ${}^n z_{\tau}$ is the nth hyperpower of z_{τ} with n variables in G_{τ} . For example $n = 2$, we have

$$\text{HT } \Delta f(z) := \frac{f((q^{\alpha} z)^{q^{\beta} z}) - f((q^{\alpha'} z)^{q^{\beta'} z})}{(q^{\alpha} z)^{q^{\beta} z} - (q^{\alpha'} z)^{q^{\beta'} z}} \text{ or}$$

$$\text{HT } \tilde{\Delta} f(z) := \frac{f((z^{q^{\alpha}})^{z^{q^{\beta}}}) - f((z^{q^{\alpha'}})^{z^{q^{\beta'}}})}{(z^{q^{\alpha}})^{z^{q^{\beta}}} - (z^{q^{\alpha'}})^{z^{q^{\beta'}}}} \text{ with}$$

4 variations $\alpha = \alpha'$ and $\beta \neq \beta'$,

$$\alpha \neq \alpha' \text{ and } \beta = \beta',$$

$$\alpha \neq \alpha' \text{ and } \beta \neq \beta',$$

$$\alpha = \beta' \text{ and } \beta = \alpha'.$$

Now, it is ready to consider the quantum motives induced by the HTD. The development is given by the exact sequence

$$(1) \xrightarrow{\text{HTD}} V \rightarrow V \rightarrow \dots \xrightarrow{\text{HTD}} V \rightarrow V \rightarrow (1),$$

where V is the fixed function space. More general setting, for physics reasons, is that

$$\text{HTD} := \Delta_{(\alpha, \beta)}^n - u(z_{\tau})o \text{ or } \text{HTD} := \tilde{\Delta}_{(\alpha, \beta)}^n - u(z_{\tau})o \text{ acting on } v(z_{\tau}) \in V$$

with $o \in \{+, \cdot, \tilde{*}, *, \circ\}$.

Therefore, the hypertranscendency is established by the cohomology indicated above.

3. EPILOGUE

The reasonable knowledge of elliptic modules [12] and motives [13] is actually required in order to understand the fundamentals of hypertranscendancy. The extensions of quantum groups and hypertranscendental differentials are also needed to construct the general framework of the hypertranscendental buildings. Therefore, the hypertranscendental motives (ht-motives) become a natural logic consequence of linking other algebraic curves than hyperelliptic curves. It is our hope that this paper gives an elementary account of the aforementioned viewpoint and to prove the theory on the reality of β calculus [14]. More accessibly, the hypertranscendental Yang-Mills equation

$$F_{\mu\nu} = \Delta_\mu A_\nu - \Delta_\nu A_\mu + cQ^{-1}[A_\mu, A_\nu]Q \text{ or}$$

$$\tilde{F}_{\mu\nu} = \tilde{\Delta}_\mu A_\nu - \tilde{\Delta}_\nu A_\mu + \tilde{c}\tilde{Q}^{-1}[A_\mu, A_\nu]\tilde{Q},$$

where c and $\tilde{c}$ are coupling constants belonging to a hypertranscendental set [15], is solvable by the hypertransor analysis.

4. PROBLEMS ON HYPERTRANSCENDENCY

A function $f(z)$ is said to be algebraic if it satisfies a polynomial equation $P(z, f(z)) = 0$, and is, otherwise, called transcendental. A function $f(z)$ is said to be differentially algebraic if it satisfies a polynomial equation $P(z, f(z), f^{(1)}(z), \dots, f^{(n)}(z)) = 0$, and is, otherwise, called hypertranscendental [16, 17].

$$\text{Let the Liouville-Riemann differential of } f(z) \text{ be } f^{(\sigma)}(z) = \frac{\Gamma(\sigma+1)}{2\pi i} \oint_C \frac{f(s)}{(s-z)^{(\sigma+1)}} ds,$$

where z is an interior point of contour C , and σ is a complex number.

$$\text{Let the Jackson differential of } f(z) \text{ be } JDf(z) = \lim_{q \rightarrow 1} \frac{f(q^{\sigma_1} z) - f(q^{\sigma_2} z)}{q^{\sigma_1} z - q^{\sigma_2} z},$$

where σ_1, σ_2 are in any number field.

$$\text{Let the Frobenius differential of } f(z) \text{ be } FDf(z) = \lim_{q \rightarrow 1} \frac{f(z^{q^{\sigma_1}}) - f(z^{q^{\sigma_2}})}{z^{q^{\sigma_1}} - z^{q^{\sigma_2}}},$$

where σ_1, σ_2 are in any number field.

Therefore, the problems are the followings

- (i) Does there exist any function $f(z)$ which satisfies the polynomial equation $P(z, f^{(\sigma)}(z)) = 0$?
- (ii) Does there exist any function $f(z)$ which satisfies the polynomial equation $P(z, JDf(z), JD^{(1)}f(z), \dots, JD^{(n)}f(z)) = 0$?
- (iii) Does there exist any function $f(z)$ which satisfies the polynomial equation $P(z, FDf(z), FD^{(1)}f(z), \dots, FD^{(n)}f(z)) = 0$?

Then, $f(z)$ is said to be hyper-hypertranscendental if it satisfies none of the aforementioned polynomial equations.

Acknowledgement. This work is performed under the auspices of the project 95W801.

REFERENCES

1. Borwein J.M. and Borwein P.B., *Pi and the AGM - A Study in Analytic Number Theory and Computational Complexity*, Canadian Math. Soc. Series of Mono. and Adv. Texts, Wiley, 1986.
2. Baker A. Ed., *New Advances in Transcendence Theory*, Camb. Univ. Press, Camb., 1986.
3. Parshin A.N. and Shafavarich I.R., Eds., *Number Theory IV, Transcendental Numbers*, Springer, 1998.
4. Manin Yu. I., *Quantum Groups and Non-commutative Geometry* CRM, Universite de Montreal, Montreal, Canada, 1988.
5. Drinfeld V.G., Hopf Algebras and the Quantum Yang-Baxter Equation, *Soviet Math. Dokl. SSSR*, 32, 1, 254-258, 1985.
6. Jackson F.H., On q-definite Integrals, *Quat. Joul. Pure and Appl. Math.*, 41, 193-203, 1910.
7. Tsallis C., "Possible Generalization of Boltzmann-Gibbs Statistics", *Joul. Stat. Phys.*, 52, 1/2, 479-487, 1988.
8. Van Assche W., Little q-Legendre Polynomials and Irrationality of Certain Lambert Series, *The Ramannujan Joul.*, 5, 295-310, 2001.
9. Chin C.H. and Chin H.H., Hypertranscendental Differential, *Proc. of IMFP 2005*, Eds. Wong C.S. et al., Malaysia, 276-279, 2007.
10. Chin C.H., Quantum Field Theory under Hypertranscendental Action, *Proc. of the Satellite Meeting of STATPHYS-19, Statistical Models, Yang-Baxter Equation and Related Topics, and Symmetry, Statistical Mechanical Models and Applications*, Eds. Ge M.L. and Wu F.Y., World Sci., Singapore, 370-377, 1995.
11. Ford L.R., Fractions, *Amer. Math. Monthly*, 45, 586-601, 1938.
12. Drinfeld V.G., Elliptic Modules, *Math. USSR Sb.*, 23, 4, 457-502, 1974.
13. Anderson G.W., t-motives, *Duke Math. Joul.*, 53, 2, 457-502, 1986.
14. Chin C.H., β Calculus, *Proc. of CMS*, Eds. Chang J.H. and Lin S.S., 237-244, 1994.
15. t Hooft G., Ed., *50 Years of Yang-Mills Theory*, World Sci., Singapore, 2005.
16. Rubel L.A., A Survey of Transcendentally Transcendental Function, *Amer. Math. Monthly*, 96, 777-788, 1989.
17. Thakur D.S., *Function Field Arithmetic*, World Sci., Singapore, 2004.

Chin: Institute of Mathematical Modeling and Scientific Computing and Department of Electrophysics, National Chiao Tung University, Hsin Chu, Taiwan, R.O.C.
E-mail: jchchin@cc.nctu.edu.tw

BIFURCATION ANALYSIS OF A NEW-PRODUCT GROWTH MODEL

JOICE RUTH JULIANA AND M. SETIJO WINARKO

Abstract. Every part of our life is a system. Day by day, the complexity of the system is growing up. In fact, problems always are found in a system. Thus, study of model dynamic system and the analysis are needed. This paper discuss dynamic model related with business problem, that is dynamical of spread of new-product in a market. The idea come from Sterman [5]. And then, Juliana [3] used the idea to propose mathematical model of a new product growth with owner and competitor treatments. It has been done the equilibrium points analysis and the economical interpretation. Now, the work will be continued by doing advanced analysis. It is high risk to neglect a little change if the little change imply the change of dynamical behavior the system. Bifurcation theory is the powerfull theory for investigation about it in advanced analysis.

First, it is traced the parameters condition in which the bifurcation occurs. Next, bifurcation codimension one related with one zero eigenvalue and a pair of pure imaginary eigenvalues are shown. Also, bifurcation codimension two related with double zero eigenvalues is found. After that, interpretation of the bifurcation analysis is done. The last, some numerical examples are given.

The result of the investigation is expected to give positive contribution in business field. The detail information of the dynamical behavior system hopefully can be used to increase the efficiency of productivity and marketing managemen

Keywords: *a new-product growth, bifurcation*

1. INTRODUCTION

We find that dynamical model has been applied in many field of real-life problems [1, 2, 4, 5, 7, 9]. We should obtain the exact solution of the dynamical model and interpret it to find out the solution of real-life problem. In fact, the closer the model to the real, the more difficult find the exact solution. However, the difficulty can be overcome by qualitative behavior analysis. This method analyze two important matters. First, it can be observed behavior of all solution curves of the system for infinite time. Usually, it is said that we observe stability of equilibrium points. Second, we can analyse the change of dynamical behaviour system because of a change of the parameters system value, it is called as bifurcation.

In this paper will be shown the bifurcation analysis of a new-product growth model. The model as proposed by Juliana [3], and we add a factor that is individual come in to the system naturally. The model become

$$\begin{aligned}\frac{dP}{dt} &= T_o(P) - T_c(P) - \alpha P + \beta P - cPA \\ \frac{dA}{dt} &= T_o(A) - T_c(A) - (\alpha + g)A + \beta A + cPA\end{aligned}\tag{1}$$

with $P(t)$ (potential adopters) is population of individual who has potency to adopt new product at time t , $A(t)$ (adopter) is population of individual who adopt new product at time t , α is the speed of individual out from the system naturally, β is the speed of individual come in to the system naturally, c is the speed of contact between individual of P and A such as individual of P adopt new product, g is the speed of individual A do not adopt new product again. Assume a reaction from competitor that is performing a special treatment like advertisement, lotery and others in such a way till make someone leave population of P ($T_c(P)$), and population of A ($T_c(A)$). In the other hand, we assume that the owner do good special treatment too in such a way till occur recruitment population of P ($T_o(P)$), and also population of A ($T_o(A)$).

It will be observed the parameter conditions in which the bifurcation of codimension one and codimension two occurs. Also, it will be given the economical interpretation of bifurcation analysis. The result of the investigation is expected to give positive contribution in business field. The detail information of the dynamical behavior system hopefully is useful increasing the efficiency of productivity and marketing management.

2. BIFURCATION

It is high risk to neglect a little change if the little change imply the change of dynamical behavior of the system. Bifurcation theory is the powerfull theory for investigation about it in advanced analysis. Here is a simple way to understand the basic concept of bifurcation. The blue curves at figure 2.1 is the solution curves(orbit) $x(t)$ of the ordinary differential equation that write below the picture.

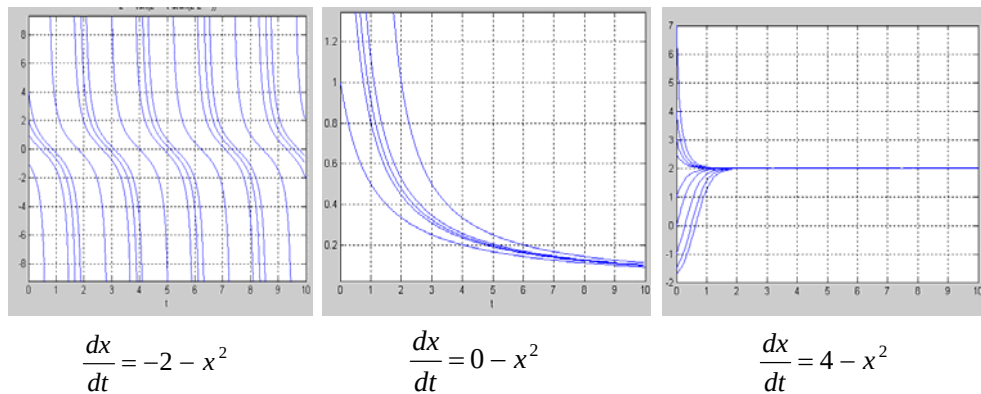


Figure 2.1. picture of the change of dynamical behavior system in time series plot under saddle-node bifurcation

It is clearly, the change of parameter value imply the change of behavior system. The first picture shows no equilibrium point. The second picture shows one equilibrium point, that is 0. The last picture shows two equilibrium points, $x = -2$ and $x = 2$, and the solution curves go away from -2 and go to 2 . It can be said that -2 is an unstable equilibrium point and 2 is a stable equilibrium point. If we find the condition, we say that saddle node bifurcation occur in the system.

:

$\frac{dx}{dt} = \mu - x^2$ is a normal form of saddle-node bifurcation. μ is called as parameter deformation because the change of the value imply the change of behavior system. Figure 2.2. shows the relation between μ and x .

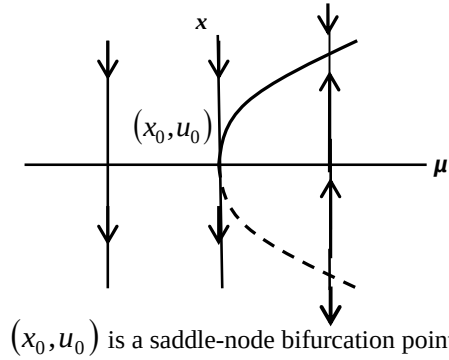


Figure 2.2. saddle-node bifurcation system dimensi one codimensi one

Now, it is shown a simple example no bifurcation occurs in a system. See figure 2.3., for $\frac{dx}{dt} = \mu - x$, the value change of parameter μ does not effect the change of behavior system. The system always has one equilibrium point and all of solution curves go to the point. We just find the change of the value of the equilibrium point, and it does not bifurcation matter.

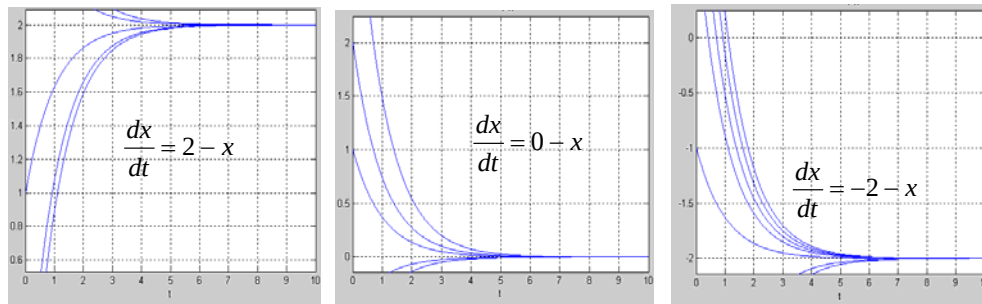


Figure 2.3. no bifurcation occur in system $\frac{dx}{dt} = \mu - x$

We just talk about one dimensional dynamical system. In a two dimensional dynamical system, we may find Hopf bifurcation. Normal form of Hopf bifurcation is,

$$\begin{cases} \dot{x} = -\text{Im}(\lambda)y + f(x, y, \alpha) \\ \dot{y} = \text{Im}(\lambda)x + g(x, y, \alpha) \end{cases}$$

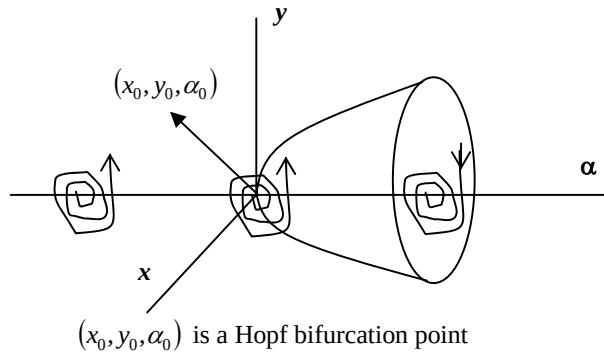
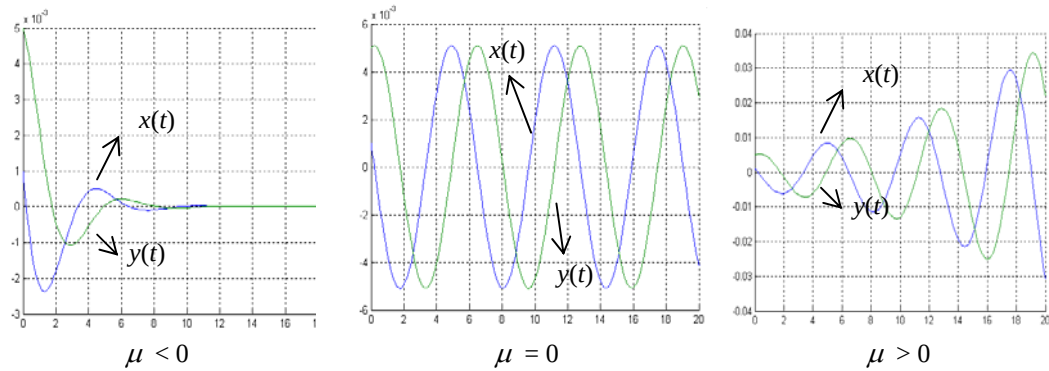


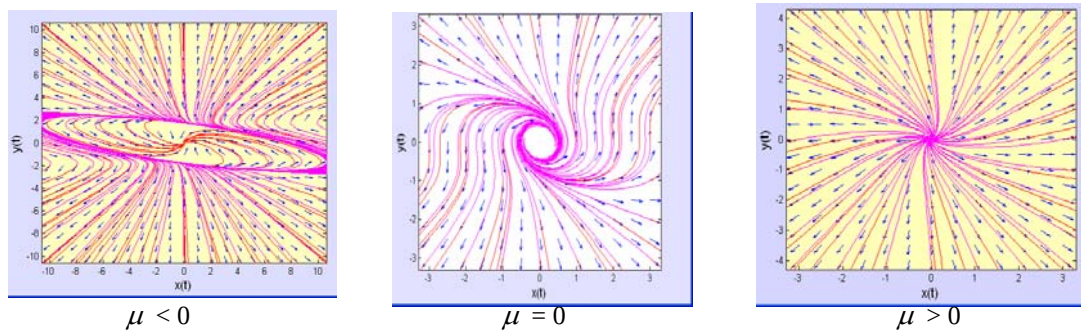
Figure 2.4. Hopf bifurcation in system dimension two codimension one

Characteristics of Hopf bifurcation are the existence of limit cycle and the change stability of the equilibrium point (see figure 2.4). Next, figure 2.5 shows a visual example Hopf bifurcation occurs in system

$$\begin{aligned}\frac{dx}{dt} &= \mu x - y + xy^2 \\ \frac{dy}{dt} &= x + \mu y + y^3\end{aligned}$$



(a). the change behavior system in time series under Hopf bifurcation



(b). the change behavior system in phase plane under Hopf bifurcation

Figure 2.5. Dynamical behavior of Hopf bifurcation

All of bifurcation that we discuss above, saddle-node and Hopf bifurcation, are codimension one bifurcation. They have one deformation parameter. Furthermore, we discuss bifurcation with two deformation parameters, that is Bogdanov-Takens bifurcation. Bogdanov-Takens bifurcation point is caused by intersection and interference saddle-node and Hopf bifurcation curve. The dynamical behavior of Bogdanov-Takens bifurcation is more complicated than saddle-node or Hopf bifurcation. See figure 2.6., it is the dynamical behavior of Bogdanov-Takens bifurcation with normal form

$$\dot{x} = y$$

$$\dot{y} = \beta_0 + \beta_1 y + x^2 + sxy + O(3)$$

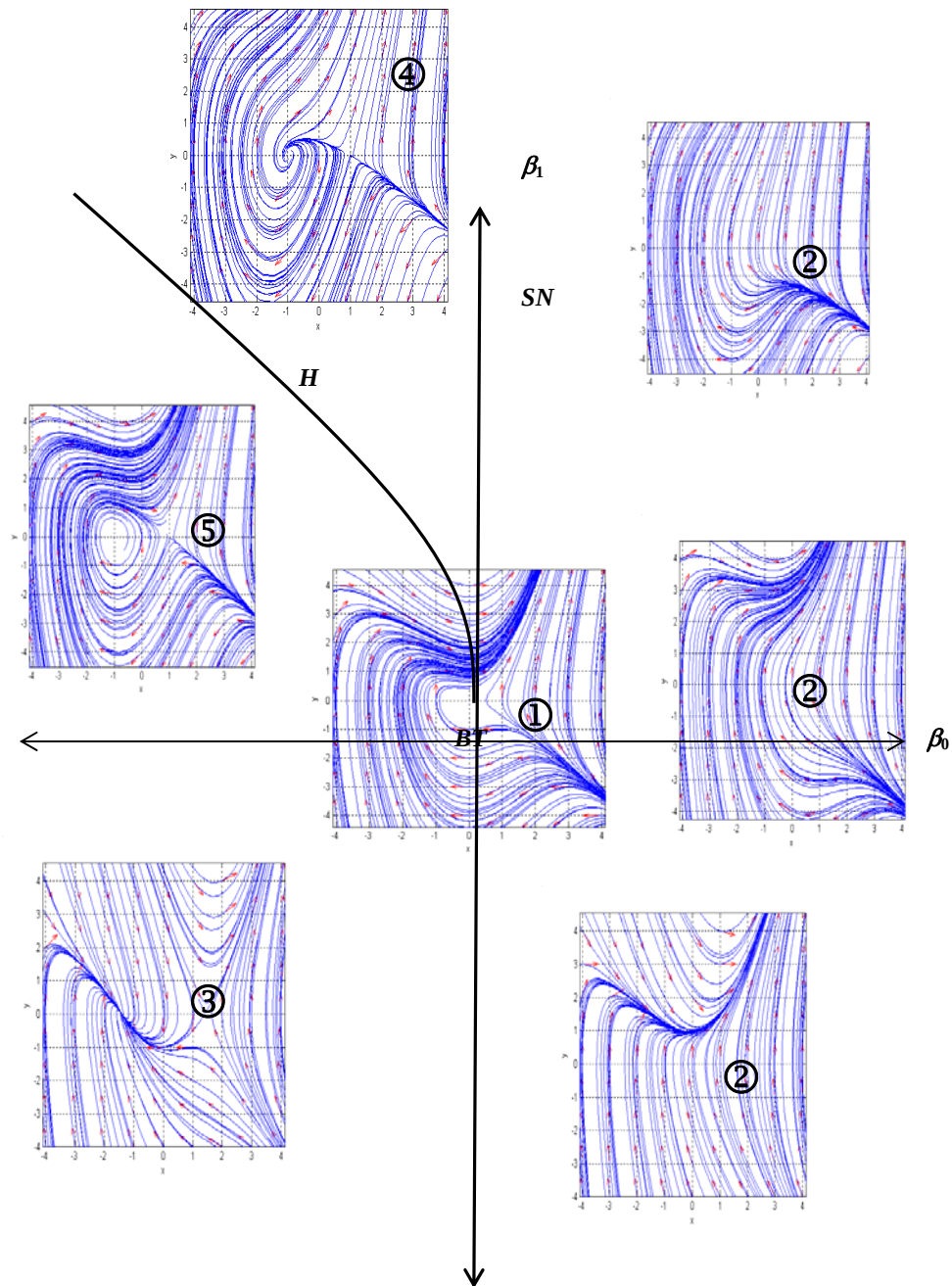


Figure 2.6. Dynamical behavior of Bogdanov-Takens bifurcation

BT is Bogdanov-Takens bifurcation point, $(x, y, \beta_0, \beta_1) = (0, 0, 0, 0)$

SN is Saddle node bifurcation curve, $\beta_0 = 0$

H is Hopf bifurcation curve, $\beta_1 = \sqrt{-\beta_0}$

- ① is potret phase in xy -plane, it is a cups
- ② is potret phase of no equilibrium point
- ③ is potret phase of one stable equilibrium point and one saddle point
- ④ is potret phase of one unstable equilibrium point and one saddle point
- ⑤ is potret phase of unstable limit cycle, with one stable equilibrium point and one saddle point

We know that we can observed local bifurcation from eigenvalues of the Jacobi of the system in the equilibrium point [6, 8]. Let we have autonomous system dimension two. Saddle-node bifurcation occurs in a system when the equilibrium point correspond to one zero eigenvalue of the Jacoby system. If we find a pair of pure imajiner eigenvalues, it means the system under Hopf bifurcation. On the other hand, if we find double zero eigenvalues, it means the system under Bogdanov-Takens bifurcation. In this paper, we will do bifurcation analysis in this way. It will be traced conditions of the parameters system such as the bifurcations is happened.

3. BIFURCATION ANALYSIS OF A NEW-PRODUCT GROWTH MODEL

In this section, we discuss dynamic model related with business problem, that is dynamical of spread of new-product in a market. The model is as mentioned in the first section (1). We assume all of the parameters are positive. Also, It is assumed the special treatments are constant, denoted by m_0 and k_0 . There are eight cases, first case: $m_0 - m_c = m > 0$, $k_0 - k_c = k > 0$ and $\beta - \alpha = -a < 0$, second case: $m_0 - m_c = m > 0$, $k_0 - k_c = k > 0$ and $\beta - \alpha = b > 0$, third case: $m_0 - m_c = -m < 0$, $k_0 - k_c = k > 0$ and $\beta - \alpha = -a < 0$, fourth case: $m_0 - m_c = -m < 0$, $k_0 - k_c = k > 0$ and $\beta - \alpha = b > 0$, fifth case: $m_0 - m_c = m > 0$, $k_0 - k_c = -k < 0$ and $\beta - \alpha = -a < 0$, sixth case: $m_0 - m_c = m > 0$, $k_0 - k_c = -k < 0$ and $\beta - \alpha = b > 0$, seventh case: $m_0 - m_c = -m < 0$, $k_0 - k_c = -k < 0$, $\beta - \alpha = -a < 0$, and the last case: $m_0 - m_c = -m < 0$, $k_0 - k_c = -k < 0$, $\beta - \alpha = b > 0$. Analysis is done at first quadrant, because P and A represent population. We get all of these theorems below.

Theorem 3.1. For $m_0 - m_c = m > 0$, $k_0 - k_c = k > 0$, $\beta - \alpha = b > 0$.

- (i) If $g - b > 0$, $b^2 - bg - kc - mc < 0$, $cA - b > 0$, $(g - b)A - m - k > 0$ and $(b^2 - bg - kc - mc)^2 - 4(g - b)cbk > 0$, model (1) has two positive equilibrium points $T_1(A_1, P_1)$ and $T_2(A_2, P_2)$. They are

$$T_{12} \left(\frac{-(b-g)b - (k+m)c \pm \sqrt{(b^2 - bg - kc - mc)^2 - 4kb(g-b)c}}{2(g-b)c}, \frac{k + (g-b)A_2 - m}{b} \right).$$

- (ii) If $g - 2b < 0$, $2b^2 - cm - ck - bg < 0$, and

:

$$k = -m - \frac{(g-2b)(g-b)}{2c} \left(1 - \sqrt{1 + \frac{4cm}{(g-b)^2}} \right), \text{ then bifurcation Hopf occurs in the } T_1.$$

Proof.

(i). We make all equations in (1) equal to zero, we get

$$P = \frac{-m}{b-cA} \quad (2)$$

$$P = \frac{(g-b)A-k}{cA} \quad (3)$$

From (2) and (3), we find

$$0 = c(g-b)A^2 + (b^2 - bg - kc - mc)A + bk. \quad (4)$$

Clearly, equation (4) has positive solution in A if $g-b > 0$ and $b^2 - bg - kc - mc < 0$. By classical computation, it is obtained

$$A_{12} = \frac{-((b-g)b - (k+m)c) \pm \sqrt{(b^2 - bg - kc - mc)^2 - 4kb(g-b)c}}{2(g-b)c},$$

$$P_{12} = \frac{k + (g-b)A_{12} - m}{b}$$

Also clearly, equation (4) has positive P if $cA - b > 0$, $(g-b)A - m - k > 0$ is fulfilled. And, we get two equilibrium point if $(b^2 - bg - kc - mc)^2 - 4(g-b)cbk > 0$.

(ii). It is proved by find a condition in which the Jacobi system has a pair of pure imaginary eigenvalues. It is shown through $\det(J) > 0$ and $\text{tr}(J) = 0$.

$$\det(J_{T_1}) = cbP - (g-b)b + c^2AP - c^2AP + c(g-b)A$$

$$= \sqrt{(b^2 - bg - kc - mc)^2 - 4kb(g-b)c} > 0.$$

$$\text{tr}(J_{T_1}) = \frac{(2b^2 - gb - mc - kc) - (2bg - g)A_1}{b},$$

substitution A_1 to $\text{tr}(J_{T_1})$, and then put $\text{tr}(J_{T_1})$ to zero, we have condition $g - 2b < 0$, $2b^2 - cm - ck - bg < 0$, and

$$k = -m - \frac{(g-2b)(g-b)}{2c} \left(1 - \sqrt{1 + \frac{4cm}{(g-b)^2}} \right).$$

The eigenvalues of J_{T_1} are

$$\lambda_{12} = \pm \sqrt{\det(J_{T_1})} i = \pm \sqrt{(b^2 - bg - kc - mc)^2 - 4kbc(g-b)} i.$$

Theorem 3.2. For $m_o - m_c = m > 0$, $k_o - k_c = -k < 0$, $\beta - \alpha = -a < 0$.

(i) If $a^2 + ga + kc - mc < 0$ and $(a^2 + ga + kc - mc)^2 - 4ka(a+g)c > 0$, model (1) has two positive equilibrium points $T_3(A_3, P_3)$ and $T_4(A_4, P_4)$. They are

$$T_{34} \left(\frac{-((a+g)a + (k-m)c) \pm \sqrt{(a^2 + ga + kc - mc)^2 - 4ka(a+g)c}}{2(a+g)c}, \frac{m - (a+g)A_{34} - k}{a} \right)$$

(ii) If $cmg - ckg - 3a^2g - g^2a - 2a^3 > 0$, and $k = m + \frac{(2a+g)(a+g)}{2c} \left(1 - \sqrt{1 + \frac{4cm}{(a+g)^2}} \right)$,

:

system is under Hopf bifurcation in T_3 .

Proof.

It is proved by the same way as Theorem 3.1.

Theorem 3.3. For $m_o - m_c = m > 0$, $k_o - k_c = -k < 0$, $\beta - \alpha = -a < 0$.

If $a^2 + ga + kc - mc < 0$ and $(a^2 + ga + kc - mc)^2 - 4ka(a + g)c = 0$, model (1)

(i) has a positive equilibrium point $T_5\left(\frac{-(a+g)a-(k-m)c}{2(a+g)c}, \frac{1}{2}\left(\frac{a+g}{c} + \frac{m-k}{a}\right)\right)$.

(ii) under Bogdanov-Takens bifurcation in the equilibrium if $cmg^2 = a(g+a)^3$

Proof.

(i) We use the same way as previous theorems. By making all of the equation in (1) equal to zero and substitution the given condition, we will get easily the equilibrium point T_5 .

(ii) Next, to prove theorem 3.3(ii). It is shown Jacobi system has double zero eigenvalues. In fact, by substitution all of given condition, we get $\det(J_{T_5}) = (-a - cA_{T_5})(cP_{T_5} - a - g) + c^2 A_{T_5} P_{T_5} = 0$. And, $\text{tr}(J_{T_5}) = -(a + cA_{T_5}) + cP_{T_5} - (a + g) = 0$. It means that Jacobi system in the equilibrium point T has double zero eigenvalues.

Theorem 3.4. For $m_o - m_c = -m < 0$, $k_o - k_c = k > 0$, $\beta - \alpha = b > 0$.

(i) If $g - b > 0$, $b^2 - bg - kc + mc < 0$, $cA - b < 0$, $(g - b)A - k > 0$ and $(b^2 - bg - kc + mc)^2 - 4(g - b)cbk > 0$, model (1) has two positive equilibrium points $T_6(A_6, P_6)$ and $T_7(A_7, P_7)$. They are

$$T_{67}\left(\frac{-((b-g)b-(k-m)c) \pm \sqrt{(b^2 - bg - kc + mc)^2 - 4kb(g-b)c}}{2(g-b)c}, \frac{m+(g-b)A_6-k}{b}\right).$$

(ii) If $g - 2b > 0$, $2b^2 + cm - ck - bg > 0$, and $k = m - \frac{(g-2b)(g-b)}{2c} \left(1 - \sqrt{1 - \frac{4cm}{(g-b)^2}}\right)$, then bifurcation Hopf occurs in the T_6 .

Proof.

It is proved by the same way as Theorem 3.1.

Theorem 3.5. For $m_o - m_c = -m < 0$, $k_o - k_c = -k > 0$, $\beta - \alpha = b > 0$.

(i) If $g - b < 0$, $b^2 - bg + kc + mc > 0$, $b - cA > 0$, $(g - b)A + k > 0$ and $(b^2 - bg + kc + mc)^2 + 4(g - b)cbk > 0$, model (1) has two positive equilibrium points $T_8(A_8, P_8)$ and $T_9(A_9, P_9)$. They are

$$T_{89}\left(\frac{(b^2 - gb + kc + mc) \pm \sqrt{(b^2 - bg + kc + mc)^2 + 4kb(g-b)c}}{-2(g-b)c}, \frac{m+(g-b)A_8+k}{b}\right).$$

:

- (ii) If $g - 2b > 0$, $2b^3 - cg(m + k) + bg(g - 3b) > 0$, and $k = -m + \frac{(g-2b)(g-b)}{2c} \left(1 - \sqrt{1 - \frac{4cm}{(g-b)^2}} \right)$, then bifurcation Hopf occurs in the T_8 .

Proof.

It is proved by the same way as Theorem 3.1.

Theorem 3.6. For $m_o - m_c = -m < 0$, $k_o - k_c = -k > 0$, $\beta - \alpha = b > 0$.

If $g - b < 0$, $b^2 - bg + kc + mc > 0$, $b - cA > 0$, $(g - b)A + k > 0$ and $(b^2 - bg + kc + mc)^2 + 4(g - b)cbk = 0$, model (1)

- (i) has a positive equilibrium point $T_{10} \left(\frac{(b^2 - gb + kc + mc)}{-2(g-b)c}, \frac{m + (g-b)A_0 + k}{b} \right)$
- (ii) under Bogdanov-Takens bifurcation in the equilibrium if $cmg^2 = b(g - b)^3$

Proof.

We use the same way as theorem 3.3. to prove this condition.

4. ECONOMICAL INTERPRETATION AND NUMERICAL EXAMPLES

As mentioned in earlier, bifurcation means the change behavior, i.e., the change amount of the equilibrium points, the change of stability of the equilibrium points, appear limit cycle and others. For stable equilibrium point, we interpret that always any adoption of the new product for infinite time. On the other hand, unstable equilibrium point, we interpret that no adoption the new product for infinite time. Thus, we always expect to find stable equilibrium point. Now, let we find a condition in which equilibrium corresponding to a pair of pure imaginary eigenvalues. It means that if we change one of the parameter system then we can change the stability of the equilibrium point of the system. Thus, we can change the owner special treatment parameter such as the unstable equilibrium point becomes a stable one. For example, input $g=1; b=.2; c=.01; m=2; k=0.449944$ to the model (1). The value of all parameters fullfill condition of theorem 3.4. We find one of the equilibrium point in which the eigenvalues of the Jacobi system is $\pm .366i$. By this condition, we know, if $m > 2$, the equilibrium becomes unstable equilibrium point. It means that the extinction of adoption the new product occurs at the system. We do not expect this condition. Thus, we can change the owner treatment such as $m < 2$. We will get the stable equilibrium (let choose $m = 1$, we get eigenvalues of the Jacobi system is $-.006 \pm .38i$). It means that the persistence of adoption the new product occurs at the system.

Further more, let Bogdanov-Takens bifurcation occurs in the new-product growth model. We find a little change of the owner and or competitor special treatment (advertising, lottrey), imply the change of behavior system. It is possible system become no equilibrium point, it means no adoption of new-product. However, when we change again the value, it can be found subcritical limit cycle (stable equilibrium point), it means inside limit cycle any adoption of new product. Little change again, we will find a stable equilibrium point, it :

means persistence of adoption the new product. Also, we can find the unstable equilibrium, it means extinction of adoption the new product at the system. Here an example, choose $c=1$; $g=1$; $a=.5$; $m=1.6875$; $k=.1875$. It is fulfilled theorem 3.3. Under the condition we get cups curve ①. See the potret phase in figure 4.1. By fixed c, g, a , we get limit cycle ② when we change the value of m and k by $m^*=.05$; $k^*=.007$. There is an unstable equilibrium ③ for $m^*=.05$; $k^*=.01684782609$. And, no equilibrium point ④ with $m^*=.2$; $k^*=.5$. On the other hand, we find a stable equilibrium point ⑤ for $m^*=.4$; $k^*=.001$. Thus, we can treat two of parameters model (1) such as we obtain our expected condition.

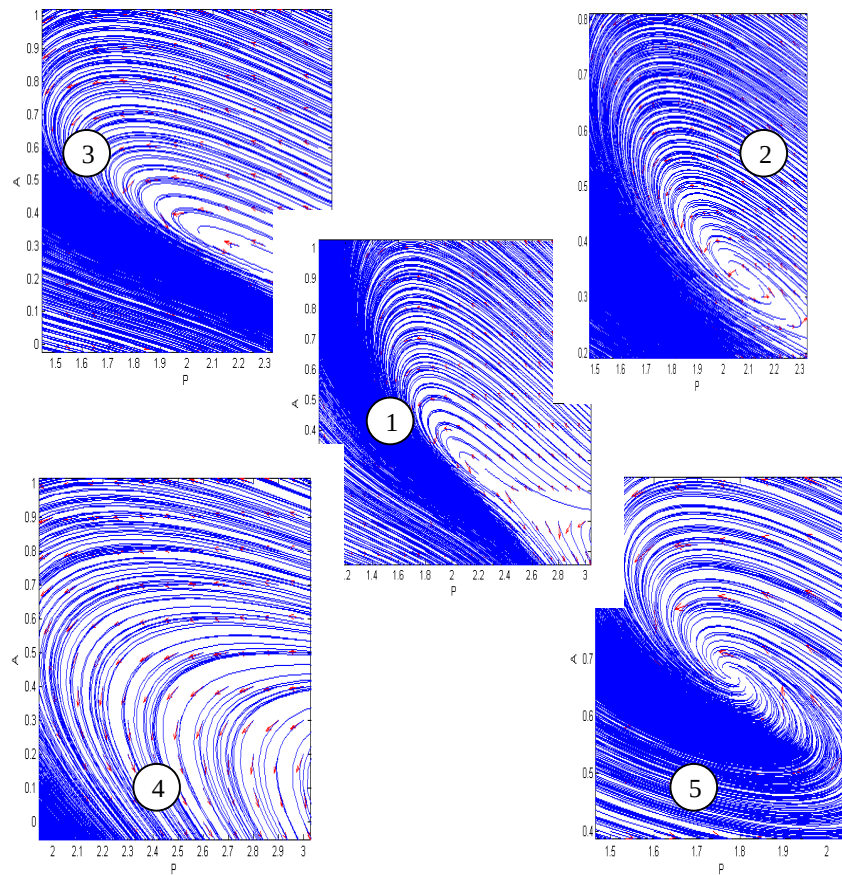


Figure 4.2 Example qualitative behavior of Bogdanov-Takens bifurcation model (1) under certain condition

- ① sign that little change will give significance change of the behavior system
- ② persistence of adoption the new product inside the limit cycle
- ③ extinction of adoption the new product at the system
- ④ extinction of adoption the new product at the system
- ⑤ always any adopter of new product for infinite time.

5. CONCLUDING REMARKS

We can get a clear picture of dynamical behaviour the new-product growth model by doing bifurcation analysis. It is expected the result of the investigation give positive contribution to business field. The detail information of the dynamical behavior system hopefully give a picture and a direction to decision maker in a firm such as it is useful increasing the efficiency of productivity and marketing management.

REFERENCES

1. Aracil, J. (1999). On the qualitative properties in system dynamics models. *European Journal of Economic and Social Systems*.
2. Dorn, J. F. C. (2002) *Nonlinear Dynamics in Homogeneous Azeotropic Distillation*, Dissertation Dr. sc. techn, Swiss Federal Institute of Technology, Zurich.
3. Juliana, J.R. (2007) *The Phenomena of A New-Product Growth with Special treatments*. Proceeding ICREM3, Malaysia.
4. Jin, Y., Wang, W. and Xiao, S. (2006) An SIRS model with a nonlinear incidence rate, *Chaos Solitons & Fractals*.
5. Sterman, J. (2004) *Business Dynamics: Systems Thinking and Modeling for a Complex World*. Mc Graw Hill.
6. Voorn, G. A. K. (2006) *PhD mini course: Introduction to Bifurcation Analysis*.
7. Wang, W, and Ruan, S. (2004) Bifurcations in an epidemic model with constant removal rate of infectives, *Journal of Mathematical Analysis and Applications*.
8. Wiggins, S. (1990) *Introduction to Applied Nonlinear Dynamical System and Chaos*, Springer-Verlag.
9. Xiao, D. et all (2001) Codimension Two Bifurcations in A Predator-Prey System with Group Defense, *Fields Institute Communications*.

Joice Ruth Juliana : Department of Mathematics and Natural Sciences,
University of Surabaya.
Email: joiceruth@yahoo.com

M. Setijo Winarko: Department of Mathematics, Institute of Technology
Sepuluh Nopember

A CONTEXTUAL APPROACH, USING PROBLEM-BASED, TO TEACHING AND LEARNING DIFFERENTIAL EQUATIONS IN THE UNIVERSITI TUN HUSSEIN MALAYSIA

NAFISAH KAMARIAH MD KAMARUDDIN AND
WAN MOHD RASHID WAN AHMAD

Abstract. Mathematics is often considered as a difficult subject in schools or universities and most students have difficulty in understanding the abstract concept in mathematics. In addition, there are many complains in Malaysia that certain fresh graduated participants who attended work interview are not able to express their ideas well even though they get good grades. Thus, the Mathematics Department, Science Studies Centre in UTHM has adopted the contextual teaching and learning approach in order to train students to make connections between what they are learning in mathematics with their engineering subjects. In the contextual approach, by using problem-based learning (PBL), each group of the students must present their ideas and solutions. Thus this research was done to study whether the contextual approach, using PBL, helped the second year technical education students who took differential equations to understand the mathematics concept and also trained them to present their ideas or concept in front of the class. A quasi-experimental research design was used in this study. The students were divided into three groups: control group, fresh PBL group and experienced PBL group. The control group followed the traditional method of teaching and learning, the fresh PBL group has just gone through the PBL method of teaching and learning and the experienced PBL group had undergone PBL training from the previous semester. Four types of instruments were designed and used for this study: questionnaires, students' presentations and 2 quizzes. The first topic that the students learned was first order differential equation. Thus the first quiz focused on the application of differential equation in the Newton's Law of Cooling and the second quiz focused on solving Newton's Law of Cooling problem. For the two quizzes, both PBL groups scored better than the control group. *One-Way Anova* was used and from the findings, there are significance differences between the control group and the fresh PBL, and also between the control group and the experienced PBL group but there is no significant difference between the fresh PBL group and the experienced PBL group for the two quizzes. Even though the student felt that the PBL method took them a longer time to understand a concept and they found too much information that make them confused, both PBL groups agreed that the PBL method helped them to be independent as well as practiced them to work as a team, motivated them to learn mathematics and they scored better in both quizzes. Lastly, the experienced PBL group gave smoother and clearer presentations and more confident when presenting compare to the fresh PBL group.

Key words and Phrases: Contextual, problem-based learning (PBL) , differential equations

1. INTRODUCTION

Mathematics is often considered as a difficult subject in schools or universities. Most students learn it very mechanistic, that is, memorizing the

formula and solving problems using the formula without knowing its application in the real world. In addition, the students have difficulty in understanding the abstract concept in mathematics. Another issue in Malaysia is that there are many complains that certain fresh graduated participants who attended work interview were not able to express their ideas well even though they got good grades. Thus, the Mathematics Department, Science Studies Centre in UTHM has adopted the contextual teaching and learning approach in order to train students to make connections between what they are learning in mathematics with their engineering subjects. In the contextual approach, watching the video and doing the lab practical helped students to connect what they learned in the class with the real world. For the lab practical, by using the problem-based learning (PBL), students engaged in problem-solving investigation that integrates skills and concept from many content areas. Once they knew the concept and able to combine all the knowledge that had been taught, they should be able to apply it in their projects, researches and future career. Two of the guidelines of contextual approach by the Center for Occupational Research and Development (CORD) that were focused in this research were learning connects with the real world and learning involves interaction with others (First Malaysian Tech Prep National Convention, 1997). In the contextual approach, by using PBL, each group of the students must present their ideas and solutions. This was important as there were students who were afraid to speak up as they have no confidence and sure of their opinions and hence they would not be good decision makers. Thus this research was done to study whether the contextual approach, using PBL, helped the second year technical education students who took differential equations to understand the mathematics concept and also trained them to present their ideas or concept in front of the class.

2. LITERATURE REVIEW

There are many definitions of the contextual learning. The definition that the Malaysian Technical Education Department uses is the one that was given by (First Malaysian Tech Prep National Convention, 1997):

The Contextual learning means learning that incorporates examples which are drawn from everyday experiences in the personal, societal and occupational life and which provide concrete hands-on applications of material to be learned.

According to CORD, one of the key elements in the contextual approach is to carry out learning in workplace setting, where possible. (First Malaysian Tech Prep National Convention, 1997). According to CORD, workplace and lab activities created by business and industry focus on measurement and computation, and help students to develop critical thinking skills (First Malaysian Tech Prep National Convention, 1997). It will be great if the students can relate the formulas and theories that are taught in the classroom to their everyday lives or their future jobs. Is not easy to bring students to the real life atmosphere, let say a company. Thus if we cannot bring them to the companies, we need to simulate the workplace. In the lab practical, besides making them understand the concept, the students also work in environment or group projects that simulate the workplace. This is a good practice because there is a lot of complain from managers from companies that not all students who graduated from higher learning situation can perform well in the workplace. In the contextual approach, by using PBL, students engage in problem-solving

investigation that integrate skills and concept from many content areas or by using the project-based learning, students works autonomously to construct their own learning, and culminate in realistic products. (Berns et al, 2001).

There are also many definitions of the problem-based learning (PBL). Fogarty defined PBL as a curriculum model designed around real-life problems that are ill-structured, open-ended, or ambiguous. In problem based learning students go through many stages which are defining the problem, gathering known facts, generating questions, hypothesizing, eventually generating alternatives, advocating solutions and justifying recommendations (Fogarty, 1997).

Many researchers have found out that PBL is very effective and helps the students, not only in understanding the concepts but also in developing themselves. In the study by Acar et al (2003), they found that PBL provided the basis for the development of systems engineers. Sometimes students are not able to connect what they learn in class with the real world. PBL will be a good method of learning as the creation of most PBL classes is built around the delivery of the real-world learning experience to the students (Cockayne et al, 2003). Traditional method is a one-way method of teaching and the students are expected to just accept what the teacher teaches. However in PBL, there is a connection between the teaching method and the depth and complexity of the learning, as the students may be expected to reach the level of analytically complex comprehension through the problem-based work that is not done in the traditional method of teaching (Graaff et al, 2003).

The success of this method will depend on many aspects such as the experience of the lecturers, the implementation of PBL, the evaluations of the students' performances, etc. This variation in PBL design would occur as the use of PBL is increasing (Zolin et al, 2003). Lecturers should be aware of the problems that may occur when implementing PBL. For example, the students would get a lot of information all at once and found difficulty to plow through the concepts or the introductory application with less theory makes the computation more difficult (Nasr et al, 2004).

3. RESEARCH METHODOLOGY

A quasi-experimental research design was used in this study. The students were divided into three groups: control group, fresh PBL group and experienced PBL group. The control group followed the traditional method of teaching and learning, the fresh PBL group has just gone through the PBL training and the experienced PBL group had undergone PBL training from the previous semester. Four types of instruments were designed and used for this study: questionnaires, students' presentations and 2 quizzes. The first topic that the students learned was first order differential equation. Thus the first quiz focused on the application of differential equation in the Newton's Law of Cooling and the second quiz focused on solving Newton's Law of Cooling problem.

Researchers designed questionnaires which are divided into four main sections : Background, Problem-based question (agree and aganist) , Problem Solving Method (Polya) and Motivation. Five level of Likert scale was used as a form of measurement. Researchers analyzed data in section B, C and D through mean score based on the table which was adapted from *Wiersma* (2000) as shown in Table 1.

TABLE 1

Mean Score Interval	Interpretation	Level
1.00 – 2.49	Do not agree	Low
2.50 – 3.49	Not totally agree	Average
3.50 – 5.00	Agree	High

The research hypothesis was as follows:

- H_0 : There is no significant difference between the quiz performances of the control group and the treated group
- H_1 : There is significant difference between the quiz performances of the control group and the treated group

4. MAIN RESULTS

From the section A of the questionnaire, the respondents background is shown in Table 2.

TABLE 2

	Category	No of students	Percent (%)
Gender	Male	18	29.0
	Female	44	71.0
Major	Civil	23	37.1
	Electrical	13	21.0
	Mechanical	26	41.9
CGPA	>3.50	22	35.5
	3.00-3.49	38	61.3
	2.50-2.99	2	3.2

In Section B, the researchers wanted to survey whether the students agree the method of learning mathematics using the PBL. There were ten items in Section B(i) and the overall value of the mean score is 4.21 ($\sigma = 0.76$) which is high. Secondly, the researchers wanted to survey whether the students the against the method of learning mathematics using the PBL. There were ten items in Section B(ii) and the overall value of the mean score is 3.58 which is high, however the $\sigma = 1.02 > 1$. In section C, the researchers wanted to survey whether the student solved the PBL problem using problem solving method. There were ten items in Section C and the overall value of the mean score is 3.99 ($\sigma = 0.70$) which is also high. Lastly the researcher wanted to survey whether the students are motivated to learn mathematics using PBL. There were ten items in Section D and the overall value of the mean score is 4.04 which is high and the $\sigma = 0.63$. All the results in section, B, C and D are given in Table 3.

TABLE 3

Section	Category	Mean score	Standard deviation	Interpretation
B (i)	Problem-based question (agree)	4.21	0.76	High

B (ii)	Problem-based question (against)	3.58	1.02	High
C	Problem Solving Method (Polya)	3.99	0.70	High
D	Motivation	4.04	0.63	High

From the two quizzes, both PBL groups scored better than the control group. For the conceptual quiz, the experienced PBL group got the highest mean score which was 8.767 ($\sigma = 1.3047$) while the mean score for the control group was only 4.844 ($\sigma = 2.8411$) and the fresh PBL group was 7.875 ($\sigma = 1.4085$). As for the procedural quiz, the fresh PBL group got the highest mean score which was 8.203 ($\sigma = 2.1132$) while the mean score for the experienced PBL group was 7.050 ($\sigma = 3.1605$) and the control group was only 4.969 ($\sigma = 2.6577$). All the results are given in Table 5.

TABLE 5

	Group	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
						Lower Bound	Upper Bound
conceptual quiz	Exp	30	8.767	1.3047	.2382	8.279	9.254
	control	32	4.844	2.8411	.5022	3.819	5.868
	fresh	32	7.875	1.4085	.2490	7.367	8.383
	Total	94	7.128	2.5954	.2677	6.596	7.659
procedural quiz	Exp	30	7.050	3.1605	.5770	5.870	8.230
	control	32	4.969	2.6577	.4698	4.011	5.927
	fresh	32	8.203	2.1132	.3736	7.441	8.965
	Total	94	6.734	2.9673	.3061	6.126	7.342

From the *One-Way Anova* analysis, there is a significant difference between the performances between the three groups since $p < 0.05$ as shown in Table 6.

Table 6

		Sum of Squares	df	Mean Square	F	Sig.
conceptual quiz	Between Groups	265.383	2	132.691	33.441	.000
	Within Groups	361.085	91	3.968		
	Total	626.468	93			
procedural quiz	Between Groups	171.778	2	85.889	12.079	.000
	Within Groups	647.073	91	7.111		
	Total	818.851	93			

Referring to Table 7, there are significance differences between the control group and the fresh PBL ($p = 0.00 < 0.05$), and also between the control group and the experienced PBL group ($p = 0.00 < 0.05$). However, there is no significant difference between the fresh PBL group and the experienced PBL group for the two quizzes ($p = 0.188 > 0.05$).

Table 7

Dependent Variable		(I) Group	(J) Group	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
conceptual quiz	Tukey HSD	exp	control	3.923(*)	.5062	.000	2.717	5.129
			fresh	.892	.5062	.188	-.314	2.098
		control	exp	-3.923(*)	.5062	.000	-5.129	-2.717
			fresh	-3.031(*)	.4980	.000	-4.218	-1.845
		fresh	exp	-.892	.5062	.188	-2.098	.314
			control	3.031(*)	.4980	.000	1.845	4.218
	LSD	exp	control	3.923(*)	.5062	.000	2.917	4.928
			fresh	.892	.5062	.082	-.114	1.897
		control	exp	-3.923(*)	.5062	.000	-4.928	-2.917
			fresh	-3.031(*)	.4980	.000	-4.020	-2.042
procedural quiz	Tukey HSD	exp	control	2.081(*)	.6777	.008	.467	3.696
			fresh	-1.153	.6777	.210	-2.768	.462
		control	exp	-2.081(*)	.6777	.008	-3.696	-.467
			fresh	-3.234(*)	.6666	.000	-4.823	-1.646
		fresh	exp	1.153	.6777	.210	-.462	2.768
			control	3.234(*)	.6666	.000	1.646	4.823
	LSD	exp	control	2.081(*)	.6777	.003	.735	3.427
			fresh	-1.153	.6777	.092	-2.499	.193
		control	exp	-2.081(*)	.6777	.003	-3.427	-.735
			fresh	-3.234(*)	.6666	.000	-4.559	-1.910
		fresh	exp	1.153	.6777	.092	-.193	2.499
			control	3.234(*)	.6666	.000	1.910	4.559

* The mean difference is significant at the .05 level.

5. CONCLUDING REMARKS

Even though the student felt that the PBL method took them a longer time to understand a concept and they found too much information that make them confused, both PBL groups agreed that the PBL method helped them to be independent as well as practiced them to work as a team, motivated them to learn mathematics and they scored better in both quizzes. Lastly, the experienced PBL group gave smoother and clearer presentations and more confident when presenting compare to the fresh PBL group.

REFERENCES

1. Acar, B. S. and Newman, I. A. Students as Tutors – Learning Problem-Solving Skills by Tutoring PBL. *International Journal Engineering Education*, Vol, 19, No.5, pp 712-716, 2003.
2. Berns, Robert G and Erikson, Patricia. Contextual Teaching and Learning. The Highlight Zone : Research @ Work No . 5. *The National Centers for Career and Technical Education*, USA, 2001.
3. Cockayne, W., Feland III, J. M. and Leifer L. Using the Contextual Skills Matrix for PBL Assessment. *International Journal Engineering Education*, Vol, 19, No.5, pp 701-705, 2003.
4. Graaff, E. and Kolmos, A. Characteristics of Problem-Based Learning. *International Journal Engineering Education*, Vol, 19, No.5, pp 657-662, 2003.
5. First Tech Prep National Convention, organized by Technical Education Department, Malaysian Ministry of Education in September, 1997.
6. Fogarty, Robin. *Problem-based Learning & Other Curriculum Models for the Multiple Intelligences Classroom*. SkyLight Training and Publishing Inc., 1997.
7. Nasr, K and Thomas, C. Student-centered, Concept-embedded Problem-based Engineering Thermodynamics. *International Journal Engineering Education*, Vol, 20, No.4, pp 660-670, 2004.
8. Zolin, R., Fruchter, R. and Levitt, R. Realism and Control: Problem-Based Learning Programs as a Data Source for Work-Related Research. *International Journal Engineering Education*, Vol, 19, No.6, pp 788-798, 2003.

Nafisah Kamariah MD Kamaruddin: Universiti Tun Hussein Onn Malaysia (Uthm)

E-mail: nafisah@uthm.edu.my

Wan Mohd Rashid Wan Ahmad: Universiti Tun Hussein Onn Malaysia (Uthm)

E-mail: mdrashid@uthm.edu.my

NONLINEAR MHD MIXED CONVECTION FLOW AND HEAT AND MASS TRANSFER OF AN ELECTRICALLY CONDUCTING VARIABLE VISCOSITY OVER A POROUS WEDGE

R. KANDASAMY, I. HASHIM, MUHAJIMIN AND RUHAILA

Abstract: *In the present study, an analysis is carried out to study the variable viscosity effects on MHD flow, heat and mass transfer characteristics in a viscous fluid over a porous wedge. The wall of the wedge is embedded in a uniform Darcian porous medium in order to allow for possible fluid wall suction or injection. The governing boundary layer equations are written into a dimensionless form by similarity transformations. The transformed coupled nonlinear ordinary differential equations are solved numerically by using R.K.Gill and shooting methods. The effects of different parameters on the dimensionless velocity, temperature and concentration profiles are shown graphically. Comparisons with previously published works are performed and excellent agreement between the results is obtained. The results are presented graphically and the conclusion is drawn that the flow field and other quantities of physical interest are significantly influenced by these parameters.*

1. INTRODUCTION

In many transport processes in nature and in industrial applications in which heat and mass transfer with heat source is a consequence of buoyancy effects caused by diffusion of heat and chemical species. The study of such processes is useful for improving a number of chemical technologies, such as polymer production and food processing. In nature, the presence of pure air or water is impossible. Some foreign mass may be present either naturally or mixed with the air or water. The present trend in the field of chemical reaction with viscosity analysis is to give a mathematical model for the system to predict the reactor performance. A large amount of research work has been reported in this field. In particular, the study of chemical reaction, heat and mass transfer with heat source is of considerable importance in chemical and hydrometallurgical industries.

The effect of the presence of foreign mass on the free convection flow past a semi – infinite vertical plate was studied by Gebhart and Pera[1]. The presence of a foreign mass in air or water causes some kind of chemical reaction. During a chemical reaction between two species, heat is also generated [2]. In most of cases of chemical reaction, the reaction rate depends on the concentration of the species itself. A reaction is said to be first order if the rate of reaction is directly proportional to concentration itself [3]. The effects of heat and mass transfer on laminar boundary layer flow over a wedge have been studied by many authors [4-14] in different situations. Chemical reaction effects on heat and mass transfer laminar boundary layer flow have been discussed by many authors [15-18] in various situations. The previous studies are based on the constant physical properties of the fluid. For most realistic fluids, the viscosity shows a rather

pronounced variation with temperature. It is known that the fluid viscosity changes with temperature[19]. Then it is necessary to take into account the variation of viscosity with temperature in order to accurately predict the heat transfer rates. The effect of temperature-dependent viscosity on the mixed convection flow from vertical plate is investigated by several authors[19-22].

NOMENCLATURE			
u, v	velocity components in x and y direction.	U	Flow velocity away from the wedge.
g	Acceleration due to gravity.	β^*	Coefficient of volume expansion.
K	Permeability of the porous medium	Pr	Prandtl number
λ	Porous medium parameter	T	Temperature of the fluid.
T_w	Temperature of the wall.	T_∞	Temperature far away from the wall.
β	Coefficient of thermal expansion	δ	Heat source
C	Species concentration of the fluid.	C_w	Species concentration along the wall.
C_∞	Species concentration away from the wall.	ρ	Density of the fluid.
σ	Electric conductivity of the fluid.	α	Thermal diffusivity

The aim of this work is to study the effects of variable viscosity, heat and mass transfer on nonlinear MHD mixed convection flow over a porous wedge in the presence of suction or injection. It is hoped that the results obtained will not only provide useful information for applications, but also serve as a complement to the previous studies.

2. MATHEMATICAL ANALYSIS

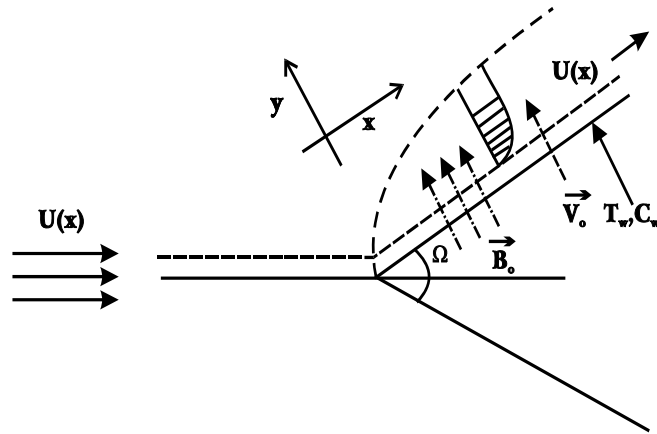


Fig.1 Flow analysis along the wall of the wedge

Let us consider a steady, laminar, hydromagnetic coupled heat and mass transfer by mixed convection flow in front of a stagnation point on a wedge plate embedded in porous medium. The fluid is assumed to be Newtonian, electrically conducting and its property variations due to temperature are limited to density and viscosity. The density variation and the effects of the buoyancy are taken into account in the momentum equation (Boussinesq's approximation) and the concentration of species far from the wall, C_∞ , is infinitesimally small [2]. In addition, there is no applied electric field and all of the Hall effects and Joule heating are neglected. Let the x-axis be taken along the direction of the wedge and y-axis normal to it. An uniform transverse magnetic

field of strength B_0 is applied parallel to the y -axis. The chemical reaction are not taking place in the flow and a constant suction or injection is imposed at the wedge surface, see Fig.1. Since the magnetic Reynolds number is very small for most used in industrial applications, we assume that the induced magnetic field is negligible. For steady, two-dimensional flow under Boussinesq's approximation including variable viscosity are

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0 \quad (1)$$

$$u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} = \frac{1}{\rho} \frac{\partial}{\partial y} \left(\mu \frac{\partial u}{\partial y} \right) + U \frac{dU}{dx} - \frac{\sigma B_0^2}{\rho} (u - U) \\ + [g\beta(T - T_\infty) + g\beta^*(C - C_\infty)] \sin \frac{\Omega}{2} - \frac{\nu}{K} (u - U) \quad (2)$$

$$u \frac{\partial T}{\partial x} + v \frac{\partial T}{\partial y} = \alpha \frac{\partial^2 T}{\partial y^2} \quad (3)$$

$$u \frac{\partial C}{\partial x} + v \frac{\partial C}{\partial y} = D \frac{\partial^2 C}{\partial y^2} \quad (4)$$

The boundary conditions are,

$$u = 0, v = -v_0, T = T_w, C = C_w \text{ at } y = 0 \quad (5)$$

$$u = U(x), T = T_\infty, C = C_\infty \text{ at } y \rightarrow \infty \quad (6)$$

Following the lines of Kafoussias et.al., [5], the following change of variables are introduced

$$\psi = \sqrt{\frac{2U\nu x}{1+m}} f(x, \eta) \quad (7)$$

$$\eta = y \sqrt{\frac{(1+m)U}{2\nu x}} \quad (8)$$

The viscosity is assumed to be an inverse linear function of temperature given by the following [24]

$$\frac{1}{\mu} = \frac{1}{\mu_a} [1 + \chi(T - T_a)] \quad (9)$$

where μ_a is the ambient fluid dynamic viscosity and χ is a thermal property of the fluid. Equ.(9) can be written as follows

$$\frac{1}{\mu} = a(T - T_r) \quad (10)$$

where $a = \frac{\chi}{\mu_a}$ and $T_r = T_a - \frac{1}{\chi}$ are constants and their values depend on the reference state and the thermal property of the fluid.

Under this consideration, the potential flow velocity can be written as

$$U(x) = A x^m, \quad \beta_1 = \frac{2m}{1+m} \quad (11)$$

where A is a constant and β_1 is the Hartree pressure gradient parameter that corresponds to

$\beta_1 = \frac{\Omega}{\pi}$ for a total angle Ω of the wedge.

The continuity equation (1) is satisfied by the stream function ψ defined by

$$u = \frac{\partial \psi}{\partial y} \text{ and } v = -\frac{\partial \psi}{\partial x} \quad (12)$$

To transform Eqs. (2), (3) and (4) into a set of ordinary differential equations, we introduce the following dimensionless parameters and variables,

$$\theta = \frac{T - T_\infty}{T_w - T_\infty} \quad (13)$$

$$\phi = \frac{C - C_\infty}{C_w - C_\infty} \quad (14)$$

$$Gr_1 = \frac{\nu g \beta (T_w - T_\infty)}{U^3} \quad (\text{Grashof number}) \quad (15)$$

$$Gc_1 = \frac{\nu g \beta^* (C_w - C_\infty)}{U^3} \quad (\text{Modified Grashof number}) \quad (16)$$

$$N = \frac{\beta^* (C_w - C_\infty)}{\beta (T_w - T_\infty)} \quad (\text{Buoyancy ratio}) \quad (17)$$

$$Re_x = \frac{Ux}{\nu} \quad (\text{Reynolds number}) \quad (18)$$

$$Pr = \frac{\nu}{\alpha} \quad (\text{Prandtl number}) \quad (19)$$

$$Sc = \frac{\nu}{D} \quad (\text{Schmidt number}) \quad (20)$$

$$M^2 = \frac{\sigma B_0^2}{\rho A} \quad (\text{magnetic parameter}) \quad (21)$$

$$S = v_0 \sqrt{\frac{(1+m)x}{2\nu U}} \quad (\text{suction or injection parameter}) \quad (22)$$

$$\lambda = \frac{\alpha}{KA} \quad (\text{porous medium parameter}) \quad (23)$$

Now the equations (2) to (4)

$$\begin{aligned}
 (\theta - \theta_r) \frac{\partial^3 f}{\partial \eta^3} = & \frac{(\theta - \theta_r)^2}{\theta_r} \left[-f \frac{\partial^2 f}{\partial \eta^2} - \frac{2m}{1+m} \left(1 - \left(\frac{\partial f}{\partial \eta} \right)^2 \right) - \frac{2}{1+m} \frac{N\phi + \theta}{1+N} Gr Re_x \sin \frac{\Omega}{2} \right. \\
 & \left. + \frac{2}{m+1} \lambda \left(\frac{\partial f}{\partial \eta} - 1 \right) + \frac{2x}{1+m} \left(\frac{\partial f}{\partial \eta} \frac{\partial^2 f}{\partial x \partial \eta} - \frac{\partial f}{\partial x} \frac{\partial^2 f}{\partial \eta^2} \right) + \frac{2x}{m+1} \frac{\sigma B_0^2}{\rho U} \left(\frac{\partial f}{\partial \eta} - 1 \right) \right] + \frac{2}{m+1} \frac{\partial \theta}{\partial \eta} \frac{\partial^2 f}{\partial \eta^2}
 \end{aligned} \quad (24)$$

$$\frac{\partial^2 \theta}{\partial \eta^2} = -Pr f \frac{\partial \theta}{\partial \eta} + \frac{2Pr}{1+m} \theta \frac{\partial f}{\partial \eta} + Pr \frac{2x}{1+m} \left(\frac{\partial f}{\partial \eta} \frac{\partial \theta}{\partial x} - \frac{\partial f}{\partial x} \frac{\partial \theta}{\partial \eta} \right) \quad (25)$$

$$\frac{\partial^2 \phi}{\partial \eta^2} = -Sc f \frac{\partial \phi}{\partial \eta} + \frac{2Sc}{1+m} \phi \frac{\partial f}{\partial \eta} + \frac{2xSc}{1+m} \left(\frac{\partial f}{\partial \eta} \frac{\partial \phi}{\partial x} - \frac{\partial f}{\partial x} \frac{\partial \phi}{\partial \eta} \right) \quad (26)$$

The boundary conditions can be written as

$$\begin{aligned}
 \eta = 0: \quad & \frac{\partial f}{\partial \eta} = 0, \quad \frac{f}{2} \left(1 + \frac{x}{U} \frac{dU}{dx} \right) + x \frac{\partial f}{\partial x} = -v_0 \sqrt{\frac{(1+m)x}{2\nu U}}, \quad \theta = 1, \quad \phi = 1 \\
 \eta \rightarrow \infty: \quad & \frac{\partial f}{\partial \eta} = 1, \quad \theta = 0, \quad \phi = 0
 \end{aligned} \quad (27)$$

where v_0 is the velocity of suction if $v_0 < 0$ and injection if $v_0 > 0$ and $Gr = Gr_1 + Gr_2$

The equations (24) to (26) and boundary conditions (27) can be written as

$$\begin{aligned}
 \frac{\partial^3 f}{\partial \eta^3} + \frac{(\theta - \theta_r)}{\theta_r} \left[\left(f + \frac{1-m}{1+m} \xi \frac{\partial f}{\partial \xi} \right) \frac{\partial^2 f}{\partial \eta^2} - \frac{1-m}{1+m} \xi \frac{\partial^2 f}{\partial \xi \partial \eta} \frac{\partial f}{\partial \eta} - \frac{2}{1+m} M^2 \xi^2 \left(\frac{\partial f}{\partial \eta} - 1 \right) + \frac{2m}{1+m} \left(1 - \left(\frac{\partial f}{\partial \eta} \right)^2 \right) \right. \\
 \left. + \frac{2}{1+m} \frac{N\phi + \theta}{1+N} Gr Re_x \sin \frac{\Omega}{2} - \frac{2}{m+1} \lambda \left(\frac{\partial f}{\partial \eta} - 1 \right) \right] - \frac{2}{1+m} \frac{1}{\theta - \theta_r} \frac{\partial \theta}{\partial \eta} \frac{\partial^2 f}{\partial \eta^2} = 0
 \end{aligned} \quad (28)$$

$$\frac{\partial^2 \theta}{\partial \eta^2} + Pr \left(f + \frac{1-m}{1+m} \xi \frac{\partial f}{\partial \xi} \right) \frac{\partial \theta}{\partial \eta} - \frac{2Pr}{1+m} \theta \frac{\partial f}{\partial \eta} - \frac{1-m}{1+m} \xi \frac{\partial \theta}{\partial \xi} \frac{\partial f}{\partial \eta} = 0 \quad (29)$$

$$\frac{\partial^2 \phi}{\partial \eta^2} + Sc f \frac{\partial \phi}{\partial \eta} + Sc \frac{1+m}{1-m} \left(\frac{\partial \phi}{\partial \eta} \xi \frac{\partial f}{\partial \xi} - \frac{\partial f}{\partial \eta} \xi \frac{\partial \phi}{\partial \xi} \right) - \frac{2Sc}{1+m} \phi \frac{\partial f}{\partial \eta} = 0 \quad (30)$$

$$\begin{aligned}
 \eta = 0: \quad & \frac{\partial f}{\partial \eta} = 0, \quad \frac{(1+m)f}{2} + \frac{1-m}{2} \xi \frac{\partial f}{\partial \xi} = -S, \quad \theta = 1, \quad \phi = 1 \\
 \eta \rightarrow \infty: \quad & \frac{\partial f}{\partial \eta} = 1, \quad \theta = 0, \quad \phi = 0
 \end{aligned} \quad (31)$$

where S is the suction parameter if $S > 0$ and injection if $S < 0$ and $\xi = k x^{\frac{1-m}{2}}$ [5], is the dimensionless distance along the wedge ($\xi > 0$). In this system of equations $f(\xi, \eta)$ is the dimensionless stream function; $\theta(\xi, \eta)$ be the dimensionless temperature; $\phi(\xi, \eta)$ be the dimensionless concentration; Pr , the Prandtl number, Re_x , Reynolds number etc. which are defined in (13) to (23). The parameter ξ indicates the dimensionless distance along the wedge ($\xi > 0$). It is obvious that to retain the ξ -derivative terms, it is necessary to employ a numerical scheme suitable for partial differential equations for the solution. In addition, owing to the coupling between adjacent stream wise location through the ξ -derivatives, a locally autonomous solution, at any given stream wise location can not be obtained. In such a case, an implicit marching numerical solution scheme is usually applied proceeding the solution in the ξ -direction, i.e., calculating unknown profiles at ξ_{i+1} when the same profiles at ξ_i are known. The process

starts at $\xi = 0$ and the solution proceeds from ξ_l to ξ_{l+1} but such a procedure is time consuming.

However, when the terms involving $\frac{\partial f}{\partial \xi}$, $\frac{\partial \theta}{\partial \xi}$ and $\frac{\partial \phi}{\partial \xi}$ and their η derivatives are deleted, the resulting system of equations resembles, in effect, a system of ordinary differential equations for the functions f , θ and ϕ with ξ as a parameter and the computational task is simplified. Furthermore a locally autonomous solution for any given ξ can be obtained because the stream wise coupling is severed. So, following the lines of [5], R.K.Gill,[23] and Shooting numerical solution scheme are utilized for obtaining the solution of the problem. Now, due to the above mentioned factors and for $\xi = 1.0$, the equations (28) to (30) are changed to

$$f''' + \frac{\theta - \theta_r}{\theta_r} f f'' + \frac{2m}{1+m} \frac{\theta - \theta_r}{\theta_r} (1 - f'^2) + \frac{2}{1+m} \frac{\theta - \theta_r}{\theta_r} \frac{N\phi + \theta}{1+N} Gr Re_x \sin \frac{\Omega}{2} - \frac{\theta - \theta_r}{\theta_r} \frac{2}{1+m} \xi^2 (M^2 + Pr \lambda) (f' - 1) - \frac{2}{1+m} \frac{1}{\theta - \theta_r} \theta' f'' = 0 \quad (32)$$

$$\theta'' + Pr f \theta' - \frac{2Pr}{1+m} \theta f' + \frac{2Pr}{m+1} \delta \theta = 0 \quad (33)$$

$$\phi'' + Sc f \phi' - \frac{2Sc}{1+m} f' \phi = 0 \quad (34)$$

with boundary conditions

$$\begin{aligned} \eta = 0: \quad f(0) = -\frac{2}{1+m} S, \quad f'(0) = 0, \quad \theta(0) = 1, \quad \phi(0) = 1 \\ \eta \rightarrow \infty: \quad f'(\infty) = 1, \quad \theta(\infty) = 0, \quad \phi(\infty) = 0 \end{aligned} \quad (35)$$

3. NUMERICAL SOLUTION AND DISCUSSION

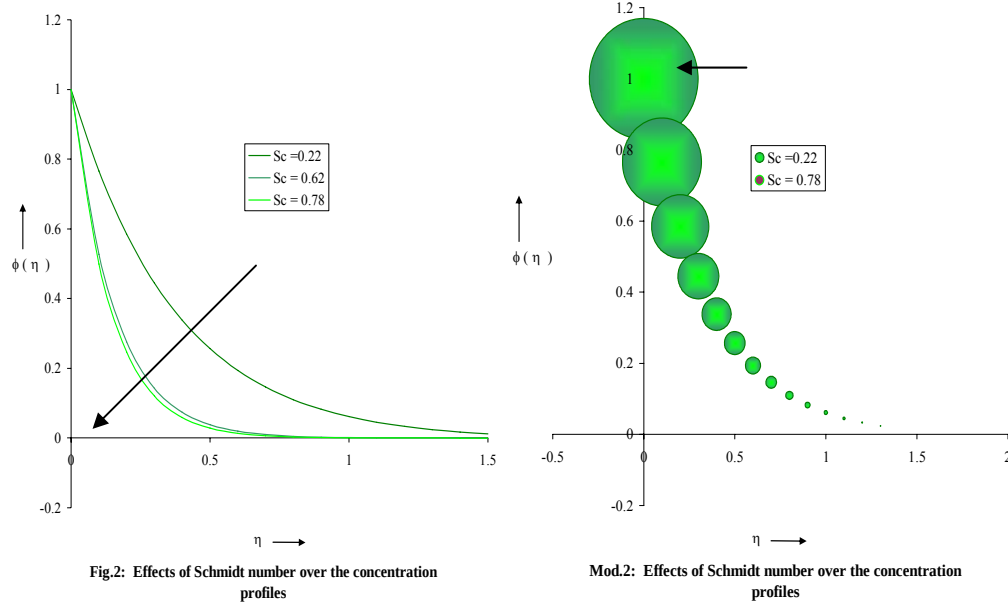
Equations (32) to (34) with boundary condition (35) were solved numerically using Runge Kutta Gill and shooting methods,[23]. The computations have been carried out for various values of variable viscosity, θ_r , magnetic parameter, M^2 and porous medium, λ . In order to validate our method, we have compared steady state results of skin friction, $f''(0)$ and rate of heat transfer $-\theta'(0)$ for various values of θ_r (Table.1) with those of [24] and found them in excellent agreement.

Table.1:

Comparison of the values of $f''(0)$ and $-\theta'(0)$ for various values of θ_r with $\lambda = 0$, $\Omega = 30^\circ$, $N = 0$, $m = 0.0909$, $Sc = 0$, $M^2 = 0$ and $\delta = S = 0$.

θ_r	Pantoskratoras,[24]		Present work		
	$f''(0)$	$-\theta'(0)$	$f''(0)$	$-\theta'(0)$	Pr
2.0	0.2642	-0.7341	0.2651	-0.7345	20.00
4.0	0.3785	-0.7888	0.3791	-0.78921	13.33
6.0	0.4145	-0.8042	0.4167	-0.8061	12.00
8.0	0.4322	-0.8112	0.4341	-0.8127	11.43
10	0.4426	-0.8162	0.4452	-0.8171	11.11

The velocity, temperature and concentration profiles obtained in the dimensionless form are presented in Figures.2-8 for $Pr = 0.71$ which represents air at temperature 20°C and $Sc = 0.62$ which corresponds to water vapor that represents a diffusion chemical species of most common interest in air. Grashof number for heat transfer is chosen to be $Gr = 1.0$, since the these values corresponds to a cooling problem, and Reynolds number $Re_x = 3.0$. The value of θ_r (for air $\theta_r > 0$) is chosen to be 1.0, 3.0 and 5.0 and the value of suction, S is chosen to be 3.0. The value of strength of the magnetic field is chosen to be 0.5, 2.0 and 6.0.



with $Pr = 0.71$, $Sc = 0.62$, $m = 0.0909$, $M^2 = 1.0$, $N = 1.0$, $\lambda = \delta = 0.1$, $S = 3.0$, $\theta_r = 0.5$, $Gr = 1.0$, $Re_x = 3.0$, $\xi = 1.0$ and $\Omega = 30^\circ$

The dimensionless concentration profiles for different values of Schmidt number are plotted in Fig.2 and Mod.2. Due to the uniform viscosity, $\theta_r = 0.1$, it is clear that the concentration of the fluid decreases with increase of Schmidt number while the velocity and temperature are not significant with increase of Schmidt number and these are displayed through Fig.2 and Mod.2.

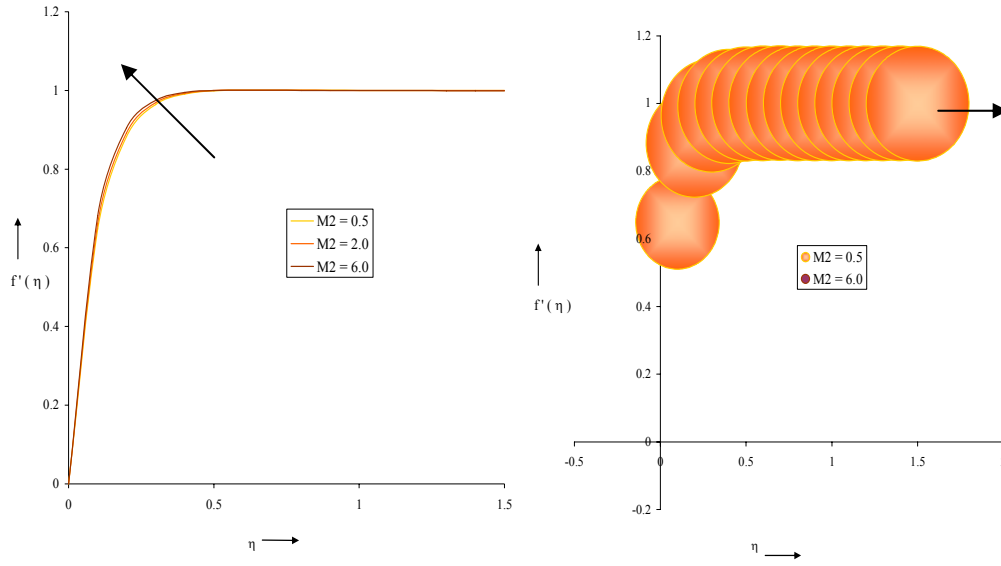


Fig.3: Influence of magnetic field over the velocity profiles

Mod.3: Influence of magnetic field over the velocity profiles

with $Pr = 0.71$, $Sc = 0.62$, $m = 0.0909$, $\gamma = 1.0$, $N = 1.0$, $\lambda = \delta = 0.1$, $S = 3.0$, $\theta_r = 0.5$, $Gr = 1.0$, $Re_x = 3.0$, $\xi = 1.0$ and $\Omega = 30^\circ$

Figure 3 and Model 3 display the influence of the magnetic effects on velocity profiles. It is clear that the velocity increases with increase of the strength of the magnetic effect, while the temperature and concentration are not significant with increase of magnetic effects. As the strength of the magnetic effect increases, the Lorentz force, which oppose the flow, also increases and leads to enhanced deceleration of the flow. This result qualitatively agrees with the expectations, since magnetic field exerts retarding force on the mixed convection flow.

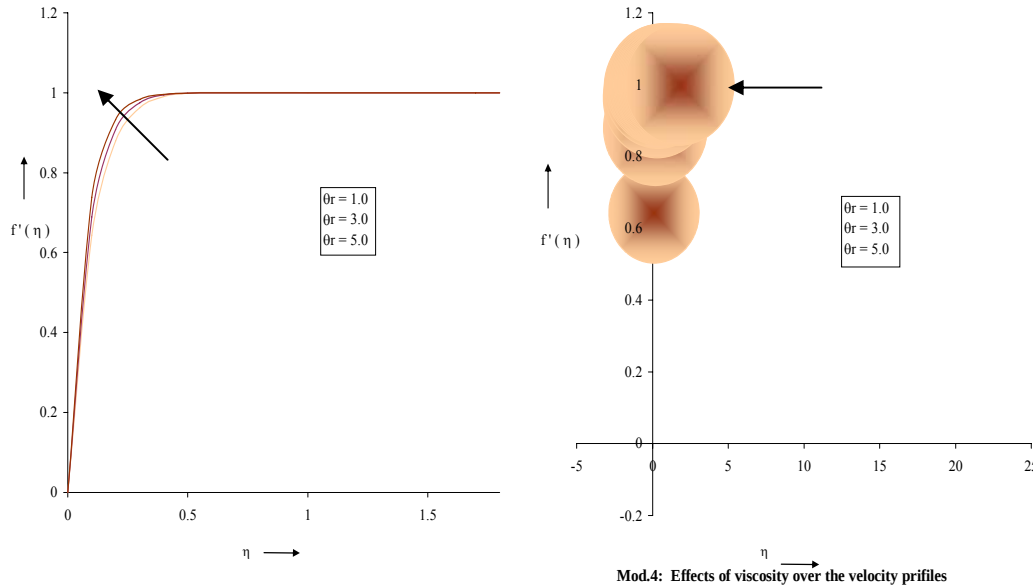


Fig.4: Effects of viscosity over the velocity profiles

Mod.4: Effects of viscosity over the velocity profiles

with $Pr = 0.71$, $Sc = 0.62$, $m = 0.0909$, $\gamma = 1.0$, $N = 1.0$, $\lambda = \delta = 0.1$, $S = 3.0$, $M^2 = 0.5$, $Gr = 1.0$, $Re_x = 3.0$, $\xi = 1.0$ and $\Omega = 30^\circ$

The effects of the viscosity parameter θ_r on velocity is shown through Fig.4 and Mod.4. It is seen that the velocity increases with increase of viscosity parameter while the temperature and concentration are not significant with increase of viscosity. Also, it is observed that the

temperature and concentration of the fluid is almost not affected with increase of the viscosity.

4. CONCLUSIONS

This article studied the effects of nonlinear MHD mixed convection flow and heat and mass transfer of an electrically conducting variable viscosity over a porous wedge in the presence of suction or injection. The results are presented graphically and the conclusion is drawn that the flow field and other quantities of physical interest are significantly influenced by these parameters. Comparisons with previously published works are performed and excellent agreement between the results are obtained.

We conclude the following from the results and discussions:

- * Concentration of the fluid decreases with increase of Schmidt number.
- * Velocity of the fluid increases with increase of the strength of the magnetic effect. As the strength of the magnetic effect increases, the Lorentz force, which opposes the flow, also increases and leads to enhanced deceleration of the flow. This result qualitatively agrees with the expectations, since magnetic field exerts retarding force on the mixed convection flow.
- * Increase of the viscosity accelerates the fluid motion along the wall of the wedge.

It is hoped that the present investigation may be useful for the study of movement of oil or gas and water through the reservoir of an oil or gas field, in the migration of underground water and in the filtration and water purification processes. The results of the problem are also of great interest in geophysics in the study of interaction of the geomagnetic field with the fluid in the geothermal region.

Acknowledgement: The authors wish to express their cordial thanks to our beloved The Rector and The Director of Science Studies Centre, UTHM, Malaysia for their encouragements and acknowledge the financial support received from MOSTI under SFRG 04-01-02-SF0115.

REFERENCES

1. Gehart, B., and Pera, L., 1971, The nature of vertical natural convection flows resulting from the combined buoyancy effects of thermal and mass diffusion, *International Journal of Heat Mass Transfer*, **14**, 2025.
2. Byron Bird, R.; Warren E. Stewart.; Edwin N. Lightfoot., 1992, *Transport phenomena*, John Wiley and sons, New York.
3. Cussler, E.L., 1988, *Diffusion Mass Transfer in Fluid Systems*, Cambridge University Press, London, UK.
4. Yih, K.A., 1998, Uniform suction / blowing effect on forced convection about wedge, *Acta Mech.*, **128**, 173.
5. Kafousias, N.G.; Nanousis, N.D., 1997, Magnetohydrodynamic laminar boundary layer flow over a wedge with suction or injection. *Can. Journal of Physics*, **75**, 733.
6. Kumari, M., 1998, Effect of large blowing rates on the steady laminar incompressible electrically conducting fluid over an infinite wedge with a magnetic field applied parallel to the wedge, *International Journal of Engng.Sci.*, **36**, 299.
7. Anjali Devi, S.P. and Kandasamy, R., 2001, Effects of heat and mass transfer on MHD laminar boundary layer flow over a wedge with suction or injection, *Journal of Energy, Heat and Mass Transfer*, **23**, 167.

8. Yih, K.A.,1998,The effect of uniform suction/blowing on heat transfer of MHD Hiemenz flow through porous media, *Acta Mech.*,**130**,147.
9. Watanabe,T.,1990,Thermal boundary layer over a wedge with uniform suction or injection in forced flow,*Acta Mechanica*, **83**,119.
10. Chamkha,A.J.;Khaled,A.R.A.,2001,Similarity solutions for hydro magnetic simultaneous heat and mass transfer,*Heat Mass Transfer*,**37**,117.
11. Hossain,M.A.,1992,Viscous and Joule heating effects on MHD free convection flow with variable plate temperature, *Int. journal of heat and mass transfer*, **35**, 3485.
12. Hakiem,M.A.EL.;Mohammadeian,A.A.;Kaheir,S.M.M.EL.Gorla,R.S.R.,1999,Joule heating effects on MHD free convection flow of a micro polar fluid, *International Comms. Heat Mass Transfer*,**26**,219.
13. Kuo Bor-Lih.,2005,Heat transfer analysis for the Falkner-Skan wedge flow by the differential transformation method, *Int.al J.of Heat Mass Transfer*,**48**,5036.
14. Cheng, W.T., Lin, H.T.,2002, Non-similarity solution and correlation of transient heat transfer in laminar boundary layer flow over a wedge, *Int. J.of Engg.Sci.*,**40**,531.
15. Apelblat, A., 1982, Mass transfer with a chemical reaction of the first order. Effects of axial diffusion, *The chemical Engineering Journal*, **23**,193.
16. Das,U.N., Deka, R and Soundalgekar, 1994, Effects of mass transfer on flow past an impulsively started infinite vertical plate with constant heat flux and chemical reaction, *Forschung im Ingenieurwesen*,**60**, 284.
17. Muthucumaraswamy, R. and Ganesan,P.,2001, Effects of the chemical reaction and injection on flow characteristics in an unsteady upward motion of an isothermal plate, *Journal of Applied Mechanics and Technical Physics*, **42**, 665
18. Kandasamy,R., Periasamy, K., and Sivagnana Prabhu,K.K.,2005, Effects of chemical reaction, heat and mass transfer along a wedge with heat source and concentration in the presence of suction or injection, *International Journal of Heat and Mass Transfer*,**48**, 1388.
19. Herwing,H. and Wickern, 1986, The effects of variable properties on laminar boundary layer flow, *Warme-und Stoffibctragung*,**20**,47.
20. Lai, E.c. and Kulacki, F.A.,1990, The effects of variable viscosity on convective heat transfer along a vertical surface in a saturated porous medium, *International Journal of Heat and Mass Transfer*, **33**, 1028.
21. Kafoussias, N.g. and Williams, E.w., 1995, The effects of temperature-dependent viscosity on free-forced convective laminar boundary layer flow past a vertical isothermal flat plate, *Acta Mechanica*, **110**, 123.
22. Hady,E.M., Bakier, A.Y., and Gorla, R.S.R.,1996, Mixed convection boundary layer flow on a continuous flat plate with variable viscosity, *Heat and Mass Transfer*,**31**, 169.
23. S.Gill.,1951, *Proceeding of Cambridge Philosophical Society*, pp.96-123.
24. Pantokratoras,A.,2006, The Falkner- Skan flow with constant wall temperature and variable viscosity, *International Journal of Thermal Sciences*, **45**,378.

R. Kandasamy: Science Studies Centre, University Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat Johor, Malaysia.
Email: kandan_kkk@yahoo.co.in

I. Hashim: School of Mathematical Sciences, Universiti Kebangsaan Malaysia,
43600 UKM Bangi Selangor, Malaysia.

Muhaimin: Science Studies Centre, University Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat Johor, Malaysia.

Ruhaila : Science Studies Centre, University Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat Johor, Malaysia.

DERIVATION OF THE NLS BREATHER SOLUTIONS USING DISPLACED PHASE-AMPLITUDE VARIABLES

NATANAEL KARJANTO AND E. VAN GROESEN

Abstract. Breather solutions of the nonlinear Schrödinger equation are derived in this paper: the Soliton on Finite Background, the Ma breather and the rational breather. A special Ansatz of a displaced phase-amplitude equation with respect to a background is used as has been proposed by van Groesen et. al. (2006). Requiring the displaced phase to be temporally independent, has as consequence that the dynamics at each position is described by the motion of a nonlinear autonomous oscillator in a potential energy that depends on the phase and on the spatial phase change. The relation among the breather solutions is confirmed by explicit expressions, and illustrated with the amplitude amplification factor. Additionally, the corresponding physical wave field is also studied and wavefront dislocation together with phase singularity at vanishing amplitude are observed in all three cases.

Key words and Phrases: displaced phase-amplitude variables, nonlinear Schrödinger equation, breather solutions, amplitude amplification factor, wavefront dislocation.

1. INTRODUCTION

The nonlinear Schrödinger (NLS) equation is a nonlinear dispersive partial differential equation that has been used as a mathematical model in many areas of Mathematical Physics, such as nonlinear water waves, nonlinear optics and plasma physics (Ablowitz and Segur, 1981; Ablowitz and Clarkson, 1991). In this letter, we will consider the spatial NLS equation, given as follows:

$$\partial_{\xi}\psi + i\beta\partial_{\tau}^2\psi + i\gamma|\psi|^2\psi = 0. \quad (1)$$

We will derive and study three exact solutions of this equation. These exact solutions are known in the literature as ‘breather’ solutions of the NLS equation (Dysthe and Trulsen, 1999; Dysthe, 2001; Grimshaw et. al., 2001). The name ‘breather’

2000 *Mathematics Subject Classification:* 37K40, 35Q55, 74J30.

reflects the behavior of the profile which is periodic in time or space and localized in space or time. The concept was introduced by Ablowitz, Kaup, Newell and Segur (Ablowitz, et. al., 1974) in the context of the sine-Gordon partial differential equation. The breather solutions are also found in other equations, for instance in the Davey-Stewartson equation (Tajiri and Arai, 2000) and modified Korteweg-de Vries (mKdV) equation (Drazin and Johnson, 1989).

To find the breather solutions of the NLS equation, we use a special Ansatz in the description with displaced phase-amplitude variables introduced by van Groesen et. al. (2006), with a displacement that depends on the background. Requiring the displaced phase to be temporally independent, has as consequence that the dynamics at each position is described by the motion of a nonlinear autonomous oscillator in a potential energy that depends on the phase and on the spatial phase change. It is remarkable that this assumption leads to the breather solutions of the NLS equation that are known in the literatures: the Soliton on Finite Background (SFB) (Akhmediev et. al., 1987), the Ma breather (Ma, 1979) and the rational breather (Peregrine, 1983). These solutions have been suggested as models for a class of freak wave events seen in $2 + 1$ dimensional simulations of surface gravity waves (Henderson et. al., 1999). However, the SFB shows many properties of extreme, or freak, rogue wave events, as observed by Osborne et. al. (2000), Calini and Schober (2002). The SFB is also a good candidate in order to generate extreme wave events in a hydrodynamic laboratory (Andonowati et. al., 2006; Huijsmans et. al., 2005).

This letter is organized as follows. Section 2 presents some exact solutions of the NLS equation. We demonstrate that three different solutions of the NLS equation are obtained within a similar formulation. Section 3 explains briefly the relationship among the solutions of the NLS equation. Section 4 will discuss the evolution of breather solutions in the Argand diagram. In Section 5 we will deal with surface water waves and observe the occurrence of wavefront dislocations in the corresponding density plots of the breather solutions. The final section draws some conclusions and remarks on the subject in this letter.

2. EXACT SOLUTIONS OF THE NLS EQUATION

The simplest nontrivial solution of the NLS equation is the *plane-wave* or the ‘continuous wave’ (cw) solution. It does not depend on the temporal variable τ and is given by $A_0(\xi) = r_0 e^{-i\gamma r_0^2 \xi}$. Another simple solution is the ‘one-soliton’ or ‘single-soliton’ solution. It can be found by the inverse scattering technique (Zakharov and Shabat, 1972), but also much simpler by seeking a travelling-wave solution. An explicit expression is given by

$$A(\xi, \tau) = A_0(\xi) \sqrt{2} \operatorname{sech} \left(r_0 \sqrt{\frac{\gamma}{\beta}} \tau \right). \quad (2)$$

It will become clear when we consider the corresponding physical wave field that

both the cw and the one-soliton solutions have coherent structures, while other solutions that we will deal with below do not.

In van Groesen et. al. (2006), a displaced phase-amplitude description has been introduced. As a special case, we will look for a solution of the NLS equation in the form

$$A(\xi, \tau) = A_0(\xi)[G(\xi, \tau)e^{i\phi(\xi)} - 1], \quad (3)$$

for which the phase ϕ , which is ‘displaced’ since the background contribution is taken out, is taken to depend only on the spatial variable ξ . Substituting (3) into the NLS equation (1), we obtain two equations that correspond to the real and the imaginary parts, respectively.

Multiplying the real part by $\cos \phi$ and the imaginary part by $\sin \phi$, adding both equations will give a Riccati-like equation for G :

$$\partial_\xi G + \gamma r_0^2 \sin 2\phi G - \gamma r_0^2 \sin \phi G^2 = 0. \quad (4)$$

Solving this equation leads us to a conclusion that G can be written in a special form. By letting $G = 1/H$, then equation (4) becomes a first order linear differential equation in H :

$$\partial_\xi H - \gamma r_0^2 \sin 2\phi H + \gamma r_0^2 \sin \phi = 0. \quad (5)$$

Let us take $\tilde{P}(\xi) = \exp(-\gamma r_0^2 \int \sin 2\phi(\xi) d\xi)$ as the integrating factor, multiply it to (5) and integrate the result with respect to ξ , then we obtain the solution for H :

$$H(\xi, \tau) = \frac{-\gamma r_0^2 \int \tilde{P}(\xi) \sin \phi(\xi) d\xi - \zeta(\tau)}{\tilde{P}(\xi)} = \frac{\tilde{Q}(\xi) - \zeta(\tau)}{\tilde{P}(\xi)}, \quad (6)$$

where $-\zeta(\tau)$ is a constant of integration that depends on τ and $\tilde{Q}(\xi)$ denotes the term with the integral sign. Assuming that the displaced phase $\phi(\xi)$ is an invertible function, we can write $\xi = \xi(\phi)$ and dropping the tilde signs to indicate that $P = P(\phi)$ and $Q = Q(\phi)$. Consequently, G is now written as follows:

$$G(\phi, \tau) = \frac{P(\phi)}{Q(\phi) - \zeta(\tau)}. \quad (7)$$

We will observe in the following subsections that by choosing three different functions of $\zeta(\tau)$ will lead to three different breather solutions of the NLS equation.

On the other hand, multiplying the real part by $\sin \phi$ and the imaginary part by $\cos \phi$, subtracting one from the other we obtain a nonlinear oscillator equation for G :

$$\beta \partial_\tau^2 G + (\phi'(\xi) + 2\gamma r_0^2 \cos^2 \phi)G - 3\gamma r_0^2 \cos \phi G^2 + \gamma r_0^2 G^3 = 0. \quad (8)$$

In deriving the breather solutions, we will compare this equation with the second order differential equation for G after choosing a particular function of $\zeta(\tau)$.

2.1 Soliton on Finite Background

Firstly, we take $\zeta(\tau) = \cos(\nu\tau)$, where ν is a modulation frequency. For the normalized quantity $\tilde{\nu} = \nu / \left(r_0 \sqrt{\frac{\gamma}{\beta}} \right)$ we consider $0 < \tilde{\nu} < \sqrt{2}$. The differential equation for G becomes:

$$\partial_\tau^2 G = -\nu^2 G + 3\nu^2 \frac{Q}{P} G^2 + 2\nu^2 \frac{1-Q^2}{P^2} G^3.$$

By comparing with (8), the solution (7) is obtained with $P(\phi) = \tilde{\nu}^2 Q(\phi) / \cos \phi$, $Q^2(\phi) = 2 \cos^2 \phi / (2 \cos^2 \phi - \tilde{\nu}^2)$ and the displaced phase satisfies $\tan \phi(\xi) = -\frac{\tilde{\sigma}}{\tilde{\nu}^2} \tanh(\sigma\xi)$, where $\tilde{\sigma} = \tilde{\nu} \sqrt{2 - \tilde{\nu}^2}$ and $\sigma = \gamma r_0^2 \tilde{\sigma}$. The corresponding solution of the NLS equation (3) with extremum at $(\xi, \tau) = (0, 0)$ can then be written after some manipulations as:

$$A(\xi, \tau) = A_0(\xi) \cdot \left(\frac{\tilde{\nu}^2 \cosh(\sigma\xi) - i\tilde{\sigma} \sinh(\sigma\xi)}{\cosh(\sigma\xi) - \sqrt{1 - \frac{1}{2}\tilde{\nu}^2} \cos(\nu\tau)} - 1 \right).$$

This solution was found by Akhmediev et. al. (1987). They suggested a method of obtaining exact solutions of the NLS equation which is based on the substitution connecting the real and imaginary parts of the solution by a linear relationship with coefficients depending only on time. The method consists in constructing a certain system of ordinary differential equations, the solutions of which determine the solutions of the NLS equation. However, the explicit expression of the SFB already appeared in the previous papers (Akhmediev et. al. 1985; Akhmediev and Korneev, 1986). The SFB is also derived by Ablowitz and Herbst (1990) using Hirota's method (Hirota, 1976) and for $\tilde{\nu} = 1$ by Osborne et. al. (2000) using the inverse scattering technique. The SFB is the single homoclinic orbit to the cw as a fixed point of the NLS equation (Ablowitz and Herbst, 1990; Li and McLaughlin, 1994, 1997; Calini and Schober, 2002). The asymptotic behaviour for $\xi \rightarrow -\infty$ of the SFB describes the linear instability of the cw solution for side-band perturbation (Akhmediev and Korneev, 1986). This modulational instability is known as the Benjamin-Feir instability in the context of water waves (Benjamin and Feir, 1967). A plot of the absolute value of this complex amplitude solution for $r_0 = 1$ and $\tilde{\nu} = 1/2$ is given in Figure 1(a).

2.2 Ma breather

Secondly, we take $\zeta(\tau) = \cosh(\mu\tau)$, where $\mu = r_0 \sqrt{\frac{\gamma}{\beta}} \tilde{\mu}$. The differential equation for G becomes:

$$\partial_\tau^2 G = \mu^2 G - 3\mu^2 \frac{Q}{P} G^2 - 2\mu^2 \frac{1-Q^2}{P^2} G^3.$$

Comparing again with (8), the solution (7) is obtained with $P(\phi) = -\tilde{\mu}^2 Q(\phi) / \cos \phi$, $Q^2(\phi) = 2 \cos^2 \phi / (2 \cos^2 \phi + \tilde{\mu}^2)$ and the displaced phase satisfies $\tan \phi(\xi) = -\frac{\tilde{\rho}}{\tilde{\mu}^2} \tan(\rho\xi)$, where $\tilde{\rho} = \tilde{\mu} \sqrt{2 + \tilde{\mu}^2}$ and $\rho = \gamma r_0^2 \tilde{\rho}$. The corresponding solution of the NLS (3)

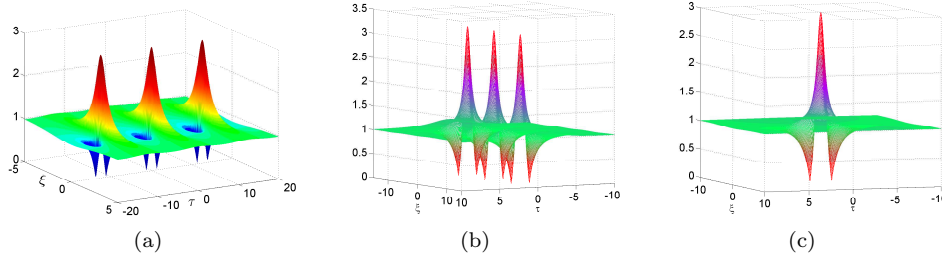


Figure 1: Plots of the absolute value of the SFB breather for $\bar{\nu} = 1/2$ (left), the Ma breather for $\tilde{\mu} = 0.4713$ (center) and the rational breather (right), for $r_0 = 1$ in all cases. For illustration purposes, the axis are scaled corresponding to $\beta = 1 = \gamma$.

can then be written after some manipulations as:

$$A(\xi, \tau) = A_0(\xi) \cdot \left(\frac{-\tilde{\mu}^2 \cos(\rho\xi) + i\tilde{\rho} \sin(\rho\xi)}{\cos(\rho\xi) - \sqrt{1 + \frac{1}{2}\tilde{\mu}^2} \cosh(\mu\tau)} - 1 \right).$$

We call this solution Ma solution or Ma breather since it has been found by Ma (1979) using the inverse scattering technique of (Gardner et. al., 1967). The same technique is also used by Osborne et. al. (2000) to arrive at this solution for $\tilde{\mu} = \sqrt{2}$. A plot of the absolute value of the Ma breather for $r_0 = 1$ and $\tilde{\mu} = 0.4713$ is given in Figure 1(b). This choice of this value of $\tilde{\mu}$ is just for our convenience, which we will use again for the corresponding physical wave field.

2.3 Rational breather

Thirdly, we take $\zeta(\tau) = 1 - \frac{1}{2}\nu^2\tau^2$. Similarly, the same solution can also be obtained by substituting $\zeta(\tau) = 1 + \frac{1}{2}\mu^2\tau^2$. As a consequence, the differential equation for G now becomes:

$$\partial_\tau^2 G = \frac{3\nu^2}{P} G^2 - 4\nu^2 \frac{Q-1}{P^2} G^3.$$

Comparing again with (8), we have $P(\phi) = \tilde{\nu}^2 / \cos \phi$, $Q(\phi) = 1 + P^2 / (4\tilde{\nu}^2)$ and the displaced phase satisfies $\tan \phi(\xi) = -2\gamma r_0^2 \xi$. Substituting into the Ansatz (7), we obtain

$$A(\xi, \tau) = A_0(\xi) \cdot \left(\frac{4(1 - 2i\gamma r_0^2 \xi)}{1 + 4(\gamma r_0^2 \xi)^2 + 2\frac{\gamma}{\beta} r_0^2 \tau^2} - 1 \right).$$

This rational solution is clearly different from the two solutions above, is derived by Peregrine (1983) as a limiting case from the Ma breather. He took the analytic expression and performed a double Taylor series expansion about the amplitude peak (which occur at $\xi = 0 = \tau$) to arrive at his solution. Since this solution is

confined in both space and time, some authors (Henderson et. al., 1999) also called it the ‘isolated Ma soliton’ and other authors (Akhmediev and Ankiewicz, 1997) called it the ‘rational solution’. Because of its soliton-like feature in ξ , (Nakamura and Hirota, 1985) called it as an ‘explode-decay solitary wave’. In this letter, we call it the ‘rational breather’. A plot of the absolute value of the rational breather for $r_0 = 1$ is given in Figure 1(c).

3. RELATION BETWEEN THE NLS BREATHING SOLUTIONS

As shown by Dysthe and Trulsen (1999), there is a relationship among the breather solutions of the NLS equation. Different parameters are used for explicit expressions of the SFB and the Ma breather in Dysthe and Trulsen (1999) as well as in Grimshaw et. al. (2001). Introducing new parameters φ and $\vartheta \in \mathbb{R}$ by the following relations: $\tilde{\nu} = \sqrt{2} \sin \varphi$, $\tilde{\sigma} = \sin 2\varphi$, $\tilde{\mu} = \sqrt{2} \sinh \vartheta$ and $\tilde{\rho} = \sinh 2\vartheta$, we can write our expressions similar to the ones in these papers.

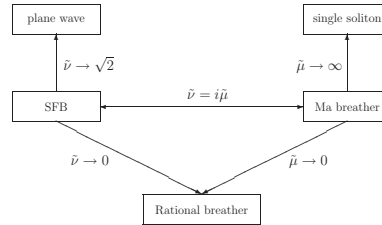


Figure 2: The schematic diagram for the derivation of the Ma breather and the rational breather from the SFB.

The SFB becomes the Ma breather if we substitute $\nu = i\mu$ and it becomes the rational breather if $\nu \rightarrow 0$. Similarly, the Ma breather becomes the rational breather for $\mu \rightarrow 0$. Interestingly, the Ma breather becomes the one-soliton solution for $\mu \rightarrow \infty$ (Akhmediev and Ankiewicz, 1997). Figure 2 explains the schematic diagram of these derivations. These relations can also be seen from the expression in the previous section. For the SFB, we take $\zeta(\tau) = \cos(\nu\tau)$, which leads to the Ma breather if we substitute $\nu = i\mu$, so that $\zeta(\tau) = \cos(i\mu\tau) = \cosh(\mu\tau)$. Taking either ν or $\mu \rightarrow 0$, we get $\zeta(\tau) = 1 - \frac{1}{2}\nu^2\tau^2$ or $\zeta(\tau) = 1 + \frac{1}{2}\mu^2\tau^2$, and both will lead to the rational breather.

The relations have also consequence for the amplitude amplification factor (AAF), defined as the ratio between the maximum amplitude and the value of its background. The expressions for the breather solutions as given in this letter were shifted such that the maximum amplitude is at $(\xi, \tau) = (0, 0)$ and the value of the background is r_0 . For the SFB, the amplification is given by $\text{AAF}_s(\tilde{\nu}) = 1 + \sqrt{4 - 2\tilde{\nu}^2}$. For $0 < \tilde{\nu} < \sqrt{2}$, the amplification is bounded and are $1 < \text{AAF}_s(\tilde{\nu}) < 3$. The AAF for the Ma breather is given by $\text{AAF}_{\text{Ma}}(\tilde{\mu}) = 1 + \sqrt{4 + 2\tilde{\mu}^2}$. Hence, for

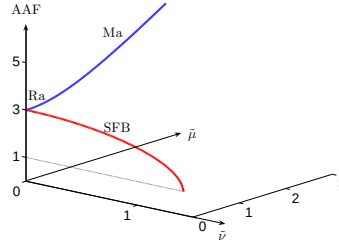


Figure 3: Plot of the amplitude amplification factor for the SFB and the Ma breather. The meeting point of the two curves is the AAF of the rational breather.

$\tilde{\mu} > 0$, we have $\text{AAF}_{\text{Ma}} > 3$. The AAF for the rational breather is exactly 3, which follows by letting the either $\tilde{\nu}$ or $\tilde{\mu}$ go to zero

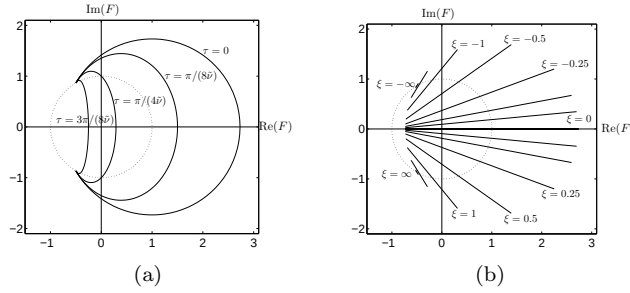
$$\text{AAF}_{\text{Ra}} = \lim_{\tilde{\nu} \rightarrow 0} \text{AAF}_{\text{S}} = 3 = \lim_{\tilde{\mu} \rightarrow 0} \text{AAF}_{\text{Ma}}. \quad (9)$$

In Onorato et. al. (2001), the AAF for the SFB is given as a function of the wave steepness and the number of waves under the modulation. The plot of the AAF for all the three breather solutions is given in Figure 3. For $\tilde{\mu} \rightarrow \infty$ in the Ma breather, the single-soliton solution is obtained (Akhmediev and Ankiewicz, 1997).

4. EVOLUTION IN THE ARGAND DIAGRAM

In this section we present the evolution of the breather solutions in a complex plane. The real and the imaginary axes of this plane are denoted by the real and the imaginary part of the breather solutions after removing the plane-wave contribution, respectively. In this paper we refer this plane as the Argand diagram. Since the breather solutions depend on the space variable ξ and the time variable τ , the evolution curves are presented by parameterizing one variable and plotting for the other at different values. We will observe that the point $(-1, 0)$ plays a significant role with respect to all evolution curves, whether they are parameterized in space or in time. A study of this evolution curves for understanding experimental results on freak wave generation using the SFB solution has been presented recently (Karjanto, 2006).

Figure 4 shows the evolution curves of the SFB for $\tilde{\nu} = 1/\sqrt{2}$. The parameterized plot in position for different times is given by Figure 4(a). For this particular value of modulation frequency $\tilde{\nu}$, all curves start from a point in the second quadrant, which corresponds to $\xi = -\infty$, and end to a point in the third quadrant, which corresponds to $\xi = \infty$. During one modulation period $0 \leq \tau < 2\pi/\tilde{\nu}$, we observe that the curve reaches its maximum value for $\tau = 0$ when the $\text{Im}(F)$ vanishes and it reaches a minimum value for $\tau = \pi/\tilde{\nu}$. The dotted circle indicates the plane wave solution that has been removed. When we allow the modulation frequency

Figure 4: The evolution of the SFB in the Argand diagram for $\tilde{\nu} = 1/\sqrt{2}$.

$\tilde{\nu} \rightarrow 0$, both the initial and the final points will coincide at $(-1, 0)$, as shown by the rational breather case in Figure 6(a). The parameterized plot in time for different positions is given by Figure 4(b). All curves are straight lines and they are passed twice during one modulation period. For $\xi \rightarrow \pm\infty$, the lines shrink to a point in the third and the second quadrants, respectively. The longest line is reached at $\xi = 0$ when it lies on the real axis. Notice that $(-1, 0)$ becomes a center for all the lines. Moreover, for $\tilde{\nu} \rightarrow 0$, the left-hand side of all the lines will meet at $(-1, 0)$, as shown by the rational breather case in Figure 6(b).

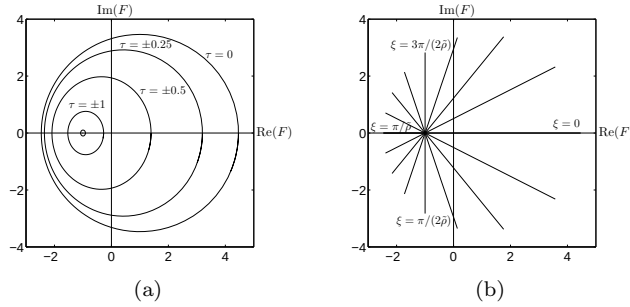
Figure 5: The evolution of the Ma breather in the Argand diagram for $\tilde{\mu} = 2$.

Figure 5 shows the evolution curves of the Ma breather for $\tilde{\mu} = 2$. The parameterized plot in position for different times is given by Figure 5(a). All the curves are elliptical, which indicate that they are ξ -periodic. The largest curve occurs at $\tau = 0$, which has a circular form. As the time moves forward and backward, the ellipses are getting smaller until eventually shrink to a point at $(-1, 0)$ for $\tau = \pm\infty$. By letting $\tilde{\mu} \rightarrow 0$, the left-hand side of the ellipses at the real axis moves accordingly toward $(-1, 0)$, and eventually the pattern becomes the one as shown by the rational breather in Figure 6(a). The parameterized plot in time for different positions is given by Figure 5(b). Similar to the previous case, all the

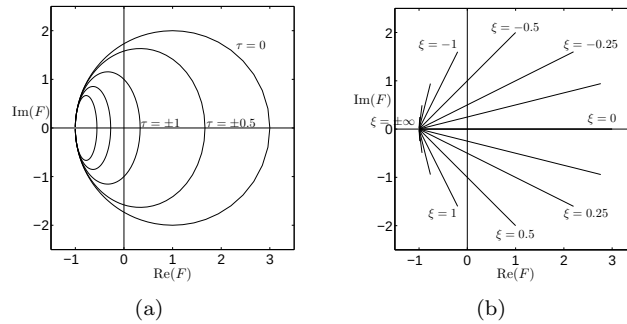


Figure 6: The evolution of the rational breather in the Argand diagram.

curves are straight lines and we notice clearly that they are centered at $(-1, 0)$. Different from the SFB solution case where the lines are at the right-hand side of $(-1, 0)$, for the Ma breather case the point radiates lines to all different direction depending on the position. For one periodic position $0 \leq \xi < 2\pi/\tilde{\rho}$, the longest line occurs at $\xi = 0$ and the shortest one occurs at $\xi = \pi/\tilde{\rho}$, being both of them lay on the real axis. When $\tilde{\mu} \rightarrow 0$, the lines which are on the left-hand side of $(-1, 0)$ will get shorter until eventually become the rational breather case as shown in Figure 6(b). Figure 6 shows the evolution curves of the rational breather, which show the limiting cases for both $\tilde{\nu} \rightarrow 0$ and $\tilde{\mu} \rightarrow 0$.

5. PHYSICAL WAVE FIELDS

In this section, we show applications in surface water waves. Since the NLS equation is an envelope equation, the corresponding physical wave fields $\eta(x, t)$ are wave packets. Considering only first order contributions, the physical wave field is given by

$$\eta(x, t) = A(\xi, \tau) e^{i(k_0 x - \omega_0 t)} + \text{c.c.} \quad (10)$$

where c.c. denotes complex conjugate of the preceding term. The complex amplitude A is described in a moving frame of reference with $\xi = x$ and $\tau = t - x/V_0$. The wavenumber k_0 and the frequency ω_0 are related by the linear dispersion relation. For surface water waves, the normalized form of the dispersion relation is given by $\omega = \Omega(k) = k \sqrt{(\tanh k)/k}$, the group velocity $V_0 = \Omega'(k_0)$.

Density plots for three breather solutions are given in Figure 7. For illustrations, we choose $r_0 = 1$, $k_0 = 2\pi$ ($\omega_0 \approx \sqrt{2\pi}$) in all cases, modulation frequency $\tilde{\nu} = 1/2$ for the SFB and $\tilde{\mu} = 0.4713$ for the Ma breather. For better representation, the physical wave fields are shown in a moving frame of reference with suitable chosen group velocity. The SFB has extreme values at $x = \xi = 0$ and is periodic in time. The Ma breather has extreme values at $\tau = 0$ and is periodic in space. The rational breather is neither periodic in time nor in space, but isolated with

its maximum at $(x, t) = (0, 0)$. Despite some differences, all the breather solutions show ‘wavefront dislocation’, when splitting or merging of waves occurs (Nye and Berry, 1974). A necessary condition for this phenomenon to occur is that the Chu-Mei quotient of the nonlinear dispersion relation at the vanishing amplitude is unbounded. When it occurs, the real-valued phase of the corresponding wave field is undefined at the vanishing amplitude, known as ‘phase singularity’. More information on wavefront dislocation in surface water waves can be found in (Tanaka, 1995; Karjanto and van Groesen, 2007).

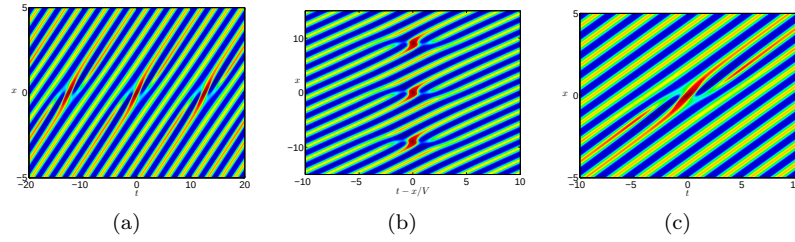


Figure 7: Density plot of the physical wave field of the SFB for $\tilde{\nu} = 1/2$ (left), the Ma breather for $\tilde{\mu} = 0.4713$ (center) and the rational breather (right) corresponding to Figure 1 for $k_0 = 2\pi$. The plots are shown in a moving frame of reference with suitable chosen velocity.

6. CONCLUSIONS

In this letter, we have shown that a description with displaced phase-amplitude variables, for which the phase is time independent, leads to the three breather solutions of the NLS equation: the Soliton on Finite Background, the Ma breather, and the rational breather. These are obtained as explicit solutions of the equation for the displaced amplitude of the NLS equation. We described the relation between these three breather solutions and presented the amplitude amplification factor of each solution as function of the parameters. For the corresponding physical wave fields, wavefront dislocation and phase singularity at vanishing amplitude are observed in all three cases.

Acknowledgement. Partly of this work is executed at University of Twente, The Netherlands as part of the project ‘Prediction and generation of deterministic extreme waves in hydrodynamic laboratories’ (TWI.5374) of the Netherlands Organization of Scientific Research NWO, subdivision Applied Sciences STW.

REFERENCES

1. M. J. ABLOWITZ, ET. AL. The inverse scattering transform-Fourier analysis for nonlinear problems, *Stud. Appl. Math.*, **53**, 249-315, 1974.

2. M. J. ABLOWITZ AND H. SEGUR, *Solitons and Inverse Scattering Transform*, Society for Industrial and Applied Mathematics, Philadelphia, 1981.
3. M. J. ABLOWITZ AND B. M. HERBST, On homoclinic structure and numerically induced chaos for the NLS equation, *SIAM J. Appl. Math.*, **50**, 2:339-351, 1990.
4. M. J. ABLOWITZ AND P. A. CLARKSON, *Solitons, Nonlinear Evolution Equations and Inverse Scattering*, volume **149** of *London Mathematical Society Lecture Note Series*. Cambridge University Press, 1991.
5. N. N. AKHMEDIEV, V. M. ELEONSKIĬ AND N. E. KULAGIN, Generation of periodic trains of picosecond pulses in an optical fiber: exact solutions, *Sov. Phys. JETP*, **62**, 5: 894-899, 1985.
6. N. N. AKHMEDIEV AND V. I. KORNEEV, Modulation instability and periodic solutions of the nonlinear Schrödinger equation, *Teoret. Mat. Fiz.*, **62**, 2:189-194, 1986. English translation: *Theoret. Math. Phys.* **69**, 1089-1092, 1986.
7. N. N. AKHMEDIEV, V. M. ELEONSKIĬ, AND N. E. KULAGIN, First-order exact solutions of the nonlinear Schrödinger equation, *Teoret. Mat. Fiz.*, **72**, 2:183-196, 1987. English translation: *Theoret. Math. Phys.*, **72**, 2:809-818, 1987.
8. N. N. AKHMEDIEV AND A. ANKIEWICZ, *Solitons—Nonlinear Pulses and Beams*, volume **5** of *Optical and Quantum Electronic Series*, Chapman & Hall, first edition, 1997.
9. ANDONOWATI, N. KARJANTO AND E. VAN GROESEN, Extreme wave phenomena in down-stream running modulated waves, *Appl. Math. Model.*, **31**, 1425-1443, 2007.
10. T. B. BENJAMIN AND J. E. FEIR, The disintegration of wave trains on deep water. Part 1. Theory., *J. Fluid Mech.*, **27**, 3:417-430, 1967.
11. A. CALINI AND C. M. SCHOBBER, Homoclinic chaos increases the likelihood of rogue wave formation, *Phys. Lett. A*, **298**, 335-349, 2002.
12. P. G. DRAZIN AND R. S. JOHNSON, *Solitons: an Introduction*, Cambridge University Press, 1989.
13. K. B. DYSTHE AND K. TRULSEN, Note on breather type solutions of the NLS as models for freak-waves, *Phys. Scripta*, **T82**, 48-52, 1999.
14. K. B. DYSTHE, Modelling a “rogue wave”—speculations or a realistic possibility? In M. Olagnon and G. A. Athanassoulis, editors, *Proceedings Rogue Waves 2000*, Ifremer, Brest, France, 2001.
15. C. S. GARDNER, J. M. GREENE, M. D. KRUSKAL AND R. M. MIURA, Method for solving the Korteweg-de Vries equation, *Phys. Rev. Lett.* **19**, 19:1095-1097, 1967.
16. R. GRIMSHAW, D. PELINOVSKY, E. PELINOVSKY AND T. TALIPOVA, Wave group dynamics in weakly nonlinear long-wave models, *Physica D* **159**, 35-37, 2001.
17. K. L. HENDERSON, D. H. PEREGRINE, AND J. W. DOLD, Unsteady water wave modulations: fully nonlinear solutions and comparison with the nonlinear Schrödinger equation. *Wave Motion*, **29**, 341-361, 1999.
18. R. HIROTA, Direct method of finding exact solutions of nonlinear evolution equations, in R. M. Miura, editor, *Bäcklund Transformations, the Inverse Scattering Method, Solitons, and Their Applications*, Springer-Verlag, Berlin, 1976.
19. R. H. M. HUIJSMANS, G. KLOPMAN, N. KARJANTO AND ANDONOWATI, Experiments on extreme wave generation using the Soliton on Finite Background, in M. Olagnon and M. Prevosto, editors, *Proceedings Rogue Waves 2004*, Ifremer, Brest, France, 2005.

20. N. KARJANTO, *Mathematical Aspects of Extreme Water Waves*, PhD thesis, University of Twente, 2006.
21. N. KARJANTO AND E. VAN GROESEN, Note on wavefront dislocation in surface water waves, *Phys. Lett. A*, **37**, 173-179, 2007.
22. Y. LI AND D. W. McLAUGHLIN, Morse and Melnikov functions for NLS PDE's, *Commun. Math. Phys.*, **162**, 175-214, 1994.
23. Y. LI AND D. W. McLAUGHLIN, Homoclinic orbits and chaos in discretized perturbed NLS systems: Part I. Homoclinic orbits, *J. Nonlinear Sci.*, **7**, 211-269, 1997.
24. Y.-C. MA, The perturbed plane-wave solutions of the cubic Schrödinger equation, *Stud. Appl. Math.*, **60**, 1:43-58, 1979.
25. A. NAKAMURA AND R. HIROTA, A new example of explode-decay solitary waves in one dimension, *J. Phys. Soc. Japan*, **54**, 2:491-499, 1985.
26. J.F. NYE AND M.V. BERRY, Dislocation in wave trains, *Proc. R. Soc. Lond. A*, **336**, 1065:165-190, 1974.
27. M. ONORATO, A. OSBORNE, M. SERIO AND T. DAMIANI, Occurrence of freak waves from envelope equations in random ocean wave simulations, in M. Olagnon and G. A. Athanassoulis, editors, *Proceedings Rogue Waves 2000*, Ifremer, Brest, France, 2001.
28. A. R. OSBORNE, M. ONORATO AND M. SERIO, The nonlinear dynamics of rogue waves and holes in deep-water gravity wave trains, *Phys. Lett. A*, **275**, 386-393, 2000.
29. D. H. PEREGRINE, Water waves, nonlinear Schrödinger equations and their solutions, *J. Austral. Math. Soc. Ser. B*, **25**, 1:16-43, 1983.
30. M. TANAKA, Disappearance of waves in modulated train of surface gravity waves, *Structure and Dynamics of Nonlinear Waves in Fluids*, in A. Mielke and K. Kirchgässner, editors, IUTAM/ISIMM Symposium Proceedings (Hannover, August 1994), volume **7** of *Advanced Series in Nonlinear Dynamics*, World Scientific, Singapore, 392-398, 1995.
31. M. TAJIRI AND T. ARAI, Periodic soliton solutions to the Davey-Stewartson equation, *Proc. Inst. Math. Natl. Acad. Sci. Ukr.*, **30**, 1:210-217, 2000.
32. E. van Groesen, Andonowati and N. Karjanto, Displaced phase-Amplitude variables for waves on finite background, *Phys. Lett. A*, **354**, 312-319, 2006.
33. V. E. Zakharov and A. B. Shabat, Exact theory of two-dimensional self-focusing and one-dimensional self-modulation of waves in nonlinear media, *Sov. Phys. JETP*, **34**, 1:62-69, 1972.

NATANAEL KARJANTO: School of Applied Mathematics, The University of Nottingham Malaysia Campus, Jalan Broga, 43500 Semenyih, Selangor Darul Ehsan, Malaysia.
E-mail: natanael.karjanto@nottingham.edu.my

E. VAN GROESEN: Department of Applied Mathematics, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
E-mail: groesen@math.utwente.nl

SYMMETRIC RANK ONE METHOD FOR UNCONSTRAINED OPTIMIZATION PROBLEMS

M.B MAMAT, K.H. SOW , Y.B DASRIL, AND I.B MOHD

Abstract. Symmetric rank one (SR1) algorithm for unconstrained nonlinear optimization generates a sequence of matrices that can be considered as approximation of the objectives function second derivatives. The SR1 update formula is used to prove that the approximation is converging to the true Hessian matrix. In this paper, the convergence rate of a positive definite scaled symmetric rank one is examined. This analysis assumed that the Hessian approximations are positive-definite and bounded asymptotically. Numerical experiments indicate that the modified SR1 (mSR1) method is better than SR1 method

Key words and Phrases: SR1 method, uniform linear independence, superlinear, positive definite, Lipschitz continuous.

1. INTRODUCTION

Quasi-Newton methods are recognized today as one of the most efficient ways to solve nonlinear unconstrained or bound constrained optimization problems. These methods are mostly used when the second derivative matrix of the objective function is either unavailable or too costly to compute. They are very similar to Newton's method, but avoid the need of computing Hessian matrices by recurring, from iteration to iteration, a symmetric matrix which can be considered as an approximation of the true Hessian. Therefore, these methods allow the curvature of the problem to be exploited in the numerical algorithm, despite the fact that only first derivatives (gradients) and function values are required.

The problem considered here is that of finding a local minimum of the unconstrained optimization problem

$$\min_{x \in \mathfrak{R}^n} f(x) \quad (1)$$

where $f(x)$ is a smooth function from $\mathfrak{R}^n$ into $\mathfrak{R}$, with $f(x)$ assumed to be at least twice continuously differentiable. Given current iteration $x_k, f(x_k), \nabla f(x_k)$ or finite difference approximation, and Hessian approximation $B_k \in \mathfrak{R}^{n \times n}$ symmetric, select new iterate x_{k+1} by a line search method. Update B_k to B_{k+1} such that B_{k+1} is symmetric and satisfies the secant equation

$$y_k = B_{k+1} s_k$$

where $s_k = x_{k+1} - x_k$, $y_k = g_{k+1} - g_k$ and $g_k = \nabla f(x_k)$.

In this paper, considered the SR1 update for the Hessian approximation as :

$$B_{k+1} = B_k + \frac{(y_k - B_k s_k)(y_k - B_k s_k)^T}{s_k^T (y_k - B_k s_k)} \quad (2)$$

Sherman-Morrison formula stated that if B is a nonsingular matrix, and if $H_k = B_k^{-1}$ then (2) is given by

$$H_{k+1} = H_k + \frac{(s_k - H_k y_k)(s_k - H_k y_k)^T}{y_k^T (s_k - H_k y_k)} \quad (3)$$

The procedure used to compute the search direction, d_k is often called the “line search”. In this context, maintaining B_k positive definite has the distinct advantage that it guarantees that d_k is a descent direction with respect to the objective function.

2. SR1 METHOD

The SR1 formula makes a symmetric rank one change to the previous Hessian approximation B_k . Compared with other secant updates, the SR1 update is simpler and may require less computation per iteration. A basic disadvantage to the SR1 update is that the SR1 update may not preserve positive-definiteness even if this is positive, for instance, when B_k is positive-definite and $s_k^T y_k > 0$. A simple remedy by Malik et al. (2001) to this problem is to replace the non-positive definite B_{k+1} with a positive multiple of the identity whenever this difficulty arises.

Recent work by Conn et al. (1988a, 1988b, and 1991) has sparked renewed interest in the SR1 update. Conn et al. (1991) proved that the sequences of matrices generated by the SR1 formula converges to the actual Hessian at the solution $\nabla^2 f(x^*)$, provided that the steps taken are uniformly linearly independent, that the SR1 update denominator is always sufficiently different from zero, and that the iterate converge to finite limit. The assumption of uniform linear independence is defined as follows.

Definition 2.1 A sequence of vectors $\{s_k\}$ is uniformly linearly independent, if there exist a $c_1 > 0$, a k_0 and an $m \geq n$ such that, for each $k \geq k_0$, one can choose n distinct indices

$$k \leq k_1 < \dots < k_n \leq k + m$$

with

$$\sigma_{\min}(S_k) \geq c_1$$

where $\sigma_{\min}(S_k)$ is the minimum singular value of the matrix

$$S_k = \left[\frac{s_{k_1}}{\|s_{k_1}\|}, \dots, \frac{s_{k_n}}{\|s_{k_n}\|} \right] \quad (4)$$

The condition of linear independence of the sequence $\{s_k\}$ under which Conn et

al. (1991) analyze the performance of the SR1 method may be too strong in practice. Therefore, in this paper, the condition of linear independence of the sequence $\{s_k\}$ is omitted, but adds instead the condition that the sequence $\{B_k\}$ remains positive-definite and bounded.

The following assumptions will frequently be made on the problems and the sequence of iterates generated by the SR1 algorithmic model.

Assumptions 2.1

- (i) Function $f(x)$ is twice continuously differentiable everywhere.
- (ii) $\nabla^2 f(x^*)$ is positive definite and its Hessian $\nabla^2 f(x) = H(x)$ is Lipschitz continuous near optimal value x^* , that is, there exists a constant $c_2 > 0$ such that for all $x, y \in \mathbb{R}^n$,
$$\|H(x) - H(y)\| \leq c_2 \|x - y\|$$
and c_2 is called Lipschitz constant for function f .
- (iii) The sequence $\{x_k\}$ converges to local finite limit point, x^* .

Due to Conn et al. (1991) analysis, if the steps used by SR1 updates are linear independent, then $\{B_k\}$ converges to $\nabla^2 f(x^*)$.

Lemma 2.1 Let $\{x_k\}$ be a sequence of iterates defined by $x_{k+1} = x_k + s_k$. Suppose that Assumptions 2.1 hold, that the sequence of matrices $\{B_k\}$ is generated from the SR1 updates and that for each iteration

$$\left| s_k^T (y_k - B_k s_k) \right| \geq r \|s_k\| \|y_k - B_k s_k\| \quad (5)$$

where $r \in (0,1)$ is a constant. Then, for each j , $\|y_j - B_{j+1} s_j\| = 0$, and

$$\|y_j - B_j s_j\| \leq \frac{\gamma}{r} \left(\frac{2}{r} + 1 \right)^{i-j-2} \eta_{i,j} \|s_j\| \quad (6)$$

for all $i \geq j+2$, where $\eta_{i,j} = \max\{\|x_q - x_t\| \mid j \leq t \leq q \leq i\}$, and γ is the Lipschitz constant from Assumptions 2.1.

Actually, it is apparent from Lemma 2.1 by Conn et al. (1991) that, if the update is skipped whenever (5) is violated, then (6) still holds for all j for which (5) is true. The following lemma by Byrd et al. (1993) proved that if the sequence of steps generated by an iterative process using the SR1 update satisfies (5), and the sequence of matrices is bounded, then out of any set of $n+1$ steps, at least one is very good. As in the previous lemma, condition (5) actually needs only hold at this set of $n+1$ steps, as long as the update is not made when that condition fails.

Lemma 2.2 Suppose the assumptions of Lemma 2.1 are satisfied for the sequences $\{x_k\}$ and $\{B_k\}$, and that in addition there exists M for which $\|B_k\| \leq M$ for all k . Then there exist $K \geq 0$ with $S = \{s_{k_j} : K \leq k_1 \leq \dots \leq k_{n+1}\}$

and an index k_m , $m \in \{1, 2, \dots, n\}$, such that

$$\frac{\|(B_{k_m} - \nabla^2 f(x^*))s_{k_m}\|}{\|s_{k_m}\|} < \bar{c} \varepsilon_s^{\frac{1}{n}}$$

where

$$\varepsilon_s = \max_{i \leq j \leq n+1} \{ \|x_{k_j} - x^*\| \}$$

and

$$\bar{c} = 4 \left[\gamma + \sqrt{n} \frac{\gamma}{r} \left(\frac{2}{r} + 1 \right)^{k_{n+1} - k_1 - 2} + M + \|\nabla^2 f(x^*)\| \right]$$

In order to use this lemma to establish a rate of convergence, the following result is needed, which is closely related to the well-known superlinear convergence characterization of Dennis and Moré (1974).

Lemma 2.3 Suppose the function f satisfies Assumption 2.1. If the quantities

$$e_k = \|x_k - x^*\| \text{ and } \frac{\|(B_k - \nabla^2 f(x^*))s_k\|}{\|s_k\|}$$

are sufficiently small, and if $B_k s_k = -\nabla f(x_k)$, then

$$\|x_k + s_k - x^*\| \leq \left\| \left(\nabla^2 f(x^*) \right)^{-1} \right\| \left[2 \frac{\|(B_k - \nabla^2 f(x^*))s_k\|}{\|s_k\|} e_k + \frac{\gamma}{2} e_k^2 \right].$$

Algorithm of mSR1 Method

Data: $x_0 \in \Re^n$, $H_0 \in R^{n \times n}$, $\varepsilon = 10^{-6} \in \Re$, ω_0 , α_0 .

1. $k = 0$
2. $g_k = \nabla f(x_k)$
3. if $\|g_k\| \leq \varepsilon$
then
 - 3.1 stop. ! print the results
4. $d_k = -H_k g_k - \omega_k g_k$
5. $\alpha_k = \arg \min_{\alpha > 0} f(x_k - \alpha g_k)$
6. $x_{k+1} = x_k + \alpha_k d_k$
7. case true of
 - 7.1 $d_k^T g_k < 0$:

$$7.1.1 \quad H_{k+1} = H_k + \frac{(s_k - H_k y_k)(s_k - H_k y_k)^T}{y_k^T (s_k - H_k y_k)}$$

$$7.2 \quad d_k^T g_k > 0 \text{ or } k = 1:$$

$$7.2.1 \quad H_{k+1} = \tilde{\delta}_k I$$

$$7.2.2 \quad \tilde{\delta}_k = \frac{s_k^T s_k}{y_k^T s_k} - \left[\frac{(s_k^T s_k)^2}{(y_k^T s_k)^2} - \frac{s_k^T s_k}{y_k^T y_k} \right]^{\frac{1}{2}}$$

$$7.2.3 \quad d_{k+1} = -\tilde{\delta}_k H_k g_k - \omega_k g_k$$

$$7.3 \quad \text{default : ! } d_k^T g_k = 0$$

$$7.3.1 \quad \text{Check computation}$$

$$7.3.2 \quad \text{stop}$$

$$8. \quad k = k + 1$$

$$9. \quad \text{return .}$$

3. NUMERICAL RESULTS

Using the above algorithm, a few unconstrained nonlinear problems are tested by using Matlab 6.5 with $\omega_k = 0.001$ and their numerical results are shown in the Table 1. We list all functions which have been so used and they will form the basis for applying the comparison.

1. Powell Function

$$f(x_1, x_2, x_3, x_4) = (x_1 + 10x_2)^2 + 5(x_3 - x_4)^2 + (x_2 - 2x_3)^4 + 10(x_1 - x_4)^4$$

2. Wood Function

$$f(x_1, x_2, x_3, x_4) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2 + 90(x_4 - x_3)^2 + 10.1[(x_2 - 1)^2 + (x_4 - 1)^2] + 19.8(x_2 - 1)(x_4 - 1)$$

3. Shallow Function

$$f(x_1, x_2, x_3, x_4) = (x_1^2 - x_2)^2 + (1 - x_1)^2 + (x_3^2 - x_4)^2 + (1 - x_3)^2$$

4. Cube Function

$$f(x_1, x_2, x_3, x_4) = 100(x_1^3 - x_2)^2 + (1 - x_1)^2 + 100(x_3^3 - x_4)^2 + (1 - x_3)^2$$

5. Rosenbrock Function

$$f(x_1, x_2, x_3, x_4) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2 + 100(x_4 - x_3)^2 + (1 - x_3)^2$$

Function and initial point, x_0	Optimal point, x^*	Number of iterations	
		SR1	mSR1

Powell (-3,-1,0,1) (-6,-2,0,2) (-9,-3,0,3)	(0,0,0,0)	33 28 36	36 38 35
Wood (-3,-1,-3,-1) (-6,-2,-6,-2) (-9,-3,-9,-3)	(1,1,1,1)	40 127 90	44 144 53
Shalow (-2,-2,-2,-2) (-20,-20,-20,-20) (-200,-200,-200,-200)	(1,1,1,1)	10 11 11	10 11 11
Cube (-1.2,1.6, -1.2,1.6) (-3.6,6.4, -3.6,6.4) (-6,8, -6,8)	(1,1,1,1)	29 59 291	26 47 189
Rosenbrock (1.2, 2, 1.2, 2) (2.4, 4, 2.4, 4)	(1,1,1,1)	20 31	20 22

Table 1: The comparison between SR1 and mSR1

4. CONCLUSION

According to the numerical result in Table 1, we can conclude that the mSR1 method is more effective to obtain the optimum solution when the initial point is far from the solution, x^* .

5. REFERENCE

1. Byrd, R.H., Khalfan, H.F., dan Schnabel, R.B. (1993). A Theoretical and Experimental Study of the Symmetric Rank One Update, *SIAM Journal on Optimization*, **3**:1-24.
2. Conn, A.R., Gould, N.I.M. dan Toint, Ph. (1988a). Global Convergence of A Class of Trust Region Algorithms for Optimization with Simple Bounds, *SIAM Journal of Numerical Analysis*, **25**(2):433-460.
3. Conn, A.R., Gould, N.I.M. dan Toint, Ph. (1988b). Testing A Class of Methods for Solving Minimization Problems with Simple Bounds on the

M.B MAMAT, K.H. SOW , Y.B DASRIL, AND I.B MOHD

Variables, *Mathematics of Computation*, **50(2)**:399-430.

4. Conn, A.R., Gould, N.I.M. dan Toint, Ph.L. (1991). Convergence of Quasi-Newton Matrices Generated by the Symmetric Rank One Update, *Mathematical Programming*, **50**:177-195.
5. Dennis, J.E. Jr. dan Moré, J.J. (1974). A Characterization of Superlinear Convergence and Its Application to Quasi-Newton Methods, *Mathematics of Computation*, **28**:549- 560.
6. Malik, A.H., Mansor, M. dan Leong, W.J. (2001). *Scaling Symmetric-Rank One Update for Unconstrained Optimization*, Tech. Report Bil 1, Universiti Putra Malaysia.

M.B Mamat, K.H. Sow , Y.B Dasril, And I.B Mohd: Department of Mathematics, Universiti Malaysia Terengganu, Malaysia
E-mail: mus@umt.edu.my

FUNCTIONAL ANALYSIS OF THE LINEARIZED STOKES PROBLEM IN A HETEROGENEOUS MEDIA

H. A. PARHUSIP

Abstract. Functional analysis of the linearized Stokes problem in a heterogeneous media is shown here. The linearized Stokes problem here means that the nonlinearized Stokes problem is linearized with a perturbation method. The governing equations are called cell problems. These cell problems are usual linear Stokes problems with an additional source term. Therefore the known theories of linear Stokes problem in a heterogeneous media can be used. We obtain the existence and uniqueness to the governing cell problems. One may find a relation between functional analysis and an application on this paper such that readers may grasp up an idea for creating a mathematical model from a real problem and giving an analysis using functional analysis.

Key words and Phrases: Stokes problem, cell problems, existence and uniqueness

1. INTRODUCTION

Polymer to control the mobility of injected water has been employed for many years in Enhanced Oil Recovery (EOR) applications [7]. Enhanced Oil Recovery (EOR) is a generic term for techniques for increasing the amount of oil that can be extracted from an oil field. Using EOR, 30-60 efficiency and reduce channeling and breakthrough, has been used by many operators to increase oil recovery. Therefore polymer flow is becoming one of important research topics. Polymer flow is considered as non-Newtonian fluid (fluid with unconstant viscosity).

We consider incompressible non-Newtonian fluid with Carreau model viscosity flows in a porous medium. We assume the porous medium is a bounded

2000 Mathematics Subject Classification: 35B27; 46E35; 46E30.

domain $\Omega \subset \mathbb{R}^n$ consisting of a periodic repetition of an elementary square cell $\mathcal{Y}^\varepsilon$ of size ε . This elementary cell contains the fluid part $\mathcal{F}^\varepsilon \subset \mathcal{Y}^\varepsilon$ and the solid part $\mathcal{S}^\varepsilon = \overline{\Omega} \setminus \mathcal{F}^\varepsilon$. If x stands for macrovariable with $x \in \Omega$ and $y = \frac{x}{\varepsilon}$ denotes microvariable then we assume that the velocity u^ε and the pressure p^ε follow the usual asymptotic expansion in the homogenization method (see for detail in [2] [5]), i.e

$$u^\varepsilon = u_0(x, y) + \varepsilon u_1(x, y) + \dots \quad (1)$$

$$p^\varepsilon = p_0(x) + \varepsilon p_1(x, y) + \dots \quad (2)$$

We obtain the nonlinear stationary two scale Stokes problem

$$\begin{aligned} \nabla_y \cdot [\eta(\mathcal{D}(u_0(x, y))) \mathcal{D}(u_0(x, y))] &= \nabla_x p_0(x) + \nabla_y p_1(x, y) && \text{in } \Omega \times \mathcal{F}, \\ \nabla_y \cdot u_0(x, y) &= 0 && \text{in } \Omega \times \mathcal{F}, \\ u_0(x, y) &= 0 && \text{on } \Omega \times \mathcal{S}, \\ \nabla_x \cdot \int_{\mathcal{Y}} u_0(x, y) dy &= 0 && \text{on } \Omega, \\ (u_0, p_1) &\text{ are periodic in } y, \end{aligned} \quad (3)$$

with viscosity is given by Carreau model, i.e

$$\eta(\mathcal{D}(u_0)) = (\eta_0 - \eta_\infty) \left(1 + \lambda \|\mathcal{D}(u_0)\|_2^2 \right)^{(r/2-1)} + \eta_\infty. \quad (4)$$

The pressure gradient $\nabla_x p_0(x)$ is usually given. The dependent variables are $u_0(x, y)$ and $p_1(x, y)$. Compared to $p_0(x)$, the pressure $p_1(x, y)$ is in the order of ε in the equation (2). The variables x and y appear on the $u_0(x, y)$ and $p_1(x, y)$ variables with $y = x/\varepsilon$ to present heterogeneities of porous medium. Both variables are treated independently. We have $\eta_0 > \eta_\infty > 0$, $\lambda \rightarrow 0$, $1 < r < 2$ and the deformation tensor

$$\mathcal{D}(u_0(x, y)) = \frac{1}{2} (\nabla_y u_0(x, y) + \nabla_y^T u_0(x, y))$$

with its norm

$$\|D(u_0)\|_2^2 = \text{Tr} (D(u_0) D^T(u_0)).$$

In all cases, we will use the definition for the real tensor norm $(.)$ as

$$\|(\cdot)\|_2^2 = \|(\cdot)\|_{L^2(\mathcal{F})^{n^2}}^2 = \text{Tr} ((\cdot)(\cdot)^T) \quad (5)$$

where Tr denote the traces of $(.)$ and $(.)^T$ denotes transpose of $(.)$.

Solution pair $(u_0(x, y), p_1(x, y))$ is the solution Stokes problem when $\varepsilon \rightarrow 0$. The uniqueness for problem (3) is shown in [5]. One way for solving the two scale problem (3) is introducing an asymptotic expansion with respect to the small parameter λ , i.e $\psi(x, y) = \psi^{(0)} + \lambda \psi^{(1)} \dots$ with $\psi(x, y) = (u_0(x, y), p_1(x, y))^T$. By

considering only uni-directional pressure gradient, the expansion leads to Newtonian cell problems (problems depend on microvariable y). This approach and including the numerical simulation for isotropic and anisotropic porous cells are shown in [2].

The purpose of this paper is to show the existence and uniqueness of the cell problem at $O(\lambda)$. The external force in this cell problem has a form $f = C \operatorname{div} F$, where F is a matrix and the constant C depends on the viscosity parameters (see Section). The analysis is similar to the Newtonian case (in the case of constant viscosity).

Existence and uniqueness are fundamental questions in most of mathematical problems. Philosophically means that we need a foundation such that a solution exists to the governed mathematical model and its trueness (it may means a uniqueness) is based on the used spaces. Spaces here roughly give places for solutions that they are correct (well defined) on these spaces.

We need to introduce some functional spaces in microvariable y . All spaces are defined for vector valued functions with periodic in y and we refer to the book of Sohr[3].

We denote

$$C_0^\infty(\mathcal{F}) = \{v \in C^\infty(\mathcal{F}); \operatorname{supp} v \text{ compact}, \operatorname{supp} v \in \mathcal{F}\},$$

and

$$C_{0,\sigma}^\infty(\mathcal{F}) = \{v \in C_0^\infty(\mathcal{F})^n; \operatorname{div}_y v = 0; v \text{ is } y\text{-periodic}\}.$$

Notation $C^\infty(\mathcal{F})$ means a space for any continuous function that is infinitely differentiable in $\mathcal{F}$. We say that $v \in C_0^\infty(\mathcal{F})^n$ means $v = (v_1, \dots, v_n)$, with $v_j \in C_0^\infty(\mathcal{F})$ for $j = 1, \dots, n$. Introducing notation $D_j := \frac{\partial}{\partial y_j}$, $j = 1, \dots, n$, and $\nabla := (D_1, \dots, D_n)$ then $\operatorname{div}_y v = \nabla \cdot v = D_1 v_1 + \dots + D_n v_n$.

The Sobolev space $W^{1,2}(\mathcal{F})^n$ is defined as

$$W^{1,2}(\mathcal{F})^n = \left\{ u : u \in L^2(\mathcal{F})^n, \nabla u \in L^2(\mathcal{Y})^{n^2}; u \text{ is } y\text{-periodic} \right\}$$

with its norm

$$\|u\|_{W^{1,2}(\mathcal{F})^n} = \left(\|u\|_{L^2(\mathcal{F})^n} + \|\nabla u\|_{L^2(\mathcal{F})^{n^2}} \right).$$

Recall definition of the Sobolev space $W^{-k,q}(\mathcal{F})$ of negative order $-k$ is defined as the dual space of $\mathcal{W}_0^{k,q'}$ and we write

$$W^{-k,q}(\mathcal{F}) := W_0^{k,q'}(\mathcal{F}) \text{ with } \frac{1}{q} + \frac{1}{q'} = 1. \quad (6)$$

This means, $W^{-k,q}(\mathcal{F})$ is the space of all linear functionals

$$f : \psi \mapsto [f, \psi], \psi \in W_0^{k,q'}(\mathcal{F})$$

which are continuous in $\|\psi\|_{k,q'}$. The function f is continuous in $\|\psi\|_{k,q'}$ iff there is a constant $C = C(f) > 0$ such that

$$|[f, \psi]| \leq C \|\psi\|_{k,q'}. \quad (7)$$

We introduce

$$W_{0,\sigma}^{1,2}(\mathcal{F}) = \{u : u \in W^{1,2}(\mathcal{F})^n, \operatorname{div} u = 0, u|_S = 0; u - \text{periodic in } y\}.$$

The $W_{0,\sigma}^{1,2}(\mathcal{F})$ becomes a Hilbert space with a scalar product

$$\langle \nabla u, \nabla v \rangle_{\mathcal{F}} = \langle \nabla u, \nabla v \rangle = \int_{\mathcal{F}} (\nabla u) \cdot (\nabla v) dy.$$

All the definitions above will be used frequently in Section .

2. Functional analysis on the cell problem $\mathcal{O}(\lambda)$

2.2. Normalized Asymptotic expansion of the problem (3)

The unknown $u_0(x, y), p_1(x, y)$ of the problem (3) are assumed to be a power series in λ , i.e

$$u_0(x, y) = u_0^{(0)} + \lambda u_0^{(1)} + \dots, \quad (8)$$

$$p_1(x, y) = p_1^{(0)} + \lambda p_1^{(1)} + \dots \quad (9)$$

The chosen expansion is not unique. One may define more generally [1]. We use the regular expansion (8)-(9) to introduce the first idea for using an asymptotic expansion. Since the viscosity Carreau in (3) depends on the unknown, then we assume that it is a power series in λ and its asymptotic expansion is

$$\eta \approx \eta^{(0)} + \lambda \eta^{(1)} = \eta_0 + \lambda(\eta_0 - \eta_\infty) \left(\frac{r}{2} - 1 \right) \left\| \mathcal{D}(u_0^{(0)}) \right\|_2^2 + \dots \quad (10)$$

Using (8)-(10) in problem (3) we get two scale Stokes problem for each order of λ . We present only up to the first order. The two scale problem at $O(1)$ is

$$\begin{aligned} \eta_0 \nabla_y \cdot \left[\mathcal{D}(u_0^{(0)}(x, y)) \right] &= \nabla_x p_0 + \nabla_y p_1^{(0)} && \text{in } \Omega \times \mathcal{F}, \\ \nabla_y \cdot u_0^{(0)}(x, y) &= 0 && \text{in } \Omega \times \mathcal{F}, \\ u_0^{(0)}(x, y) &= 0 && \text{on } \Omega \times \mathcal{S}, \\ \nabla_x \cdot \int_{\mathcal{Y}} u_0^{(0)}(x, y) dy &= 0 && \text{on } \Omega, \\ (u_0^{(0)}, p_1^{(0)}) &&& \text{are periodic in } y. \end{aligned} \quad (11)$$

The Stokes problem at $O(\lambda)$ is

$$\begin{aligned} \eta_0 \nabla_y \cdot \left[\mathcal{D}(u_0^{(1)}(x, y)) \right] &= \nabla_y p_1^{(1)} + C_\eta \operatorname{div}_y F && \text{in } \Omega \times \mathcal{F}, \\ \nabla_y \cdot u_0^{(1)}(x, y) &= 0 && \text{in } \Omega \times \mathcal{F}, \\ u_0^{(1)}(x, y) &= 0 && \text{on } \Omega \times \mathcal{S}, \\ \nabla_x \cdot \int_{\mathcal{Y}} u_0^{(1)}(x, y) dy &= 0 && \text{on } \Omega, \\ (u_0^{(1)}, p_1^{(1)}) &&& \text{are periodic in } y, \end{aligned} \quad (12)$$

with

$$C_\eta = -(\eta_0 - \eta_\infty) \left(\frac{r}{2} - 1 \right) \quad \text{and} \quad F = \left(\left\| \mathcal{D}u_0^{(0)} \right\|_2^2 \mathcal{D}u_0^{(0)} \right). \quad (13)$$

For deriving the cell problem we linearize the problems (11)-(12) by introducing an ansatz to the velocity, that is

$$u_0(x, y) = u_0^{(0)}(x, y) + \lambda u_0^{(1)}(x, y) = w_k^{(0)}(y) g(\partial_k p_0)^{(0)} + \lambda w_k^{(1)}(y) g(\partial_k p_0)^{(1)}, \quad (14)$$

with

$$g(\partial_k p_0)^{(0)} = \partial_k p_0 \quad \text{and} \quad g(\partial_k p_0)^{(1)} = (\partial_k p_0)^3. \quad (15)$$

The subscript k denotes that we choose a pressure gradient in k -direction nonzero, i.e. $\partial_k p(x) \neq 0$ (see [4] for its application). Equation (15) is a *trick* that one may decouple the two scale problem (problem with two variables, i.e. (11)-(12)) into a cell problem (problem that contains only microvariable). It reduces the dimension of problem (11)-(12). The cell problem at $\mathcal{O}(1)$ reads as

$$\begin{aligned} \eta_0 \Delta_{yy} w_k^{(0)}(y) &= \nabla_y \pi_k^{(0)}(y) - e_k && \text{for } y \in \mathcal{F}, \\ \nabla_y \cdot w_k^{(0)}(y) &= 0 && \text{for } y \in \mathcal{F}, \\ w_k^{(0)}(y) &= 0 && \text{for } y \in \mathcal{S}, \\ (w_k^{(0)}, \pi_k^{(0)}) &&& \text{are periodic in } y. \end{aligned} \quad (16)$$

The cell problem at $\mathcal{O}(\lambda)$ reads as

$$\begin{aligned} \eta_0 \Delta_{yy} w_k^{(1)}(y) &= \nabla_y \pi_k^{(1)}(y) + f(w_k^{(0)}) && \text{for } y \in \mathcal{F}, \\ \nabla_y \cdot w_k^{(1)}(y) &= 0 && \text{for } y \in \mathcal{F}, \\ w_k^{(1)}(y) &= 0 && \text{for } y \in \mathcal{S}, \\ (w_k^{(1)}, \pi_k^{(1)}) &&& \text{are periodic in } y, \end{aligned} \quad (17)$$

where

$$\begin{aligned} f(w_k^{(0)}) &= -(\eta_0 - \eta_\infty) \left(\frac{r}{2} - 1 \right) \nabla_y \left(\left\| \mathcal{D}w_k^{(0)}(y) \right\|_2^2 \mathcal{D}w_k^{(0)}(y) \right) \\ &= -C_\eta \operatorname{div}_y F(y). \end{aligned} \quad (18)$$

Remark: If we add the assumption that the pressure $p_0(x)$ has also regular expansion with respect to λ then we have the cell problem at $\mathcal{O}(\lambda)$ as

$$\begin{aligned}\eta_0 \Delta_{yy} w_k^{(1)}(y) &= \nabla_y \pi_k^{(1)}(y) - e_k + f(w_k^{(0)}) & \text{for } y \in \mathcal{F}, \\ \nabla_y \cdot w_k^{(1)}(y) &= 0 & \text{for } y \in \mathcal{F}, \\ w_k^{(1)}(y) &= 0 & \text{for } y \in \mathcal{S}, \\ (w_k^{(1)}, \pi_k^{(1)}) & \text{ are periodic in } y.\end{aligned}$$

It is wellknown from [6] that the cell problem (16) admits a unique pair solution $(w_k^{(0)}(y), \pi_k^{(0)}(y))$ in $(W_{0,\sigma}^{1,2}(\mathcal{F}) \times L_0^2(\mathcal{F}))$. We have the external force $f = (f_1, \dots, f_n)$ in this cell problem with a special form

$$f = C \operatorname{div} F \quad (19)$$

and the constant $C > 0$ depends on the parameters of the fluid. We have solution from the cell problem $\mathcal{O}(1)$ that $w = (w_1, w_2)$ in $W_{0,\sigma}^{1,2}(\mathcal{F})$. Hence F is well defined and therefore $F \in L^2(\mathcal{F})^4$.¹ This paper is devoted to discuss the analysis for the cell problem (17)-(18). We use mainly the result in [3] for this purpose. Without losing its generality, some considerations are designed for only two dimensional case.

2.2. Existence and Uniqueness of the solution of the cell problem $\mathcal{O}(\lambda)$

In the following section, we write the solution of the cell problem at $\mathcal{O}(1)$ as $w_k^{(0)}(y) = w_k$. Let us define the functional

$$f : \psi \mapsto [f, \psi], \quad \forall \psi \in C_0^\infty(\mathcal{Y})^n.$$

The notion $[f, \psi]$ means the value f at ψ and it is defined as

$$[f, \psi] := \langle f, \psi \rangle = \int_{\mathcal{F}} f \cdot \psi dy. \quad (20)$$

To start the analysis for the cell problem (17) we use Lemma in [3], pp 61. For any bounded fluid domain $\mathcal{F} \subseteq R^n$, the external force in the cell problem (17) has a representation

$$f = -C_\eta \operatorname{div}_y F \text{ and } F = \|\mathcal{D}(w_k)\|_2^2 \mathcal{D}(w_k) \quad (21)$$

¹ $F \in L^2(\mathcal{F})^{n^2}$ iff $\int_{\mathcal{F}} |f_{ij}|^2 dy < +\infty$

in the sense of distributions. The functional $f = -C_\eta \operatorname{div}_y F$ in the sense of distributions means that for $v \in C_0^\infty(\mathcal{F})^n$

$$\begin{aligned} [f, v] &= -[C_\eta \operatorname{div}_y F, v] \\ &= -C_\eta \langle \operatorname{div}_y \|\mathcal{D}(w_k)\|_2^2 \mathcal{D}(w_k), v \rangle \\ &= C_\eta \langle \|\mathcal{D}(w_k)\|_2^2 \mathcal{D}(w_k), \nabla_y v \rangle \\ &= C_\eta \int_{\mathcal{F}} \|\mathcal{D}(w_k)\|_2^2 \mathcal{D}(w_k) \cdot \nabla_y v dy \end{aligned} \quad (22)$$

with

$$\|\mathcal{D}(w_k)\|_2^2 = \operatorname{Tr}(\xi \xi^T), \quad \xi = \mathcal{D}(w_k) = \frac{1}{2} (\nabla_y w_k + \nabla_y^T w_k). \quad (23)$$

By definition, equation (22) for two dimensional case with $v = (v_1, v_2)$ and $w_k = (w_1, w_2)$ means

$$\begin{aligned} \|\mathcal{D}(w_k)\|_2^2 \mathcal{D}(w_k) \cdot \nabla_y v &= \|\mathcal{D}(w_k)\|_2^2 \frac{\partial w_1}{\partial y_1} \frac{\partial v_1}{\partial y_1} \\ &+ \frac{1}{2} \|\mathcal{D}(w_k)\|_2^2 \left(\frac{\partial w_1}{\partial y_2} + \frac{\partial w_2}{\partial y_1} \right) \left(\frac{\partial v_1}{\partial y_2} + \frac{\partial v_2}{\partial y_1} \right) \\ &+ \|\mathcal{D}(w_k)\|_2^2 \left(\frac{\partial w_2}{\partial y_2} \frac{\partial v_2}{\partial y_1} \right). \end{aligned} \quad (24)$$

By using (6)-(7), the functional f is continuous in $\|\nabla_y v\|_2$ since

$$\begin{aligned} |[f, v]| &= |C_\eta \langle \|\mathcal{D}(w_k)\|_2^2 \mathcal{D}(w_k), \nabla_y v \rangle| \\ &\leq |C_\eta| \left\| \|\mathcal{D}(w_k)\|_2^2 \mathcal{D}(w_k) \right\| \|\nabla_y v\|_2. \end{aligned} \quad (25)$$

By setting $C = C(f)$ with $C(f) := |C_\eta| \left\| \|\mathcal{D}(w_k)\|_2^2 \mathcal{D}(w_k) \right\|$ we get

$$|[f, v]| \leq C(f) \|\nabla_y v\|_2. \quad (26)$$

Hence $f \in W^{-1}(\mathcal{F})^2$ by following the definition (6).

Now we will show the existence of the weak solution of the cell problem (17). Before doing that, we recall some important facts without proofs from [3].

Lemma 0.0.1. ([3], pp 73) *Let $\mathcal{F} \subseteq \mathcal{R}^n, n \geq 2$, be an arbitrary fluid domain. Let $\mathcal{F}_0 \subseteq \mathcal{F}$ be a bounded subdomain with $\overline{\mathcal{F}_0} \subseteq \mathcal{F}$, $\mathcal{F}_0 \neq \emptyset$, and let $1 < q < \infty$. Notation $\overline{\mathcal{F}_0}$ is the closure of $\mathcal{F}_0$, that is $\mathcal{F}_0$ together with its boundary: $\overline{\mathcal{F}_0} = \mathcal{F}_0 \cup \partial \mathcal{F}_0$. Suppose $f \in W_{loc}^{-1,q}(\mathcal{F})^n$ satisfies*

$$[f, v] = 0 \text{ for all } v \in C_{0,\sigma}^\infty(\mathcal{F})$$

Then there exists a unique $\pi_k^{(1)} \in L_{loc}^q(\mathcal{F})$ satisfying $\nabla_y \pi_k^{(1)} = f$ in the sense of distributions and

$$\int_{\mathcal{F}_0} \pi_k^{(1)} dy = 0. \quad (27)$$

Proof. See [3], pp 73.

The next Lemma shows that adding a regularity in a domain, we obtain an estimate for the pressure $\pi_k^{(1)}$.

Lemma 0.0.2. *Let $\mathcal{F} \subseteq \mathcal{R}^n, n \geq 2$, be a bounded Lipschitz domain, let $\mathcal{F}_0 \subseteq \mathcal{F}$ be any fluid subdomain and let $1 < q < \infty$. Suppose $f \in W^{-1,q}(\mathcal{F})^n$ satisfies*

$$[f, v] = 0 \text{ for all } v \in C_{0,\sigma}^\infty(\mathcal{F}).$$

Then there exists a unique $\pi_k^{(1)} \in L^q(\mathcal{F})$ satisfying $\nabla_y \pi_k^{(1)} = f$ in the sense of distributions and

$$\int_{\mathcal{F}_0} \pi_k^{(1)} dy = 0.$$

The estimate for the pressure $\pi_k^{(1)}$ is

$$\left\| \pi_k^{(1)} \right\|_{L^q(\mathcal{F})} \leq C_1 \|f\|_{W^{-1,q}(\mathcal{F})^n} \leq C_1 C_2 \left\| \pi_k^{(1)} \right\|_{L^q(\mathcal{F})}$$

with constant $C_1 = C_1(q, \mathcal{F}_0, \mathcal{F}) > 0$ and $C_2 = C_2(n) > 0$.

Proof. See [3], pp 76-77.

Using Lemma (0.0.1)-(0.0.2) and for a given force f defined by equation (18), the velocity cell $w_k^{(1)}$ satisfies Stokes equation with the pressure $\pi_k^{(1)}$ is defined up to constant.

Lemma 0.0.3. *Let $\mathcal{F} \subseteq \mathcal{R}^n$ be any fluid domain, let $\mathcal{F}_0 \subseteq \mathcal{F}, \mathcal{F}_0 \neq \emptyset$, be a bounded fluid subdomain with $\overline{\mathcal{F}_0} \subseteq \mathcal{F}$ and let $f = -C_\eta \operatorname{div}_y F = -C_\eta \operatorname{div}_y \left(\|\mathcal{D}(w_k)\|_2^2 \mathcal{D}(w_k) \right)$*

Suppose $w_k^{(1)} \in W_{0,\sigma}^{1,2}(\mathcal{F})$ is a weak solution of the cell problem (17) with force f . Then there exists a uniquely determined $\pi_k^{(1)} \in L_{loc}^2(\mathcal{F})$ satisfying $\int_{\mathcal{F}_0} \pi_k^{(1)} dy = 0$ and

$$-\eta_0 \Delta_{yy} w_k^{(1)} + \nabla_y \pi_k^{(1)} = f \quad (28)$$

in the sense of distributions.

A pair $(w_k^{(1)}, \pi_k^{(1)}) \in W_{0,\sigma}^{1,2}(\mathcal{F}) \times L_{loc}^2(\mathcal{F})$ of Stokes cell problem (17) with force f exists iff (28) holds in the sense of distributions; $\pi_k^{(1)}$ is uniquely determined by $w_k^{(1)}$ under the additional condition $\int_{\mathcal{F}_0} \pi_k^{(1)} dy = 0$.

Proof. Let $w_k^{(1)} \in W_{0,\sigma}^{1,2}(\mathcal{F})$ be a weak solution and consider the functional $G : v \mapsto [G, v], v \in C_0^\infty(\mathcal{F})^n$, defined by

$$[G, v] := [f + \eta_0 \Delta_{yy} w_k^{(1)}, v] = [f, v] + \eta_0 [\Delta_{yy} w_k^{(1)}, v].$$

Using the given f and the weak derivatives definition, we get

$$\begin{aligned} [G, v] &= -C_\eta \langle \operatorname{div}_y \|\mathcal{D}(w_k)\|_2^2 \mathcal{D}(w_k), v \rangle - \eta_0 \langle \nabla_y w_k^{(1)}, \nabla_y v \rangle \\ &= C_\eta \langle \|\mathcal{D}(w_k)\|_2^2 \mathcal{D}(w_k), \nabla_y v \rangle - \eta_0 \langle \nabla_y w_k^{(1)}, \nabla_y v \rangle. \end{aligned} \quad (29)$$

Then for each bounded subdomain $\mathcal{F}' \subseteq \mathcal{F}$ with $\overline{\mathcal{F}'} \subseteq \mathcal{F}$ we obtain

$$\begin{aligned} |[G, v]| &\leq |C_\eta| \|\mathcal{D}(w_k)\|_2^3 \|\nabla_y v\|_2 + \eta_0 \left\| \nabla_y w_k^{(1)} \right\|_2 \|\nabla_y v\|_2 \\ &= \left(|C_\eta| \|\mathcal{D}(w_k)\|_2^3 + \eta_0 \left\| \nabla_y w_k^{(1)} \right\|_2 \right) \|\nabla_y v\|_2 \end{aligned} \quad (30)$$

for all $v \in C_0^\infty(\mathcal{F}')^n$. We define $C(G) = \left(|C_\eta| \|\mathcal{D}(w_k)\|_2^3 + \eta_0 \left\| \nabla_y w_k^{(1)} \right\|_2 \right)$, equation (30) becomes

$$|[G, v]| \leq C(G) \|\nabla_y v\|_2. \quad (31)$$

This shows that $G \in W_{loc}^{-1,2}(\mathcal{F}')^n$ by following definition equation (7). Then by Lemma (0.0.1) we obtain a unique $\pi_k^{(1)} \in L_{loc}^2(\mathcal{F}')$ with $G = \nabla_y \pi_k^{(1)}$ in the sense of distributions and with $\int_{\mathcal{F}_0} \pi_k^{(1)} dy = 0$.

Finally we use the theorem in ([3], pp 113) for existence and uniqueness of the solution for the cell problem $\mathcal{O}(\lambda)$, i.e (17).

Theorem 0.0.1. *Let $\mathcal{F} \subseteq \mathcal{R}^n$ in Ω be any fluid domain, let $\mathcal{F}_0 \subseteq \mathcal{F}$, $\mathcal{F}_0 \neq \emptyset$, be bounded fluid subdomain with $\overline{\mathcal{F}_0} \subseteq \mathcal{F}$, and let $f = -C_\eta \operatorname{div}_y \|\mathcal{D}(w_k)\|_2^2 \mathcal{D}(w_k)$. Then there exists a unique pair $(w_k^{(1)}, \pi_k^{(1)}) \in W_{0,\sigma}^{1,2}(\mathcal{F}) \times L_{loc}^2(\mathcal{F})$ satisfying $\int_{\mathcal{F}_0} \pi_k^{(1)} dy = 0$ and*

$$-\eta_0 \Delta_{yy} w_k^{(1)} + \nabla_y \pi_k^{(1)} = f, \quad \operatorname{div}_y w_k^{(1)} = 0 \quad (32)$$

in the sense of distributions; $w_k^{(1)}$ is a weak solution and $\pi_k^{(1)}$ an associated pressure of the Stokes cell problem (17) with force f .

Moreover, $w_k^{(1)}$ satisfies the inequality

$$\left\| \nabla_y w_k^{(1)} \right\|_{L^2(\mathcal{F})^4} \leq \eta_0^{-1} C_\eta \|\mathcal{D}(w_k)\|_2^3. \quad (33)$$

If $\mathcal{F}$ is a bounded Lipschitz domain, then $\pi_k^{(1)} \in L^2(\mathcal{F})$ and

$$\left\| \pi_k^{(1)} \right\|_{L^2(\mathcal{F})} \leq C \|\mathcal{D}(w_k)\|_2^3 \quad (34)$$

with $C = C(\mathcal{F}_0, \mathcal{F}, C_\eta) > 0$.

Proof. Following equations (22) and (26), we define the functional

$$\begin{aligned} f : v \mapsto [f, v] &= [-C_\eta \operatorname{div}_y \|\mathcal{D}(w_k)\|_2^2 \mathcal{D}(w_k), v] \quad v \in C_0^\infty(\mathcal{F})^n \\ &= C_\eta \langle \|\mathcal{D}(w_k)\|_2^2 \mathcal{D}(w_k), \nabla_y v \rangle \end{aligned} \quad (35)$$

This is the first term of equation (29). It is continuous in $\|\nabla_y v\|_2$ since

$$|[f, v]| \leq |C_\eta| \|D(w_k)\|_2^3 \|\nabla_y v\|_2.$$

The second term of equation (29) is considered as the scalar product in $W_{0,\sigma}^{1,2}(\mathcal{F})$. Therefore we use in $W_{0,\sigma}^{1,2}(\mathcal{F})$ the scalar product

$$\eta_0 \langle \nabla_y w_k^{(1)}, \nabla_y v \rangle = \eta_0 \int_{\mathcal{F}} (\nabla_y w_k^{(1)}) \cdot (\nabla_y v) dy.$$

This expression reminds us the Riesz representation theorem [8], which yields a uniquely determined $w_k^{(1)} \in W_{0,\sigma}^{1,2}(\mathcal{F})$ satisfying

$$\eta_0 \langle \nabla_y w_k^{(1)}, \nabla_y v \rangle = C_\eta \langle \|D(w_k)\|_2^2 D(w_k), \nabla_y v \rangle, \quad v \in W_{0,\sigma}^{1,2}(\mathcal{F}).$$

Setting $v = w_k^{(1)}$ yields

$$\eta_0 \left\| \nabla_y w_k^{(1)} \right\|_2^2 \leq C_\eta \|D(w_k)\|_2^3 \left\| \nabla_y w_k^{(1)} \right\|_2$$

and therefore

$$\eta_0 \left\| \nabla_y w_k^{(1)} \right\|_2 \leq C_\eta \|D(w_k)\|_2^3 \quad (36)$$

which proves (33).

The pressure part $\pi_k^{(1)}$ in the Stokes cell problem (17) is constructed in this following way. We start with $w_k^{(1)} \in W_{0,\sigma}^{1,2}(\mathcal{F})$, we consider the functional $v \mapsto [G, v]$ with $v \in C_0^\infty(\mathcal{F})^n$ is defined by

$$\begin{aligned} [G, v] &:= [f + \eta_0 \Delta_{yy} w_k^{(1)}, v] \text{ with } v \in C_0^\infty(\mathcal{F})^n \\ &= [f, v] + \eta_0 [\Delta_{yy} w_k^{(1)}, v] \\ &= [f, v] - \eta_0 \langle \nabla_y w_k^{(1)}, \nabla_y v \rangle \\ &= C_\eta \langle \|D(w_k)\|_2^2 D(w_k), \nabla_y v \rangle - \eta_0 \langle \nabla_y w_k^{(1)}, \nabla_y v \rangle \end{aligned} \quad (37)$$

using equation (35)). Then for each bounded subdomain $\mathcal{F}_0 \subseteq \mathcal{F}$ with $\overline{\mathcal{F}_0} \subseteq \mathcal{F}$ we obtain

$$|[G, v]| \leq |C_\eta| \|D(w_k)\|_2^3 \|\nabla_y v\|_2 + \eta_0 \left\| \nabla_y w_k^{(1)} \right\|_2 \|\nabla_y v\|_2, \quad \forall v \in C_0^\infty(\mathcal{F}_0)^n. \quad (38)$$

Using (33) in equation (38), we get

$$|[G, v]| \leq 2|C_\eta| \|D(w_k)\|_2^3 \|\nabla_y v\|_2.$$

This shows that $G \in W_{loc}^{-1,2}(\mathcal{F})$. Then from Lemma (0.0.1) we get a unique $\pi_k^{(1)} \in L_{loc}^2(\mathcal{F})$ satisfying $\nabla_y \pi_k^{(1)} = G$ in the sense of distributions, i.e

$$[\nabla_y \pi_k^{(1)}, v] = - \langle \pi_k^{(1)}, \operatorname{div}_y v \rangle = 0$$

since $\operatorname{div}_y v = 0$ for all $v \in C_{0,\sigma}^\infty(\mathcal{F})$ then (28) is satisfied.

To prove the uniqueness of $(w_k^{(1)}, \pi_k^{(1)})$, let $(\tilde{u}, \tilde{p}) \in W_{0,\sigma}^{1,2}(\mathcal{F}) \times L_{loc}^2(\mathcal{F})$ be another pair solving (28) i.e :

$$-\eta_0 \Delta_{yy} \tilde{u} + \nabla_y \tilde{p} = f, \quad \operatorname{div}_y \tilde{u} = 0. \quad (39)$$

Hence for all $v \in C_{0,\sigma}^\infty(\mathcal{F})$, and again using the weak derivatives definition

$$\eta_0 \langle \nabla_y w_k^{(1)}, \nabla_y v \rangle + \langle \nabla_y \pi_k^{(1)}, v \rangle = \langle f, v \rangle \quad (40)$$

$$\eta_0 \langle \nabla_y \tilde{u}, \nabla_y v \rangle + \langle \nabla_y \tilde{p}, v \rangle = \langle f, v \rangle. \quad (41)$$

Then we see that

$$\eta_0 \langle \nabla_y (w_k^{(1)} - \tilde{u}), \nabla_y v \rangle + \langle (\pi_k^{(1)} - \tilde{p}), \operatorname{div}_y v \rangle = 0, \quad \text{for all } v \in C_{0,\sigma}^\infty(\mathcal{F}).$$

Since $\operatorname{div}_y v = 0$ because $v \in C_{0,\sigma}^\infty(\mathcal{F})$, we apply Lemma (0.0.1) to get

$$\eta_0 \langle \nabla_y (w_k^{(1)} - \tilde{u}), \nabla_y v \rangle = 0.$$

Setting with $v = w_k^{(1)} - \tilde{u}$, we have $\left\| \nabla_y (w_k^{(1)} - \tilde{u}) \right\|_2^2 = 0$. Therefore $w_k^{(1)} = \tilde{u}$. Now we need to find the estimate for the pressure $\pi_k^{(1)}$. If $\mathcal{F}$ is a bounded Lipschitz domain then

$$\begin{aligned} \left\| \pi_k^{(1)} \right\|_{L^2(\mathcal{F})} &\leq C \left\| \nabla_y \pi_k^{(1)} \right\|_{W^{-1,2}(\mathcal{F})^n} \quad (\text{by Poincare inequality}) \\ &= C \left\| \eta_0 \Delta_{yy} w_k^{(1)} + f \right\|_{W^{-1,2}(\mathcal{F})^n} \quad (\text{by Lemma (0.0.2)}) \\ &= C \sup_{0 \neq v \in C_{0,\sigma}^\infty(\mathcal{F})^n} \left(\left| [\eta_0 \Delta_{yy} w_k^{(1)} + f, v] \right| / \left\| \nabla_y v \right\|_2 \right) \quad (\text{by definition}) \left\| \cdot \right\|_{W^{-1,2}(\mathcal{F})^n} \\ &= C \sup_{0 \neq v \in C_{0,\sigma}^\infty(\mathcal{F})^n} \frac{\left| -\eta_0 \langle \nabla_y w_k^{(1)}, \nabla_y v \rangle + C_\eta \langle \mathcal{D}(w_k), \nabla_y v \rangle \right|}{\left\| \nabla_y v \right\|_2} \\ &\leq C \left(\eta_0 \left\| \nabla_y w_k^{(1)} \right\|_2 + C_\eta \left\| \mathcal{D}(w_k) \right\|_2^3 \right), \end{aligned} \quad (42)$$

and from (33) we have $\eta_0 \left\| \nabla_y w_k^{(1)} \right\|_2 \leq C_\eta \left\| \mathcal{D}(w_k) \right\|_2^3$. Therefore (42) becomes

$$\left\| \pi_k^{(1)} \right\|_{L^2(\mathcal{F})} \leq C \left\| \mathcal{D}(w_k) \right\|_2^3$$

with $C = C(\mathcal{F}_0, \mathcal{F}, C_\eta)$.

3. REFERENCES

1. A. MIKELIĆ, Homogenization Theory and Applications to Filtration through Porous Media, *Lecture Notes*, France 2000.
2. H. A. PARHUSIP, *Darcy's law and permeability for the non-Newtonian fluid in porous media with two scale asymptotic expansion*, Dissertation, Institut Teknologi Bandung, 2005.
3. H. SOHR, *The Navier Stokes equations : an elementary functional analytic approach*, Berlin: Birkhaeuser Verlag, 2001.
4. T. GOETZ AND H. A. PARHUSIP, On an Asymptotic Expansion for Carreau Fluids in Porous Media, *Journal of Engineering Mathematics*, Springer Science + Business Media B.V **51 No. 1**, 351–365, 2005.
5. U. HORNUNG, *Homogenization and Porous Media*, Springer-Verlag, New-York, 1997.
6. R. TEMAM, *Navier Stokes Equations, Theory and Numerical Analysis, Studies in Mathematics and its Applications*, North Holland Publishing Company, 1977.

WEB REFERENCES

7. <http://www.snf-oil.com/Function.html>.
8. <http://mathworld.wolfram.com/RieszRepresentationTheorem>.

H. A. PARHUSIP: Department of Industrial Mathematics and Statistics,
Satya Wacana Christian University, Salatiga, Indonesia.
E-mail: pressure1733@hotmail.com or hanna@uksw.edu

THE APPLICATION MATHEMATICS FOR INVERSE ACOUSTICS PROBLEM

RATNA DEWI

Abstract. The problem in acoustics can be divided become two, there are direct problem and inverse problem. Inverse problem is the reverse of direct problem. In direct problem, the information at the surface of acoustic source is known, and the acoustic information at the field is desired. In acoustic inverse problem, the information at the acoustic field is known, and the acoustic information at the source is desired.

This paper is concerned optimization of the solution the acoustic inverse problem involving axisymmetric source in full space. The optimization Conjugate Gradient is used to solve the matrix equation, because of fast convergence. The acoustic parameters at the source are pressure on the surface.

The solution of direct problem and inverse problem may not be obtained analytically, especially when they involve irregular shape bodies. In this respect, the solution may be obtained numerically, for example using the Boundary Element Method. The advantage of using BEM is the reduction of the dimension of the problem being solved, for example three-dimensional problem can be solved using two-dimensional treatment. For axisymmetric sources, the dimension of the problem can be further reduced and one deals only with one-dimensional computation.

For axisymmetric sources, the surface integral appearing in the calculation is reduced to a line integral. By taking advantage of the axisymmetric property of the body in question, the modeling is simplified in that only the generator of the body needs to be discretized. Matrix singularity occurring in the inverse solution is overcome using Singular Value Decomposition (SVD).

The source being evaluated is a radiating half sphere of 0.08-meter radii and the measurement points are given at a distance of 0.01 m from the surface object. The half spherical body was discretized into 4 elements and 9 nodes, with the wave number 0.01. The case with optimization CG solved in fourth iteration, to look at the convergence of this method. Test cases show that, the solution using optimization CG yields better result than the solution with SVD.

Key words: *Inverse acoustics problem, BEM, Axisymmetric, SVD, Optimization Conjugate Gradient*

1. INTRODUCTION

In acoustic inverse problem, the acoustic information at the acoustic field is known, and the acoustic information at the source is desired. Inverse problem may appear in the design involving the determination of acoustic parameters at the source, such as acoustic pressure and impedance on the surface.

The solution of direct problem and inverse problem may not be obtained analytically, especially when they involve irregular shape bodies. In this respect, the solution may be obtained numerically, for example using the Boundary Element Method. The advantage of using BEM is the reduction of the dimension of the problem being solved, for example three-dimensional problem can be solved using two-dimensional treatment[2].

This paper is concerned initial value in optimization of the solution the acoustic inverse problem. The optimization Conjugate Gradient is used to solve the matrix equation,

because of fast convergence [6].

2. METHOD

2.1. The Boundary Element Formulation

The direct boundary element method is based on the Helmholtz integral equation[1,2]. The axisymmetric body in a full space is illustrated in Figure 1.

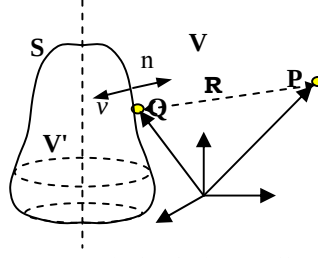


Figure 1. An axisymmetric body in a full space

Where Q is any point on S and P may be on V, S, or V' (the region inside the body bounded by S), $R \equiv |P-Q|$, n is the unit normal on S directed away from the acoustic domain, v_n and p_s , are the surface normal velocity and sound pressure on S. $C(P)$ is a coefficient. The value of $C(P)$ depends on the location of point P. $C(P)$ has the value 4π for $P \in V$, the value 2π for $P \in S$ provided there is a unique tangent to S at such a P, and the value 0 for $P \in V'$. For P at an edge or corner of S, $C(P)$ has the value[1,2] :

$$C(P) = 4\pi + \int_S \frac{\partial}{\partial v} \left(\frac{1}{R(P, Q)} \right) dS(Q) \quad (1)$$

The expression of $C(P)$ in Eq. (1) is valid for all P for arbitrary nonsmooth S.

For axisymmetric body, a cylindrical coordinate system may be used[1-3], that is

$$C(P)\phi(P) = \int_L \left(\phi(Q)K^B(P, Q) - \frac{\partial \phi(Q)}{\partial v} K^A(P, Q) \right) \rho(Q) dL(Q) + 4\pi\phi^I(P) \quad (2)$$

where

$$K^A(P, Q) = \int_0^{2\pi} \frac{e^{-ikR(P, Q)}}{R(P, Q)} d\theta(Q) \quad (3)$$

$$K^B(P, Q) = \int_0^{2\pi} \frac{\partial}{\partial v} \left(\frac{e^{-ikR(P, Q)}}{R(P, Q)} \right) d\theta(Q)$$

in which ϕ is the total scalar velocity potential satisfying the Helmholtz differential equation $\nabla^2 \phi + k^2 \phi = 0$, k is the wave number ω/c of the time harmonic waves present in the region V exterior to the surface S of an arbitrary body, ω is the circular frequency and c is the speed of sound, while ϕ^I is the incoming wave.

It can be shown that evaluations of $K^A(P, Q)$, $K^B(P, Q)$ and $C(P)$ can be performed partly numerically and partly analytically in terms of elliptic integrals. Eq. (2) is the BEM formulation for axisymmetric bodies in a full space.

2.2. An Inverse Solution Formulation [3]

The numerical implementation of Eq. (2) includes cases where P is on the surface S and P outside of S. For P on the surface, Eq. (2) can be written in matrix form:

$$[A]\Phi = [B] \Phi' + \Phi_S^I \quad (4)$$

where [A] and [B] are NxN matrices, the element of vectors Φ , Φ' , and Φ_S^I are values of ϕ , ϕ' , and $-4\pi\phi^I$ on the surface. For P outside S (field point in V) the same procedure can be followed to obtain a matrix form of Eq. (2):

$$[C]\Phi + [D]\Phi' = \Phi_F + \Phi_F^I \quad (5)$$

where [C] and [D] are NFxN matrices, with NF is the total number of field points, the element of vectors Φ_F and Φ_F^I are the value of $4\pi\phi$ and $-4\pi\phi^I$ on each field point. Eq. (4) and (5) are used in solving inverse problem, wherein Φ_S^I , Φ_F and Φ_F^I are known while Φ and Φ' are to be solved.

In the direct BEM, either p (or ϕ) or ϕ' or the relation between them is known. Therefore by using Eq. (4), ϕ can be determined when ϕ' is known or vice versa. In the inverse BEM, no boundary condition is known, so there are more unknowns (2N) than the equations (N) in Eq. (3). To fully specify the problem, at least N more equations are needed, which are given by Eq. (4) with $NF \geq N$.

To calculate ϕ and ϕ' , rewrite Eq. (4) as

$$\Phi = [A]^{-1}[B] \Phi' + [A]^{-1}\Phi_S^I \quad (6)$$

Substituting Eq. (6) into Eq. (5) yields

$$([C][A]^{-1}[B] + [D]) \Phi' = \Phi_F + \Phi_F^I + [C][A]^{-1}\Phi_S^I \quad (7)$$

which can be rewritten as

$$[G] \Phi' = Y \quad (8)$$

where Y is the right-hand side vector. Since the field points pressure and the incoming wave are known, vector Y can be determined. Then, Φ' and Φ can be found by solving Eq. (8) and (6).

Golub[4] used SVD of matrix [A], which is defined as

$$[A] = [U][\Sigma][V]^T \quad (9)$$

where [U] and [V] are unitary matrices, satisfying the condition

$$[U][U]^T = [U]^T[U] = [I] \quad (10)$$

$$[V][V]^T = [V]^T[V] = [I] \quad (11)$$

with $[.]^T$ indicates a conjugate transpose matrix,

$$[\Sigma] = \text{diag}[\sigma_1 \dots \sigma_r, 0 \dots 0] \quad (12)$$

in which σ_r is called the singular value of [A], and r is the rank of [A]. If [A] is singular then one or several of σ_i are zero, so that $r < N$.

The pseudo inverse of [A], denoted as $[A]^+$, can be calculated using the relation $[A]^+ = [V][\Sigma]^+[U]^T$, with

$$[\Sigma]^+ = \begin{cases} \frac{1}{\sigma_i}, & \sigma_i \neq 0 \\ 0, & \sigma_i = 0 \end{cases} \quad (13)$$

where $[\Sigma]^+$ is the pseudo inverse of $[\Sigma]$.

2.3. Conjugate gradient [6]

From equation (8) can formulate inverse problem as

$$\Phi' = [G]^{-1}Y \quad (14)$$

The goal of inverse is to determine velocity (Φ') at surface source for radiation in full space, but inverse matrix $[G]^{-1}$ can't be obtained, so that equation (9) can't be used. There are two method to obtained inverse solution, the first is used SVD method or used Conjugate Gradient method so velocity value (Φ') can be obtained from some iteration, the velocity for each iteration j noted by $\Phi^{(j)}$. The error between velocity actually Φ' , and velocity from iteration $\Phi^{(j)}$ as :

$$r^{(j)} = Y - \hat{Y} = Y - [G] \Phi^{(j)} \quad (15)$$

If $r^{(j)}$ small enough, the process stopped, but if still big continue the next iteration. Initial value Conjugate Gradient methode for inverse probelm can be formulate:

$$\Phi^{(0)} = 0 \quad (16)$$

$$d^{(0)} = r^{(0)} = Y - [G] \Phi^{(0)} \quad (17)$$

$$\alpha^{(j)} = (r^{(j)T} r^{(j)}) / (d^{(j)T} G d^{(j)}) \quad (18)$$

Velocity renew with

$$\Phi^{(j+1)} = \Phi^{(j)} + \alpha^{(j)} d^{(j)} \quad (19)$$

when $\alpha^{(j)}$ is step size and $d^{(j)}$ is the new direction.

Then error and β calculate with

$$r^{(j+1)} = r^{(j)} - \alpha^{(j)} G d^{(j)} \quad (20)$$

$$\beta^{(j+1)} = (r^{(j+1)T} r^{(j+1)}) / (r^{(j)T} r^{(j)}) \quad (21)$$

The new direction calculate with:

$$d^{(j+1)} = r^{(j+1)} + \beta^{(j+1)} d^{(j)} \quad (22)$$

3. RESULTS

The axisymmetric source can be built from generator body as shown in Figure 2 for a radiating dome (radius $a = 0.08m$, 4 elements, 9 nodes). The source can be discretised into a number of elements, which are assumed to radiate as point sources, and the sound field was measured.

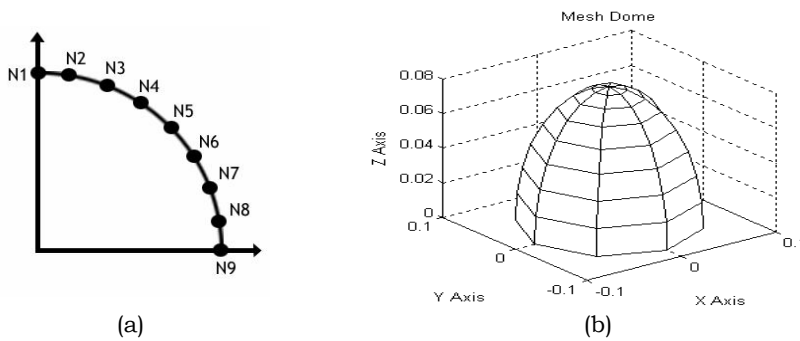


Figure 2. (a) Generator body (b) Axisymmetric source of a radiating dome

Figure 3 shows all of the field point distributions have the distance from the surface equal to 0.01m for radiating dome.

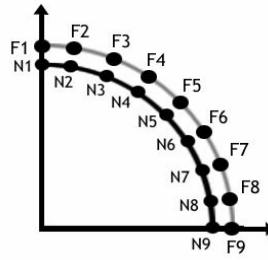


Figure 3. Fields point are distributed on the surface of imaginary space dome

The direct BEM calculation gives the surface pressure data. The prescribed surface data are referred to as the original data. Then the pressures at the selected field points are calculated using direct BEM. These pressures are then used as the given information to be employed in the present inverse formulation. Using this information as the input data, the present formulation calculates back the surface data. Figure 4 shows the real pressure for dome with inisialization from 0 to 1.4 . Figure 5 for imaginary pressure, and Figure 6 for magnitudo pressure.

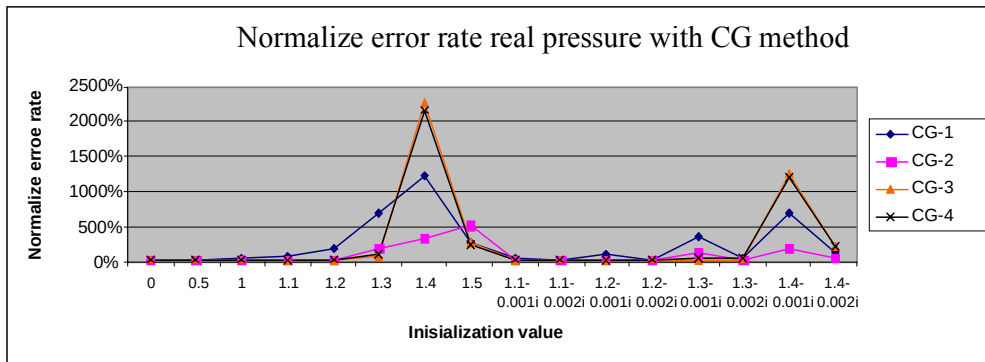


Figure 4. Comparison graph real pressure each iteration with Conjugate Gradient

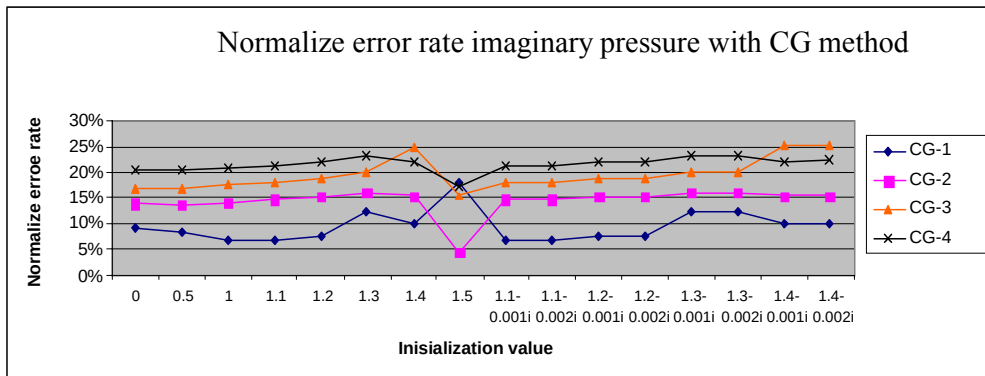


Figure 5. Comparison graph imaginary pressure each iteration with Conjugate Gradient

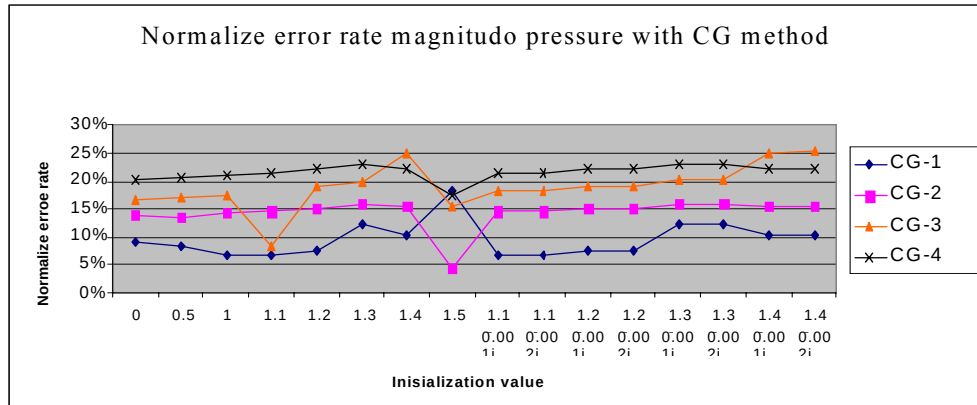


Figure 6. Comparison graph magnitudo pressure each iteration with Conjugate Gradient

Figure 7 shows the visualization of real pressure, imaginary pressure and magnitudo pressure for dome.

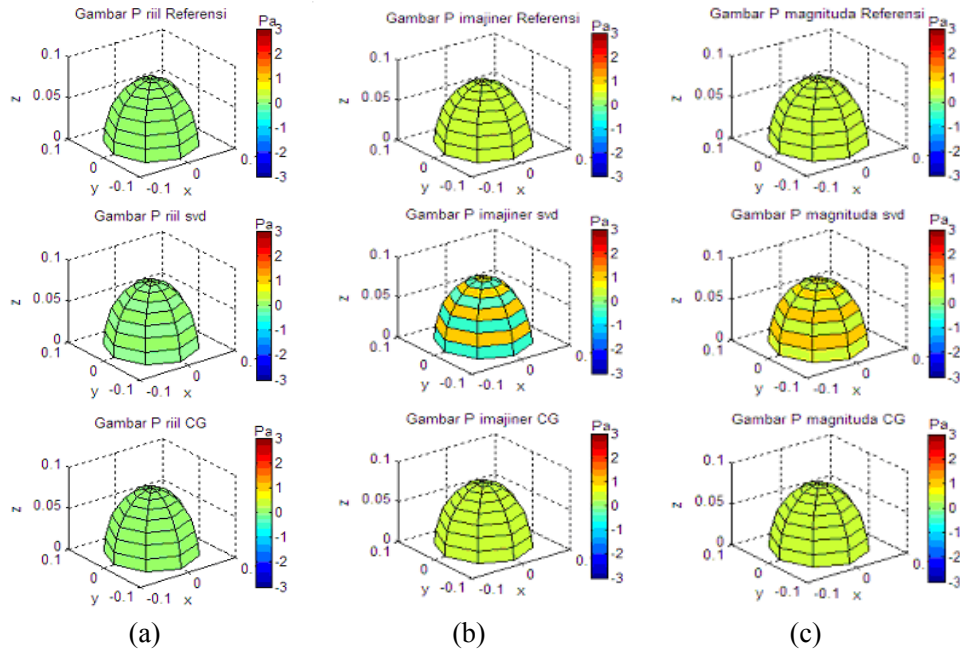


Figure 7 Visualization reference, solution with svd and solution with CG for (a)real (b)imaginary (c) magnitudo pressure

4. CONCLUSION

Inverse problems in acoustics may be solved with inverse BEM formulation using Singular Value Decomposition (SVD) and Conjugate Gradient. Test cases were conducted dome geometry. The result obtained shows a good agreement with the available analytical values and apriorily known quantities. The error solution is less than 17%. Using CG, less iteration produced optimum result. Initial condition affect the solution in the calculation with CG.

Acknowledgement

The authors would like to acknowledge the support of Prof. Benjamin Soenarko, Deddy Kurniadi, and Wono Setya Budhi, for their support.

REFERENCES

1. Benjamin Soenarko (1983), An Advanced Boundary Element Formulation for Acoustic Radiation and Scattering in Three Dimensions, *Ph.D Disertation at the University of Kentucky, America*.
2. B. Soenarko, Dwi Urika (2003), An inverse BEM formulation in acoustics for axisymmetric bodies and boundary conditions in a half space, *Proceedings of the 3rd International Conference on Numerical Analysis in Engineering*, **3**, 3.17 – 3.21
3. A.F. Seybert, B. Soenarko, F.J. Rizzo, and D.J. Shippy (1986), A Special integral equation formulation for acoustic radiation and scattering for axisymmetric bodies and boundary conditions, *J. Acoust. Soc. Am.*, **80 (4)**, 1241 – 1247
4. Gene H. Golub & Charles F. Van Loan (1996), Matrix Computations, *The John Hopkins University Press*.
5. Ratnadewi, and Soenarko, B. (2005) An Inverse Three Dimensional Acoustic Problem Solution for Axisymmetric Source in Full Space Using Boundary Element Method and Generalized Cross-Validation, *Proceedings of International Conference on Applied Mathematics 2005* (Bandung) August 22-26, pp CP9, 1-13
6. Ratnadewi, Soenarko, B., Kurniadi, D. Budhi, W.S. (2006) Optimasi Solusi Masalah Inversi Akusik untuk Sumber Simetri Sumbu dengan Conjugate Gradient dan Metode Elemen Batas, *Prosiding Konferensi Matematika XIII* (Semarang) July 24-27, pp 1-7

Ratna Dewi: Department of Electrical Engineering, Faculty of Engineering, Maranatha Christian University. Jl. Prof. Drg. Suria Sumantri no 65, Bandung 40164, Indonesia
E-mail : ratnadewi@engineer.com and rdewi@bdg.centrin.net.id

ON THE RELATION BETWEEN CONFIGURATION AND PHASE SPACE QUANTIZATION II

M. F. ROSYID

Abstract. The $-(\frac{1}{2} + i\gamma)$ -densities geometric quantization which is mathematically justified in [13] is applied to the cotangent bundles of smooth manifolds. By the use of a kind of gauge transformation, it is pointed out that such consideration leads to the quantum Borel-kinematics on the manifolds. The significance of gauge potentials in the quantum theory is rediscovered.

Key words and Phrases: geometric quantization, Borel quantization, gauge transformation

1. INTRODUCTION

In [7] the existence of the so-called Groenewold-van Hove-”no-go” theorem in quantizing the cylinder T^*S^1 is studied in the framework of a quantization method which is geometric in nature. There, a subalgebra $\Sigma(T^*S^1)$ of the algebra

$$\mathcal{F}(T^*S^1) = \{f(\theta)l + g(\theta) \in C^\infty(T^*S^1, \mathbb{R}) | (\theta, l) \in T^*S^1, f, g \in C^\infty(S^1, \mathbb{R})\}$$

of all quantizable real-valued functions on T^*S^1 is quantized through

$$\mathcal{Q}_{\nu, \eta}(f(\theta)l + g(\theta)) = -if(\theta)\frac{d}{d\theta} + \nu f(\theta) + (\eta + \frac{1}{2i})\frac{df}{d\theta} + g(\theta), \quad (1)$$

where η is a real number and $\nu \in [0, 1]$ represents a flat connection. The subalgebra $\Sigma(T^*S^1)$ is isomorphic to the *kinematical algebra* $\mathcal{S}(S^1) = C^\infty(S^1, \mathbb{R}) \oplus_s \mathfrak{X}_c(S^1)$ of S^1 . Two representations $\mathcal{Q}_{\nu, \eta'}$ and $\mathcal{Q}_{\nu, \eta}$ with two different real numbers η' and η are inequivalent. The real numbers η therefore can be understood as quantum numbers,

i.e. nontopological quantum numbers. Therefore the inequivalent quantizations of the algebra $\Sigma(T^*S^1)$ are characterized by $\pi_1^*(S^1) \times \mathbb{R}$. Moreover, Eq.(1) is equivalent to the expression of $\mathbb{Q}(f)$ and $\mathbb{P}(X)$, $\forall(f, X) \in \mathcal{S}(S^1)$, of quantum Borel-kinematics on S^1 , namely the kinematical part of Borel-quantization (BQ) of S^1 (a brief exposition of BQ will be given at the end of this Section, while the detail account can be found in [1, 2, 5, 12]).

The relation between Borel-quantization of smooth manifolds and the geometric quantizations (GQ) of their cotangent bundles is studied concisely in [12, 4, 8]. In [8] $-(\frac{1}{2} + i\gamma)$ -densities are suggested in the GQ of cotangent bundles. Still in the context of the relationship between BQ and GQ, the implementation of $-(\frac{1}{2} + i\gamma)$ -densities in the geometric quantization of arbitrary symplectic manifolds is mathematically justified in [13]. The present article is assumed to be the continuation of the exposition given in [13]. Here, we shall show that the $-(\frac{1}{2} + i\gamma)$ -densities geometric quantization on the cotangent bundles of smooth manifolds yields quantum Borel-kinematics on the manifolds. The main results of the present work are Theorem 2.1 and Theorem 2.2.

Geometric Quantization : Now we revisit shortly $-(\frac{1}{2} + i\gamma)$ -densities geometric quantization on symplectic manifolds. For the detail account the reader can consult [13] and some references therein. We begin with a general definition of quantization under which the geometric quantization is subordinated : Let $(C^\infty(M, \mathbb{R}), \{\cdot, \cdot\}_\omega)$ be a Poisson algebra of a symplectic manifold (M, ω) and $\mathcal{H}$ be a separable Hilbert space. A linear mapping $\mathcal{Q} : C^\infty(M, \mathbb{R}) \rightarrow \text{Sym}(\mathcal{H})$ from the Poisson algebra $C^\infty(M, \mathbb{R})$ into the set $\text{Sym}(\mathcal{H})$ of all symmetric operators in $\mathcal{H}$ is called *prequantization* if it satisfies both the following two axioms :

- (a) $\mathcal{Q}(c) = c\mathbb{I}$, for every $c \in \mathbb{R}$, where $\mathbb{I}$ denotes the identity operator in $\mathcal{H}$, and
- (b) $\mathcal{Q}(\{f, g\}) = i[\mathcal{Q}(f), \mathcal{Q}(g)]$ for every $f, g \in C^\infty(M, \mathbb{R})$, where $[\cdot, \cdot]$ is commutator in $\mathcal{H}$.

Whenever $\mathcal{Q}$ is also an irreducible representation of $C^\infty(M, \mathbb{R})$ in $\mathcal{H}$, we call it *quantization*. However, such irreducible map does not exist in general, even for the case of $\mathbb{R}^2$ with the canonical Poisson structure ([9, 6]).

The first step to geometric quantization is Kostant-Souriau-prequantization or, briefly, KS-prequantization. For a symplectic manifold (M, ω) , the prequantization consists of choosing a *prequantization line bundle*, i.e. complex line bundle $(L, \nabla, \langle \cdot, \cdot \rangle)$ with hermitian connection so that its curvature $\text{Cur}(L, \nabla)$ in L is equal to ω . However, not all symplectic manifold admits such line bundle. A symplectic manifold is called *prequantizable* if the manifold admits prequantization line bundle. In [11, 14] it is shown how to characterize the prequantizability of a symplectic manifold : There exists a complex line bundle $(L, \nabla, \langle \cdot, \cdot \rangle)$ with hermitian connection over (M, ω) so that $\text{Cur}(L, \nabla) = \omega$ if and only if ω is integral.

For a prequantizable manifold (M, ω) , the algebra $(C^\infty(M, \mathbb{R}), \{\cdot, \cdot\}_\omega)$ is then represented in the Hilbert space $\mathcal{H} = L^2(L, \omega^n, \langle \cdot, \cdot \rangle)$ of all square ω^n -integrable

smooth sections of L via

$$C^\infty(M, \mathbb{R}) \ni f \longmapsto \mathcal{Q}_{KS}(f) = -i \nabla_{X_f} + f \in \text{Sym}(\mathcal{H}).$$

Clearly, the map $\mathcal{Q}_{KS}$ is prequantization in the above sense. Nevertheless, it is in general not a quantization. With the help of Stone's theorem one can show that $\mathcal{Q}_{KS}(f)$ is self-adjoint whenever X_f is complete. To express $\mathcal{Q}_{KS}(f)$ locally, we need the following facts [11, 14]

Proposition 1.1 *Let (L, ∇) be a complex line bundle with connection over a symplectic manifold (M, ω) and $\{(U_\alpha, s_\alpha)\}$, where α 's are in an index set, is a local system of L with transition functions $\{c_{\alpha\beta}\}$. Then on every $U_\alpha \subset M$ there exists a smooth complex 1-form θ_α so that $\nabla_X s_\alpha = -i(X \lrcorner \theta_\alpha) s_\alpha$ for every vector field X on U_α . Furthermore we have the gauge transformation*

$$\theta_\beta = \theta_\alpha - \frac{dc_{\alpha\beta}}{ic_{\alpha\beta}}. \quad (2)$$

Conversely, if a set of 1-forms $\{\theta'_\alpha\}$ satisfies the gauge transformation Eq.(2) then there exists a unique connection ∇' so that $\nabla'_X s_\alpha = -i(X \lrcorner \theta'_\alpha) s_\alpha$ for every vector field X on U_α .

Proposition 1.2 *Let L^* be the frame bundle associated to L . Then there is a unique connection form $\theta \in \Omega^1(L^*)$ on L^* so that*

$$\theta^s = s^*(\theta) \quad (3)$$

for every open subset $U \subset M$ and every $s \in \text{Sec}^\infty(L^*|U)$, where θ^s is defined by $\nabla s = -i\theta^s s$. Conversely, if $\theta' \in \Omega^1(L^*)$ a connection form on L^* , then there exists a unique connection ∇' in L so that $\nabla' s = -i\theta'^s s$ for every $s \in \text{Sec}^\infty(L^*|U)$ and every open subset U of M , where $\theta'^s = s^*(\theta')$.

Proposition 1.3 *The connection ∇ in L is hermitean whenever the real 1-form $i(\theta - \bar{\theta})$ is exact.*

Remark 1.1 *For a prequantization line bundle $(L, \nabla, \langle \cdot, \cdot \rangle)$ every local 1-form θ_α on U_α which satisfies $\nabla_X s_\alpha = -i(X \lrcorner \theta_\alpha) s_\alpha$, for every vector field X on U_α , is a local symplectic potential, $d\theta_\alpha = \omega \upharpoonright U_\alpha$.*

Hence, for an observable $f \in C^\infty(M, \mathbb{R})$ we have locally

$$\mathcal{Q}_{KS}(f)s \upharpoonright U_\alpha = -iX_f(\psi)s_\alpha - (X_f \lrcorner \theta_\alpha)\psi s_\alpha + f\psi s_\alpha$$

for every $s \in \mathcal{H}$, where ψ is a complex function on U_α so that $s \upharpoonright U_\alpha = \psi s_\alpha$.

In the case of $Q = \mathbb{R}^n$, one already encounters some essential problems in obtaining the Schrödinger representation. For instance, the representation Hilbert space is to 'large'. It contains wave functions which depend on both position and momentum. Then, one has to reduce the size of the representation Hilbert space $\mathcal{H}$. The reduction of $\mathcal{H}$ is controled from 'below' by *polarizing* the classical phase space (M, ω) , namely a foliation of M whose leaves are *Lagrangian submanifolds* with respect to ω . In the present article we will restrict our consideration in the case of real polarization P . Let M/P denote the space of all leaves of P and π_P

the canonical projection $\pi_P : M \longrightarrow M/P$. A polarization P is called *reducible* whenever M/P is a smooth manifolds. Then, the smooth manifold M/P is called the *generalized configuration space extracted from* (M, ω) by P . But there are symplectic manifolds which do not admit a real polarization.

With the polarization at hand one constructs the so-called $-(\frac{1}{2} + i\gamma)$ - P -densities line bundle $|\wedge^n|^{-(\frac{1}{2} + i\gamma)} B^P(M)$ with a partial flat connection $\check{\nabla}$ [13]. Furthermore, the quantum bundle $\mathcal{B}_{i\gamma, P}$ is defined as the tensor product

$$(L \otimes |\wedge^n|^{-(\frac{1}{2} + i\gamma)} B^P(M), \nabla \otimes \check{\nabla})$$

where (L, ∇) is the prequantization line bundle. Then the Hilbert space $\mathcal{H}_{i\gamma, P}$ of representation is constructed from the polarized sections of $\mathcal{B}_{i\gamma, P}$. i.e. $\mathcal{H}_{i\gamma, P}$ is the completion of

$$\mathcal{W}_{i\gamma, P} := \{\psi \in Sec^\infty(\mathcal{B}_{i\gamma, P}) : (\nabla_P \otimes \check{\nabla}_P)\psi = 0 \text{ and } \int_{M/P} (\psi, \psi)_{i\gamma, P} < \infty\}$$

endowed with scalar product $\langle \psi, \psi' \rangle_{i\gamma, P} := \int_{M/P} (\psi, \psi')_{i\gamma, P}$, for all $\psi, \psi' \in \mathcal{H}_{i\gamma, P}$, where $(\cdot, \cdot)_{i\gamma, P}$ is a 1-density-valued pairing in the set of all polarized sections of $\mathcal{B}_{i\gamma, P}$ defined by

$$(\varsigma, \varsigma')_{i\gamma, P}(m) := \langle s, s' \rangle(m) \nu(m) \bar{\nu}'(m) |\varepsilon_\omega|, \quad (4)$$

for any two sections $\varsigma = s \otimes \nu$ and $\varsigma' = s' \otimes \nu'$ of $\mathcal{B}_{i\gamma, P}$, where ε_ω is Liouville measure corresponding to ω . An observable $f \in C^\infty(M, \mathbb{R})$ is called *P-quantizable* whenever $[X_f, Y] \in \mathfrak{X}_P(M)$ for every $Y \in \mathfrak{X}_P(M)$. The observable f is then quantized via

$$\mathcal{Q}_{i\gamma, P}(f)\varsigma = -i(\nabla_{X_f} s) \otimes \nu + f(s \otimes \nu) - is \otimes \mathfrak{L}_{X_f} \nu, \quad \forall \varsigma = s \otimes \nu \in \mathcal{H}_{i\gamma, P}$$

where $\mathfrak{L}$ is Lie derivative in $Sec^\infty(|\wedge^n|^{-(\frac{1}{2} + i\gamma)} B^P(M))$. If $g \in C^\infty(M, \mathbb{R})$ is quantizable, then $\mathcal{Q}_{i\gamma, P}(g)$ satisfies $\mathcal{Q}_{i\gamma, P}(g)\mathcal{W}_{i\gamma, P} \subseteq \mathcal{W}_{i\gamma, P}$ [13].

Let $\mathcal{F}_P(M, \mathbb{R})$ denote the set of all P -quantizable observables. In general $\mathcal{F}_P(M, \mathbb{R})$ is a proper subset of $C^\infty(M, \mathbb{R})$. Even in the case of the cotangent bundle T^*Q of a smooth orientable n -dimensional manifold Q with the vertical polarization P_v the quantizable subalgebra $\mathcal{F}_{P_v}(T^*Q, \mathbb{R})$ contains of functions of the form $\zeta \circ pr + \sum_i^n (\zeta^i \circ pr) p_i$, where ζ, ζ^i ($i = 1, \dots, n$) are functions in $C^\infty(Q, \mathbb{R})$ and $(p_1, \dots, p_n, q^1, \dots, q^n)$ is a (local) canonical coordinat system on T^*Q .

Borel Quantization : In contrast to GQ, Borel quantization is modelled on configuration spaces, i.e. arbitrary smooth manifolds Q . Then in the dynamical part it takes cotangent bundles associated to such manifolds as phase spaces. The kinematical part of BQ is called quantum Borel-kinematics (QBK). The first concise description of QBK is given in [1] and is discussed intensively in [2, 5].

Let Q be a smooth manifold and $\mathfrak{X}_c(Q)$ the set of all smooth complete vector fields on Q . The kinematical algebra of Q is the set $C^\infty(Q, \mathbb{R}) \times \mathfrak{X}_c(Q)$ equipped with bracket $[(f, X), (g, Y)]_s := (X(g) - Y(f), [X, Y])$. The kinematical algebra

of Q is therefore the semidirect sum $C^\infty(Q, \mathbb{R}) \oplus_s \mathfrak{X}_c(Q)$ of the abelian algebra $C^\infty(Q, \mathbb{R})$ as the ideal and the partial Lie-algebra $\mathfrak{X}_c(Q)$. In the sequel the kinematical algebra of Q will be denoted by $\mathcal{S}(Q)$

Let Φ be a closed integral real 2-form on Q and $(\eta, \langle \cdot, \cdot \rangle)$ a hermitean line bundle over Q . Further, ϑ^∞ denote the set of all compact supported smooth sections of η and $\mathfrak{H}$ the Hilbert space $L^2(\eta, \langle \cdot, \cdot \rangle, \nu)$ with respect to smooth Borel measure ν on Q . An elementary differentiable Φ -compactible quantum Borel kinematic is a quadruple $(\mathfrak{H}, \mathbb{Q}, \mathbb{P}, \theta^\infty)$ where

1. $\mathbb{Q} : C^\infty(Q, \mathbb{R}) \rightarrow SA(\mathfrak{H})$ a map defined by $\mathbb{Q}(\zeta)\sigma = \zeta\sigma$ for every $\zeta \in C^\infty(Q, \mathbb{R})$ and every $\sigma \in \text{Dom}(\mathbb{Q}(\zeta)) = \{\sigma' \in \mathfrak{H} \mid \int_Q |\zeta(x)|^2 \|\sigma'\|^2 d\nu(x) < \infty\}$.
2. $\mathbb{P} : \mathfrak{X}_c(Q) \rightarrow SA(\mathfrak{H})$ a map from $\mathfrak{X}_c(Q)$ into $SA(\mathfrak{H})$ defined by $\mathbb{P}(X)\sigma = -i\nabla_X \sigma + (\frac{1}{2i} + \gamma)[\text{div}_\nu(X)]\sigma$, $\forall \sigma \in \vartheta^\infty$, $\forall X \in \mathfrak{X}_c(Q)$, where ∇ is a hermitean connection with curvature $i\Phi$ und $\gamma \in \mathbb{R}$ is a real number.

It is easy to show that the kinematical algebra $\mathcal{S}(Q)$ is partially Lie-isomorphic to the subalgebra

$$\Sigma(T^*Q) = \{\zeta \circ \pi + (\zeta^i \circ \pi)p_i : \zeta, \zeta^i \in C^\infty(Q, \mathbb{R}) \text{ with } \zeta^i \frac{\partial}{\partial q^i} \in \mathfrak{X}_c(Q)\}$$

of $C^\infty(T^*Q, \mathbb{R})$, where $(p_1, \dots, p_n, q^1, \dots, q^n)$ is a (local) canonical coordinat system on T^*Q . The isomorphism is given by

$$\zeta \longmapsto q_\zeta = \zeta \circ \pi, \quad X \longmapsto p_X = \theta_c(\bar{X}) \quad (5)$$

for every $(\zeta, X) \in \mathcal{S}(Q)$, where θ_c is the canonical symplectic potential and $\bar{X}$ is the natural lift of X to T^*Q .

2. MAIN RESULTS

We shall show how a QBK on a Riemannian manifold emerges naturally from the $-(\frac{1}{2} + i\gamma)$ -densities geometric quantization of cotangent bundles of the manifold. Our approach is in some aspects different from that of [8].

Let (Q, g) be an n -dimensional orientable Riemannian manifold. Then the symplectic manifold (T^*Q, ω) is prequantizable as the canonical symplectic form ω is globally exact. Now choose the vertical polarization D of T^*Q . The generalized configuration space T^*Q/D is clearly Q . From the relation $T_q Q = T_m M/D(m)$, $\forall q \in Q$ mit $q = \pi_D(m)$, one have

$$|\wedge^n|^{-(1/2+i\gamma)} B^D(m) = |\wedge^{2n}|^{-(1/2+i\gamma)} B_m M \otimes |\wedge^n|^{(1/2+i\gamma)} B_q Q, \quad (6)$$

where $|\wedge^n|^{\pm(1/2+i\gamma)} B^D$, $|\wedge^n|^{\pm(1/2+i\gamma)} B M$, and $|\wedge^n|^{\pm(1/2+i\gamma)} B Q$ is the $\pm(1/2+i\gamma)$ -densities bundle associated to D , TM and TQ respectively [13]. Eq.(6) leads to the natural section

$$\nu = |\varepsilon_\omega|^{-(1/2+i\gamma)} |\varepsilon_g|^{(1/2+i\gamma)} \quad (7)$$

of the $-(1/2 + i\gamma)$ -densities bundle $|\wedge^n|^{-(1/2+i\gamma)}B^D$, where ε_g is the canonical measure on Q , i.e. $\varepsilon_g = (\det(g))^{1/2}dq^1 \wedge \cdots \wedge dq^n$ for local coordinates $(q^1 \cdots q^n)$ of Q . It is clear that the density field ν is covariantly constant along D relative to the connection ∇ . Since the canonical symplectic potential θ_c which satisfies $d\theta_c = \omega$ is globally defined, one can choose the trivial bundle $T^*Q \times \mathbb{R} \rightarrow T^*Q$ as the prequantization line bundle L . This means that the space $Sec^\infty(L)$ of all smooth sections of L is isomorphic with $C^\infty(T^*Q, \mathbb{R})$. It implies that every section s of L which are covariantly constant along D relative to the connection ∇ can be identified with one and only one smooth complex-valued function defined on Q . Hence, every wave function $\psi = s \otimes \nu \in \mathcal{H}_{i\gamma, D}$ can be written as $\psi(m) = (\phi \circ \pi_D)(m)(\mathbf{1} \otimes \nu(m))$, for every $m \in T^*Q$, where $\mathbf{1}$ is the unit section of L . The inner product in $\mathcal{H}_{i\gamma, D}$ is given by $\langle \psi, \psi' \rangle = \int_Q \phi \bar{\phi}' |\varepsilon_g|$, where ϕ' is complex-valued function on Q corresponding to $\psi' \in \mathcal{H}_{i\gamma, D}$.

For a quantizable function $h \in C^\infty(T^*Q, \mathbb{R})$ let $Y_h(\pi_D(m)) := (\pi_D)_* X_h(m)$, for every $m \in T^*Q$. It is clear that $q \mapsto Y_h(q)$ is a smooth vector field on Q . The $-(\frac{1}{2} + i\gamma)$ - D -densities quantization of h is given by

$$\begin{aligned} \mathcal{Q}(h)(\phi \circ \pi_D)(\mathbf{1} \otimes \nu) &= \{-iY_h(\phi) \circ \pi_D + h \cdot (\phi \circ \pi_D) - (\theta_c(Y_h) \cdot \phi) \circ \pi_D \\ &\quad + (\frac{1}{2i} + \gamma)[div_{\varepsilon_g}(Y_h) \circ \pi_D](\phi \circ \pi_D)\} \mathbf{1} \otimes \nu. \end{aligned}$$

Let $(p_1, \dots, p_n, q^1, \dots, q^n)$ be a canonical coordinates system of T^*Q . Every quantizable observable f can be written as

$$f(p_1, \dots, p_n, q^1, \dots, q^n) = h(q^1, \dots, q^n) + \sum_{j=1}^n k^j(q^1, \dots, q^n) p_j,$$

where $h, k^j \in C^\infty(Q, \mathbb{R})$. Therefore, the quantizable algebra is the direct sum $\mathcal{F}(T^*Q) := \mathcal{F}_0(T^*Q) \oplus_s \mathcal{F}_1(T^*Q)$, where $\mathcal{F}_r(T^*Q)$ denotes the space of all smooth functions on T^*Q which is r -homogeneous in p 's. According to the previous section, the image $\Sigma(T^*Q)$ of $\mathcal{S}(Q)$ under the isomorphism defined by Eq.(5) is contained in $\mathcal{F}(T^*Q)$. Thus, every elements of $\Sigma(T^*Q)$ is quantizable and the $-(\frac{1}{2} + i\gamma)$ - D -densities quantization of $\Sigma(T^*Q)$ is locally given by

$$\mathcal{Q}_{i\gamma}(q_f)\psi = f\psi$$

und

$$\mathcal{Q}_{i\gamma}(p_X)\psi = \{-iX(\varphi) \circ \pi_D + (\frac{1}{2i} + \gamma)[div_{\varepsilon_g}(X) \circ \pi_D](\varphi \circ \pi_D)\} \mathbf{1} \otimes \nu,$$

for all $(f, X) \in \mathcal{S}(Q)$ and all $\psi = s \otimes \nu = (\varphi \circ \pi_D)\mathbf{1} \otimes \nu \in \mathcal{H}_{i\gamma, D}$.

Let M_1 and M_2 be two smooth manifolds and $\pi_2 : L_2 \rightarrow M_2$ a complex line bundle over M_2 . Furthermore, let $\rho : M_1 \rightarrow M_2$ be a smooth mapping from M_1 into M_2 . Then $\rho^*L_2 \rightarrow M_1$, where

$$\rho^*L_2 = \bigcup_{m_1 \in M_1} (L_2)_{\rho(m_1)} \times \{m_1\},$$

is a complex line bundle over M_1 . The mapping $\tau_\rho : \rho^*L_2 \rightarrow L_2$, defined by $\tau_\rho(z, m_1) = z$ for every $(z, m_1) \in \rho^*L_2$ is a mapping of bundles. Now we appreciate the following facts. Let θ_2 be a connection form in L_2 . If L_1 is a complex line bundle over M_1 and $\tau : L_1 \rightarrow L_2$ a lift of $\rho : M_1 \rightarrow M_2$, then $\tau^*\theta_2$ a connection in L_1 and τ is a mapping of line bundles with connections. Especially, $(\rho^*L_2, \tau_\rho^*\theta_2)$ is a complex line bundle with connection over M_1 and τ_ρ is a mapping of line bundles with connections. The bundle ρ^*L_2 is called the *pull back* of L_2 and the connection $\tau_\rho^*\theta_2$ the *pull back* of θ_2 .

Lemma 2.1 *Let $\{(V_\alpha, s_\alpha)\}$ be a local system of L_2 with transition functions $c_{\alpha\beta}$ and π the bundle projection $\rho^*L_2 \rightarrow M_1$. The mapping ρ induces then a local system $\{(\rho^{-1}(V_\alpha), \mathfrak{s}_\alpha)\}$ of ρ^*L_2 with transition functions $\mathfrak{c}_{\alpha\beta}$, where $\mathfrak{s}_\alpha$ and $\mathfrak{c}_{\alpha\beta}$ are defined by*

$$\mathfrak{s}_\alpha : \rho^{-1}(V_\alpha) \ni m \mapsto (s_\alpha \circ \rho(m), m) \in \pi^{-1}(\rho^{-1}(V_\alpha))$$

and

$$\mathfrak{s}_\beta(m') = \mathfrak{c}_{\alpha\beta}(m')\mathfrak{s}_\alpha(m') = ((c_{\alpha\beta} \cdot s_\alpha) \circ \rho(m'), m'), \quad \forall m' \in \rho^{-1}(V_\alpha) \cap \rho^{-1}(V_\beta),$$

for non empty intersections $\rho^{-1}(V_\alpha) \cap \rho^{-1}(V_\beta)$ respectively. Let $\{(\theta_2)_\alpha\}$ be the set of 1-forms defined by $(\theta_2)_\alpha = s_\alpha^*\theta_2$. Then, the connection ∇^1 associated with the form $\tau_\rho^*\theta_2$ is given by $\nabla^1\mathfrak{s}_\alpha = -i\rho^*(\theta_2)_\alpha\mathfrak{s}_\alpha$.

Proof: That the functions $\mathfrak{c}_{\alpha\beta}$ and the sections $\mathfrak{s}_\beta$ are smooth is already clear from the smoothness of $c_{\alpha\beta}$, s_α and of ρ . Now for the first statement it is sufficient to show that the functions $\mathfrak{c}_{\alpha\beta}$ satisfy $\mathfrak{c}_{\alpha\beta} = (\mathfrak{c}_{\beta\alpha})^{-1}$ and $\mathfrak{c}_{\alpha\beta}\mathfrak{c}_{\beta\gamma}\mathfrak{c}_{\gamma\alpha} = 1$ if $\rho^{-1}(V_\alpha) \cap \rho^{-1}(V_\beta) \cap \rho^{-1}(V_\gamma)$ is not empty. But it follows easily from their definition. For the second statement, define $(\tau_\rho^*\theta_2)_\alpha$ for every α in the index set by $(\tau_\rho^*\theta_2)_\alpha = \mathfrak{s}_\alpha^*(\tau_\rho^*\theta_2)$. Then, by making use of Proposition 1.2 we have

$$\nabla^1\mathfrak{s}_\alpha = -i(\tau_\rho^*\theta_2)_\alpha\mathfrak{s}_\alpha = -i\mathfrak{s}_\alpha^*(\tau_\rho^*\theta_2)\mathfrak{s}_\alpha = -i(s_\alpha \circ \rho)^*\theta_2\mathfrak{s}_\alpha = -i\rho^*(\theta_2)_\alpha\mathfrak{s}_\alpha. \quad \square$$

Remark 2.1 *Because $\tau_\rho^* \circ d = d \circ \tau_\rho^*$ the connection ∇^1 is hermitean if and only if ∇^2 is hermitean. From Proposition 1.1 we have $\rho^*(\theta_2)_\beta = \rho^*(\theta_2)_\alpha - (dc_{\alpha\beta}/ic_{\alpha\beta})$.*

2.1 When external fields do not present

Lemma 2.2 *Let L' be a complex line bundle over Q , $\{(U_\alpha, s_\alpha); \alpha \in \mathbb{I}\}$ a local system of L' with transition functions $c_{\alpha\beta}$, where $\mathbb{I}$ is an index set, and $\{\Phi_\alpha\}$ a set of closed 1-forms $\Phi_\alpha \in \Omega^1(U_\alpha)$, $\forall \alpha \in \mathbb{I}$, which satisfy $\Phi_\beta = \Phi_\alpha - (dc_{\alpha\beta}/ic_{\alpha\beta})$.*

Furthermore, let Φ be the connection form on L'^* associated with the set $\{\Phi_\alpha\}$ (see Proposition 1.2), where L'^* is the frame bundle corresponding to L' . If the equivalent class $[i(\Phi - \bar{\Phi})] \in H^1(L'^*, \mathbb{R})$ vanishes, then there exists a hermitean connection ∇^Θ of curvature ω in $\mathcal{L} := L \otimes \pi_D^*L'$ so that

$$\nabla^\Theta(\mathbf{1} \otimes \mathfrak{s}_\alpha) = -i\theta_c \mathbf{1} \otimes \mathfrak{s}_\alpha - i\mathbf{1} \otimes (\pi_D^*\Phi_\alpha)\mathfrak{s}_\alpha, \quad \forall \alpha \in \mathbb{I}.$$

Proof: According to Proposition 1.1 there exists a connection $\check{\nabla}$ in L' so that $\check{\nabla}s_\alpha = -i\Phi_\alpha s_\alpha$, $\forall \alpha \in \mathbb{I}$. Then by Proposition 1.2 there is a unique connection

form Φ on L'^* which satisfies $s_\alpha^* \Phi = \Phi_\alpha$ for all $\alpha \in \mathbb{I}$. If the real 1-form $i(\Phi - \bar{\Phi})$ is exact or equivalently the class $[i(\Phi - \bar{\Phi})]$ in $H^1(L'^*, \mathbb{R})$ vanishes, then the connection $\check{\nabla}$ is hermitean by Proposition 1.3.

The pullback bundle $\pi_D^* L'$ over T^*Q of the line bundle L' relative to π_D is a bundle with total space given by $\{(m, l) | m \in T^*Q, l \in L'_{\pi_D(m)}\}$. From Lemma 2.1, the local system $\{(U_\alpha, s_\alpha); \alpha \in \mathbb{I}\}$ of the line bundle L' with transition functions $c_{\alpha\beta}$ induces the local system $\{(\pi_D^{-1}(U_\alpha), \mathfrak{s}_\alpha); \alpha \in \mathbb{I}\}$ with transition functions $\mathfrak{c}_{\alpha\beta}$. Therefore, there is a hermitean connection $\check{\nabla}$ in $\pi_D^* L'$ so that $\check{\nabla} \mathfrak{s}_\alpha = -i(\pi_D^* \Phi_\alpha) \mathfrak{s}_\alpha$, $\alpha \in \mathbb{I}$.

Since the bundle (L, ∇) is trivial, the forms $\Theta_\alpha := \theta_c + \pi_D^* \Phi_\alpha$ according to Remark 2.1 satisfy $\Theta_\beta = \Theta_\alpha - (d\mathfrak{c}_{\alpha\beta}/i\mathfrak{c}_{\alpha\beta})$, for all $\alpha, \beta \in \mathbb{I}$. Thus, from Proposition 1.1 there exists a connection ∇^Θ in product bundle $\mathcal{L} = L \otimes \pi_D^* L'$, such that

$$\nabla^\Theta(\mathbf{1} \otimes \mathfrak{s}_\alpha) = -i\theta_c \mathbf{1} \otimes \mathfrak{s}_\alpha - i\mathbf{1} \otimes (\pi_D^* \Phi_\alpha) \mathfrak{s}_\alpha$$

$\forall \alpha \in \mathbb{I}$, where $\{(\pi_D^{-1}(U_\alpha), \mathbf{1} \otimes \mathfrak{s}_\alpha); \alpha \in \mathbb{I}\}$ is the lokal system of $\mathcal{L}$ with transition functions $\mathfrak{c}_{\alpha\beta}$. By direct calculation it is easy to show that ∇^Θ is hermitean. $\square$

Let $\sigma = s \otimes \varsigma$ a smooth section of $\mathcal{L}$. Locally, σ can be written as $\sigma|_{\pi_D^{-1}U_\alpha} = s \otimes \varsigma|_{\pi_D^{-1}U_\alpha} = \varphi(\mathbf{1} \otimes \mathfrak{s}_\alpha)$, where $\varphi \in C^\infty(\pi_D^{-1}U_\alpha, \mathbb{C})$ is the *coordinates* of σ on $\pi_D^{-1}U_\alpha$ relative to the *frame* $\mathbf{1} \otimes \mathfrak{s}_\alpha$. Then we have

$$\nabla_X^\Theta \sigma|_{\pi_D^{-1}U_\alpha} = \nabla_X^\Theta [\varphi(\mathbf{1} \otimes \mathfrak{s}_\alpha)] = X(\varphi)(\mathbf{1} \otimes \mathfrak{s}_\alpha) - i(\theta_c + \pi_D^* \Phi_\alpha)(X)\varphi(\mathbf{1} \otimes \mathfrak{s}_\alpha),$$

for every $X \in \text{Sec}^\infty(TT^*Q)$.

It should be noted that $(\mathcal{L}, \nabla^\Theta)$ is a complex line bundle with hermitean connection of curvature ω . Thus, $(\mathcal{L}, \nabla^\Theta)$ is also a prequantization line bundle over (T^*Q, ω) . The prequantization map is given by

$$C^\infty(T^*Q, \mathbb{R}) \ni f \longmapsto \mathcal{Q}_{KS}(f) = -i \nabla_{X_f}^\Theta + f \in \text{Sym}(\mathcal{H}).$$

Putting the prequantization bundle $(\mathcal{L}, \nabla^\Theta)$ in the place of (L, ∇) one gets

$$\mathcal{B}_{i\gamma, D}^\Theta := \mathcal{L} \otimes |\wedge^n|^{-(1/2+i\gamma)} B^D(T^*Q)$$

as the new quantum bundle with connection $\nabla^\Theta \otimes \check{\nabla}$. Then the Hilbert space $\mathcal{H}_{i\gamma, D}^\Theta$ of quantization is the completion of the set

$$\mathcal{W}_{i\gamma, D}^\Theta := \{\psi \in \text{Sec}^\infty(\mathcal{B}_{i\gamma, D}^\Theta) | \nabla_X^\Theta \otimes \check{\nabla}_X \psi = 0, \forall X \in D \text{ and } \int_Q (\psi, \psi)_{i\gamma} < \infty\}. \quad (8)$$

From Eq.(5) we have

$$\Theta_\alpha(X_{q_f}) = 0 \quad (9)$$

and

$$\Theta_\alpha(X_{p_X}) = \theta_c(X_{p_X}) + \pi_D^* \Phi_\alpha(X_{p_X}) = p_X + \Phi_\alpha(X) \circ \pi_D, \quad (10)$$

for all $(f, X) \in \mathcal{S}(Q)$. Therefore, the $-(\frac{1}{2} + i\gamma)$ - D -densities quantization of the algebra $\Sigma(T^*Q) \cong \mathcal{S}(Q)$ is given locally by

$$\mathcal{Q}_{i\gamma}(q_f)\varphi \circ \pi_D = f\varphi \circ \pi_D \quad (11)$$

and

$$\mathcal{Q}_{i\gamma}(p_X)\psi = \{[-iX(\varphi) - \Phi_\alpha(X)\varphi + (\frac{1}{2i} + \gamma)(\text{div}_{\varepsilon_g}(X))\varphi] \circ \pi_D\}(\mathbf{1} \otimes \mathfrak{s}_\alpha \otimes \nu), \quad (12)$$

for all $(f, X) \in \mathcal{S}(Q)$, $\psi = \varphi(\mathbf{1} \otimes \mathfrak{s}_\alpha) \otimes \nu \in \mathcal{H}_{i\gamma, D}^\Theta$ and $\forall \alpha \in \mathbb{I}$.

Now $\omega|\pi_D^{-1}(q)$ vanishes for every $q \in Q$. Since $\pi_D^{-1}(q)$ is simply connected, then parallel displacements on every leave $\pi_D^{-1}(q)$ of D do not depend on the choosen contour. Then we define an equivalent relation $\sim$ as follows. Let $l_1, l_2 \in \mathcal{L}$ and $\pi : \mathcal{L} \rightarrow T^*Q$ the projection of the prequantization bundle. l_1 is equivalent to l_2 (written as $l_1 \sim l_2$) if and only if (a) $\pi(l_1)$ and $\pi(l_2)$ lie in a same leave of D , i.e. $\pi_D(\pi(l_1)) = \pi_D(\pi(l_2))$, and (b) l_2 can be reached from l_1 by ∇^Θ -parallel displacement within the leave $\pi^{-1}(\pi_D^{-1}(q))$.

The following proposition adapted from [3] is of importance in the sequel.

Proposition 2.1 *The equivalent relation $\sim$ is differentiable and conserves the hermitean structure of the bundle $\mathcal{L}$. The space $\mathcal{L}/\sim$ of all equivalent classes with respect to $\sim$ is hermitean complex line bundle over Q . If η denotes the space $\mathcal{L}/\sim$, then the diagram*

$$\begin{array}{ccc} \mathcal{L} & \xrightarrow{pr} & \eta \\ \pi \downarrow & & \downarrow \pi_\eta \\ T^*Q & \xrightarrow{\pi_D} & Q \end{array} \quad (13)$$

is commutative, where pr denotes the natural projection from $\mathcal{L}$ onto η .

Straightforward calculation shows that $\eta = \eta' \otimes L'$, where $\eta' = L/\sim$ is the quotient bundle of L with respect to the similar equivalent relation induced by the connection ∇ . Since the bundle η' is trivial, the set $\{(U_\alpha, \mathbf{1}_{\eta'} \otimes s_\alpha)\}$ forms a local system with transition functions $c_{\alpha\beta}$. Then there is a hermitean connection ∇^Φ in η so that $\nabla^\Phi(\mathbf{1}_{\eta'} \otimes s_\alpha) = -i\Phi_\alpha(\mathbf{1}_{\eta'} \otimes s_\alpha)$, for every $\alpha \in \mathbb{I}$. The connection ∇^Φ is clearly flat.

Let $\text{Sec}_D^\infty(\mathcal{L})$ denote the set of all smooth sections of $\mathcal{L}$ that are covariantly constant along D , and $\text{Sec}^\infty(\eta)$ the set of all smooth sections of η . Every $\sigma \in \text{Sec}_D^\infty(\mathcal{L})$ determines then a mapping $s_\sigma : Q \rightarrow \eta$ so that the diagram

$$\begin{array}{ccc} T^*Q & \xrightarrow{\sigma} & \mathcal{L} \\ \pi_D \downarrow & & \downarrow pr \\ Q & \xrightarrow{s_\sigma} & \eta \end{array}$$

is commutative.

Lemma 2.3 *The assignment $\sigma \mapsto s_\sigma$ defines an isomorphism from $Sec_D^\infty(\mathcal{L})$ onto $Sec^\infty(\eta)$.*

Proof : The mapping s_σ is obviously differentiable. From the diagram Eq.(13)

$$\pi_\eta \circ s_\sigma(q) = \pi_\eta \circ pr \circ \sigma(m) = q,$$

holds for each $q \in Q$, where m are points in T^*Q so that $\pi_D(m) = q$. The mapping $s_\sigma : Q \rightarrow \eta$ is therefore a smooth section of η .

Let $\sigma', \sigma'' \in Sec_D^\infty(\mathcal{L})$ so that σ' and σ'' determine the same section $s' \in Sec^\infty(\eta)$. Equation () leads then to $pr \circ \sigma' = pr \circ \sigma''$. Since $\omega|\pi_D^{-1}(q)$ vanishes for all $q \in Q$, we have $\sigma' = \sigma''$. This shows that the mapping $\sigma \mapsto s_\sigma$ is an injection.

Now let $s \in Sec^\infty(\eta)$. For every $m \in T^*Q$ define $l_m \in \mathcal{L}$ by $\pi(l_m) = m$ and $l_m \in s(\pi_D(m))$. Since the equivalent relation $\sim$ is differentiable, the mapping $T^*Q \ni m \mapsto \varsigma(m) = l_m \in \mathcal{L}$ is also differentiable. The mapping ς is a differentiable section of $\mathcal{L}$ which is covariantly constant along D . ς belongs to the set $Sec_D^\infty(\mathcal{L})$ and satisfies $pr \circ \varsigma = s \circ \pi_D$. The mapping $\sigma \mapsto s_\sigma$ is therefore surjective.

For an arbitrary point $m \in T^*Q$ let $W \subset T^*Q$ be a neighborhood of m and $u : \pi^{-1}(W) \rightarrow W \times \mathbb{C}$ a local trivialisation of $\mathcal{L}$. For every $a \in \mathbb{C}$ and $l, l_1, l_2 \in \pi^{-1}(m)$ define $a \cdot l$ and $l_1 + l_2$ respectively as $u(a \cdot l) = (m, a \cdot z)$ and $u(l_1 + l_2) = (m, z_1 + z_2)$, where z, z_1 and z_2 are complex numbers which satisfy

$$u(l) = (m, z), \quad u(l_1) = (m, z_1), \quad \text{and} \quad u(l_2) = (m, z_2).$$

Let m' be a point in T^*Q so that m' satisfies $\pi_D(m) = \pi_D(m')$. Further, let $l', l'_1, l'_2 \in \pi^{-1}(m')$ so that $l' \sim l, l'_1 \sim l_1$ and $l'_2 \sim l_2$. Since a parallel displacement relative to a connection in a vector bundle generates an isomorphism between two fibres of the bundle, we have then

$$a \cdot l \sim a \cdot l' \quad \text{and} \quad l_1 + l_2 \sim l'_1 + l'_2$$

for each $a \in \mathbb{C}$. The last identities lead to the linearity and therefore to the isomorphy of the mapping $\sigma \mapsto s_\sigma$. $\square$

Proposition 2.2 *Let $Sec_0^\infty(\eta)$ denotes the set of all sections $s \in Sec^\infty(\eta)$ which satisfy*

$$\int_Q \langle s, s \rangle_\eta |\varepsilon_g| < \infty,$$

and $(\cdot, \cdot)_\eta$ the scalar product defined in $Sec_0^\infty(\eta)$ by

$$(s', s)_\eta = \int_Q \langle s', s \rangle_\eta |\varepsilon_g|, \quad \forall s', s \in Sec_0^\infty(\eta).$$

Then the completion $\mathcal{H}$ of the scalar product space $(Sec_0^\infty(\eta), (\cdot, \cdot)_\eta)$ is a Hilbert space which is isometrically isomorphic to the Hilbert space $\mathcal{H}_{i\gamma, D}^\Theta$.

Proof : Let ψ and ψ' be any two elements of $\mathcal{H}_{i\gamma, D}$. Then there exists two

sections $\sigma, \sigma' \in \text{Sec}_D^\infty(\mathcal{L})$ so that $\psi = \sigma \otimes \nu$ and $\psi' = \sigma' \otimes \nu$. From Eq.(4) and Eq.(7) we have

$$(\psi, \psi')_{i_\gamma}(m) = \langle \sigma, \sigma' \rangle(m) |\varepsilon_g| \circ \pi_D(m); \quad \forall m \in T^*Q.$$

Next let s_σ and $s_{\sigma'}$ be a section of η which is the image of σ and σ' under the isomorphism $\text{Sec}_D^\infty(\mathcal{L}) \rightarrow S^\infty(\eta)$ respectively. Then it follows that

$$\langle s_\sigma, s_{\sigma'} \rangle_\eta \circ \pi_D(m) |\varepsilon_g| \circ \pi_D(m) = \langle pr \circ \sigma(m), pr \circ \sigma'(m) \rangle |\varepsilon_g| \circ \pi_D(m),$$

for every $m \in T^*Q$. According to Proposition 2.1 the projection pr conserves the hermitean structure of the bundle $\mathcal{L}$. Therefore

$$(\psi, \psi')_{i_\gamma}(m) = \langle s_\sigma, s_{\sigma'} \rangle_\eta \circ \pi_D(m) |\varepsilon_g| \circ \pi_D(m)$$

holds. However, it means $\int_Q (\psi, \psi')_{i_\gamma} = \int_Q \langle s_\sigma, s_{\sigma'} \rangle_\eta |\varepsilon_g|$. Consequently, s_σ lies in $\text{Sec}_0^\infty(\eta)$ if and only if $\sigma \otimes \nu$ is contained in $\mathcal{W}_{i_\gamma, D}^\Theta$ (see Eq.(8)). Thus the mapping $\mathcal{H}_{i_\gamma, D}^\Theta \ni \psi = \sigma \otimes \nu \mapsto s_\sigma \in \mathcal{H}$ is an isometric isomorphism from $\mathcal{H}_{i_\gamma, D}^\Theta$ onto $\mathcal{H}$. $\square$

Since $pr \circ [\mathbf{1}(m) \otimes (s_\alpha \circ \pi_D(m), m)] = \mathbf{1}_{\eta'}(q) \otimes s_\alpha(q)$ holds for every $q \in Q$ and $m \in T^*Q$ so that $\pi_D(m) = q$, the assignment $\mathbf{1} \otimes \mathfrak{s}_\alpha \mapsto \mathbf{1}_{\eta'} \otimes s_\alpha$ generates the isomorphism just obtained in the last proposition. Let σ be a section of $\text{Sec}_D^\infty(\mathcal{L})$ so that $\sigma \otimes \nu$ lies in $\mathcal{W}_{i_\gamma, D}^\Theta$. On each $\pi_D^{-1}(U_\alpha)$ the section σ can be written locally as $\sigma|_{\pi_D^{-1}(U_\alpha)} = (\varphi \circ \pi_D)(\mathbf{1} \otimes \mathfrak{s}_\alpha)$, where $\varphi \in C^\infty(U_\alpha, \mathbb{R})$. Then the isomorphism defined in Lemma 2.3 maps the section σ exactly in a section s_σ in $\text{Sec}_0^\infty(\eta)$ by $s_\sigma|_{U_\alpha} = \varphi(\mathbf{1}_{\eta'} \otimes s_\alpha)$ for all α in the index set. Conversely, for each $s' \in \text{Sec}_0^\infty(\eta)$ so that $s'|_{U_\alpha} = \phi(\mathbf{1}_{\eta'} \otimes s_\alpha)$ on every U_α there exists exactly an element $\sigma' \otimes \nu \in \mathcal{W}_{i_\gamma, D}^\Theta$, so daß $\sigma'|_{\pi_D^{-1}(U_\alpha)} = (\phi \circ \pi_D)(\mathbf{1} \otimes \mathfrak{s}_\alpha)$ on each $\pi_D^{-1}(U_\alpha)$. Therefore from Eq.(12) we obtain

$$\mathcal{Q}_{i_\gamma}(p_X)\varphi(\mathbf{1}_{\eta'} \otimes s_\alpha) = \{-iX(\varphi) - \Phi_\alpha(X)\varphi + (\frac{1}{2i} + \gamma)(\text{div}_{\varepsilon_g}(X))\varphi\}(\mathbf{1}_{\eta'} \otimes s_\alpha).$$

Finally, the quantization Eq.(11) and Eq.(12) can be written as

$$\mathcal{Q}_{i_\gamma}(q_f)s = fs \quad \text{and} \quad \mathcal{Q}_{i_\gamma}(p_X)s = -i \nabla_X^\Phi s + (\frac{1}{2i} + \gamma)(\text{div}_{\varepsilon_g}(X))s,$$

for all $s \in \mathcal{H}$. We summarize this in

Theorem 2.1. *Let (Q, g) be an orientable n -dimensional Riemannian manifold and θ_c the canonical symplectic potential on its cotangent bundle T^*Q so that $\omega := d\theta_c$ is the canonical symplectic structure. Furthermore, let (L', ∇^Φ) be a complex line bundle over Q with a hermitean connection ∇^Φ and $\{(U_\alpha, s_\alpha) : \alpha \in \mathbb{I}\}$ a local system of (L', ∇^Φ) so that $\nabla^\Phi s_\alpha = -i\Phi_\alpha s_\alpha$ holds, where Φ_α is a closed 1-form defined on U_α for every α in the index set $\mathbb{I}$.*

Then there exists a prequantization line bundle $\mathcal{L}$ with a hermitean connection ∇^Θ so that $\nabla^\Theta \varsigma_\alpha = -i\Theta_\alpha \varsigma_\alpha$ for a local system $\{(\pi_D^{-1}(U_\alpha), \varsigma_\alpha)\}$ of $\mathcal{L}$, where $\Theta_\alpha :=$

$\theta_c + \pi_D^* \Phi_\alpha$ are the potentials corresponding to the connection ∇^Θ and $\pi_D : T^*Q \rightarrow Q$ is the canonical projection.

The $-(\frac{1}{2} + i\gamma)$ -D-densities geometric quantization of the algebra $\Sigma(T^*Q)$ yields a differentiable 0-compatible elementary quantum Borel-kinematics on Q : for every $(f, X) \in \mathcal{S}(Q)$ and for all $s \in \mathcal{H} := \overline{\text{Sec}_0^\infty(L')}$

$$\mathbb{Q}(f)s = fs \quad \text{and} \quad \mathbb{P}(X)s = -i \nabla_X^\Phi s + (\frac{1}{2i} + \gamma)(\text{div}_{\varepsilon_g}(X))s.$$

2.2 When external fields present

If we give up the condition $d\Phi_\alpha = 0$, $\forall \alpha \in \mathbb{I}$, in the preceding section, i.e. the closedness of the potentials, then there is a global closed 2-form $\tilde{\Phi} \in \Omega^2(Q)$ which satisfies $\tilde{\Phi} = d\Phi_\alpha$, $\forall \alpha \in \mathbb{I}$. This means the presence of an external field surrounding our physical system.

As we saw in the preceding section, the closedness of the potentials guarantees that the bundle $(\mathcal{L}, \nabla^\Theta)$ is a prequantization line bundle as the curvature of ∇^Θ agrees with the symplectic structure ω . Now if an external field $\tilde{\Phi}$ presents, the curvature of ∇^Θ is $\omega + \pi_D^* \tilde{\Phi}$. Fortunately, the form $\tilde{\omega} := \omega + \pi_D^* \tilde{\Phi}$ constitutes also a symplectic structure on T^*Q . The modification of the symplectic structure alters the classical picture of our physical system. For instance, the Hamiltonian vector field generated by an observable $h \in C^\infty(T^*Q, \mathbb{R})$ is given by

$$X_h = \frac{\partial h}{\partial p_i} \frac{\partial}{\partial q^i} - (\frac{\partial h}{\partial q^j} + \tilde{\Phi}_{ij} \frac{\partial h}{\partial p_i}) \frac{\partial}{\partial p_j},$$

where $\tilde{\Phi}_{ij}$ are the components of $\tilde{\Phi}$. But the subbundle D of $T(T^*Q)$ which constitutes vertical polarization in the preceding subsection is still a polarization with the same leaves and generalized configuration space. Furthermore, for every $(f, X) \in \mathcal{S}(Q)$ the expression for $\Theta_\alpha(X_{qf})$ and $\Theta_\alpha(X_{p_X})$ do not change. It is still given by Eq.(9) and Eq.(10). We have therefore

Theorem 2.2. *Let (Q, g) be an orientable n -dimensional Riemannian manifold and θ_c the canonical symplectic potential on its cotangent bundle T^*Q so that $\omega = d\theta_c$ is the canonical symplectic structure. Furthermore, let (L', ∇^Φ) be a complex line bundle over Q with a hermitean connection ∇^Φ and $\{(U_\alpha, s_\alpha) : \alpha \in \mathbb{I}\}$ a local system of (L', ∇^Φ) so that $\nabla^\Phi s_\alpha = -i\Phi_\alpha s_\alpha$ holds, where Φ_α is a 1-form defined on U_α for every α in the index set $\mathbb{I}$.*

The cotangent bundle T^*Q endowed with 2-form $\tilde{\omega} := \omega + \pi_D^* \tilde{\Phi}$, where $\tilde{\Phi}$ is the unique field corresponding to the potentials Φ_α , constitutes a symplectic manifold. Then there exists a prequantization line bundle $\mathcal{L}$ over $(T^*Q, \tilde{\omega})$ with a hermitean connection ∇^Θ so that $\nabla^\Theta s_\alpha = -i\Theta_\alpha s_\alpha$ for a local system $\{(\pi_D^{-1}(U_\alpha), s_\alpha)\}$ of $\mathcal{L}$, where $\Theta_\alpha := \theta_c + \pi_D^* \Phi_\alpha$ are the potentials corresponding to the connection ∇^Θ .

The $-(\frac{1}{2} + i\gamma)$ -D-densities geometric quantization of the algebra $\Sigma(T^*Q)$ yields a differentiable $\tilde{\Phi}$ -compatible elementary quantum Borel-kinematics on Q .

Acknowledgement. The author would like to thank H.D. Doebner and J.D. Hennig for interesting discussion and to the Laboratory of Atomic and Nuclear Physics, Department of Physics, Gadjah Mada University, for wonderful atmosphere.

REFERENCES

1. B. ANGERMANN, H. D. DOEBNER, AND J. TOLAR, Quantum kinematics on smooth manifolds, in *Lect. Not. Math.* 1037, Springer-Verlag, Berlin, 1983.
2. B. ANGERMANN, *Über Quantisierungen lokalisierter Systeme - Physikalisch interpretierbare mathematische Modelle*, PhD thesis, Technische Universität Clausthal, 1983.
3. R.J. BLATTNER, QUANTIZATION AND REPRESENTATION THEORY, in *Proc. Sympos. Pure Math.*, vol. 26, Amer. Math. Soc., Providence, R.I., 147–165, 1973.
4. H.D. DOEBNER AND P. NATTERMANN, Borel Quantization : Kinematics and Dynamics, in *Acta Physica Polonica B*, **27**(10), 2327–2340, 1996.
5. M. DREES, *Zur Kinematik lokalisierter quantenmechanischer Systeme unter Berücksichtigung innerer Freiheitsgrade und äußerer Felder*, PhD thesis, Technische Universität Clausthal, 1993.
6. M. J. GOTAY, Obstructions to Quantization, in J.E. Marsden and S. Wiggins (editors), *The Juan Simo Memorial Volume*, Springer-Verlag, New York, 1998.
7. M.J. GOTAY AND H.B. GRUNDLING, On Quantizing T^*S^1 , *Rep. Math. Phys.* **40**, 1997.
8. J. D. HENNIG, Schrödinger equations and Hamiltonian mechanics, in H.D. Doebner, V.K. Dobrev, and P. Nattermann (editors), *Non-linear, deformed and irreversible quantum systems*, World Scientific, Singapore, 1995.
9. L. VAN HOVE, Sur certaine représentations unitaires d'un groupe de transformations, *Mem. Cl. Sci., Collect. Octavo, Acad. R. Belg.* 26, **6**, 1951.
10. A. A. KIRRILOV, Unitary representation of nilpotent Lie Group, *Usp. Mat. Nauk.* 17, **4**, 53–104, 1962.
11. B. KOSTANT, Quantization and Unitary Representation, in C. T. Taam (editor), *Lectures in Modern Analysis III*, **170**, Springer-Verlag, Berlin, 87–208, 1970.
12. P. NATTERMANN, *Dynamics in Borel-Quantization : Non-linear Schrödinger Equations vs. Master Equations*, PhD thesis, Technische Universität Clausthal, 1997.
13. M. F. ROSYID, On the Relation between Configuration and Phase Space Quantization I, *Journal of the Indonesian Mathematical Society*, **9**(1), 2003.
14. N. M. J. WOODHOUSE, *Geometric Quantization*, 2nd ed., Clarendon Press, Oxford, 1992.

M. F. ROSYID: Department of Physics, Gadjah Mada University and Institut for Science in Yogyakarta, Yogyakarta, Indonesia.
E-mail: farchani@ugm.ac.id.

ON THE GQ-CONSISTENT QUANTUM BOREL KINEMATICS : Unitary Representations of the Groups of Canonical Transformations

M. F. ROSYID

Abstract. The behaviors of the completeness of GQ-consistent kinematical algebras as well as of GQ-consistent quantum Borel kinematics under symplectomorphisms (canonical transformations) are studied. The results are implemented to some cases. The groups of canonical transformations on cotangent bundles are unitarily represented as the intertwining maps between quantum Borel kinematics on the configuration manifolds.

Key words and Phrases : kinematics, quantization, representation

1. INTRODUCTION

The present article is still in the spirit of the theory of quantization, notably the geometric quantization [10], the Borel quantization [5] and some references therein, and the relation among them [2, 6, 7]. It constitutes a continuation of [8] in which I have proposed a formulation of quantum Borel kinematics for symplectic manifolds with polarization. The formulation is referred to as GQ-consistent quantum Borel kinematics. In the article I have also pointed out that such formulation leads to ordinary quantum Borel kinematics on the generalized configuration spaces when it is applied to a certain class of symplectic manifolds with polarizations to which the application of the $-(\frac{1}{2} + i\gamma)$ -densities geometric quantization also leads to ordinary quantum Borel kinematics on the generalized configuration spaces.

In the present article, we consider symplectic manifolds with polarizations and the symplectomorphisms among them which are polarizations preserving. We study

2000 *Mathematics Subject Classification:* 53Z05, 81S10, 53D50

the behaviors of the completeness of GQ-consistent kinematical algebras as well as of GQ-consistent quantum Borel kinematics under such symplectomorphisms. We will also apply the results to some cases, for instance cotangent bundles with polarizations induced from the vertical ones. Our main results are Theorem 2.1 and Theorem 2.2.

Let (M, ω) be a quantizable symplectic manifold of dimension $2n$ and P a strong integrable complex polarization so that $0 \leq \dim(D) \leq n$, where $D^{\mathbb{C}} = P \cap \overline{P}$ is the isotropic distribution associated with P . In the $-(\frac{1}{2} + i\gamma)$ - P -densities quantization, a function $f \in C^\infty(M, \mathbb{R})$ is said to be quantizable if the Hamiltonian vector field X_f generated by f preserves the polarization P in the sense that¹ $[X_f, P] \subseteq P$. Let $\mathcal{F}_P(M, \mathbb{R})$ be the set of all real quantizable functions and $\mathcal{F}(M, \mathbb{R}; D) := \{f \in \mathcal{F}_P(M, \mathbb{R}) \mid [X_f, D] \subseteq D\} \subseteq \mathcal{F}_P(M, \mathbb{R})$ as the set of all quantizable functions preserving the distribution D .

Let Z be a vector field on M . The mapping $\pi_{D*}Z : M/D \ni \pi_D(m) \mapsto \pi_{D*}|_m Z \in T_{\pi_D(m)}(M/D)$ defines a smooth vector field on M/D if and only if Z preserves the distribution D , i.e. $[Z, D] \subseteq D$ [8]. It is clear therefore that a quantizable function f belongs to $\mathcal{F}(M, \mathbb{R}; D)$ if and only if there exists a differentiable vector field X on M/D , so that X_f and X are π_D -related, i.e. $\pi_{D*}X_f$ is a differentiable vector field on M/D .

The set $\mathfrak{X}(M; D)$ of all vector fields on M preserving the distribution D forms a Lie subalgebra of $\mathfrak{X}(M)$, the set of all smooth vector fields on M . Now let $\mathfrak{X}_{\mathcal{F}}(M; D)$ denote the set of all Hamiltonian vector fields on M generated by the functions in the set $\mathcal{F}(M, \mathbb{R}; D)$ and $\pi_{D*}^{\mathcal{F}}$ be defined as the restriction of the Lie homomorphism $\pi_{D*} : \mathfrak{X}(M; D) \rightarrow \mathfrak{X}(M/D)$ to $\mathfrak{X}_{\mathcal{F}}(M; D)$. The set $\mathfrak{X}_{\mathcal{F}}(M; D)$ is clearly a Lie subalgebra of $\mathfrak{X}(M; D)$ and $\pi_{D*}^{\mathcal{F}}$ is a Lie homomorphism. Let $\mathfrak{X}_{\mathcal{F}}^{\sim}(M; D)$ denote the quotient algebra of $\mathfrak{X}_{\mathcal{F}}(M; D)$ relative to the kernel $\mathcal{K}(\pi_{D*}^{\mathcal{F}})$ of $\pi_{D*}^{\mathcal{F}}$, and $[\cdot, \cdot]^{\sim}$ be the quotient bracket in $\mathfrak{X}_{\mathcal{F}}^{\sim}(M; D)$. If $\mathfrak{X}_{\mathcal{F},c}^{\sim}(M; D)$ is the subset of $\mathfrak{X}_{\mathcal{F}}^{\sim}(M; D)$ defined by

$$\mathfrak{X}_{\mathcal{F},c}^{\sim}(M; D) = \{[X_f] \in \mathfrak{X}_{\mathcal{F}}^{\sim}(M; D) \mid \pi_{D*}[X_f] \in \mathfrak{X}_c(M/D)\},$$

then in general, the identity $\pi_{D*}\mathfrak{X}_{\mathcal{F},c}^{\sim}(M; D) = \mathfrak{X}_c(M/D)$ is not respected, where $\pi_{D*}[X_f]$ is defined as $\pi_{D*}X_{f'}$ for arbitrary $X_{f'} \in [X_f]$ and $\mathfrak{X}_c(M/D)$ the set of all complete vector fields on M/D . Furthermore, the mapping $\pi_{D*}^c : \mathfrak{X}_{\mathcal{F},c}^{\sim}(M; D) \rightarrow \mathfrak{X}_c(M/D)$ which is defined by $\pi_{D*}^c[X_f] = \pi_{D*}X_{f'}$ for arbitrary $X_{f'} \in [X_f]$ is an injective partial Lie homomorphism.

Define now $\mathfrak{X}$ as the assignment $\mathcal{F}(M, \mathbb{R}; D) \ni f \mapsto X_f \in \mathfrak{X}_{\mathcal{F}}(M; D)$ and consider the kernel $\mathcal{K}(\pi_{D*}^{\mathcal{F}})$ of the Lie homomorphism $\pi_{D*}^{\mathcal{F}}$ in $\mathfrak{X}_{\mathcal{F}}(M; D)$. Further, let $\mathcal{F}_{\mathcal{K}}(M, \mathbb{R}; D)$ be the inverse image of $\mathcal{K}(\pi_{D*}^{\mathcal{F}})$ under the mapping $\mathfrak{X}$. If g is contained in $\mathcal{F}_{\mathcal{K}}(M, \mathbb{R}; D)$, then $\pi_{D*}^{\mathcal{F}}X_g = 0$. Therefore, there exists uniquely a real smooth function $\zeta_g \in C^\infty(M/D, \mathbb{R})$ so that $g = \zeta_g \circ \pi_D$. However, the converse is in general not correct. If $C_D^\infty(M, \mathbb{R})$ is the set of all real smooth functions on M which are constant on every leaf of D and P is real (so that $D = D^\perp$), then

¹The expression $[X_f, P] \subseteq P$ means that $[X_f, Y]$ is a cross section of P for every cross section Y of P . A cross section Y of P is a vector field on M so that $Y(m) \in P(m)$ for every $m \in M$.

$\mathcal{F}_K(M, \mathbb{R}; D) = C_D^\infty(M, \mathbb{R})$. Since, $C_D^\infty(M, \mathbb{R})$ is equal to

$$\pi_D^* C^\infty(M/D, \mathbb{R}) := \{\zeta \circ \pi_D \in C^\infty(M, \mathbb{R}) \mid \zeta \in C^\infty(M/D, \mathbb{R})\},$$

it follows also that $\mathcal{F}_K(M, \mathbb{R}; D) = \pi_D^* C^\infty(M/D, \mathbb{R})$.

Now let Ξ denote the injection $\mathcal{F}_K(M, \mathbb{R}; D) \ni g \mapsto \zeta_g \in C^\infty(M/D, \mathbb{R})$ and define for every equivalent class $[X_f] \in \mathfrak{X}_{\mathcal{F},c}^\sim(M; D)$ a linear operator $\mathcal{L}_{[X_f]}$ on the algebra $\mathcal{F}_K(M, \mathbb{R}; D)$ so that $\mathcal{L}_{[X_f]}g = X_{f'}(g)$, for every $g \in \mathcal{F}_K(M, \mathbb{R}; D)$ and arbitrary $X_{f'} \in [X_f]$.

Let $\mathcal{S}_{GQ}(M; D)$ be the semidirect sum $\mathcal{F}_K(M, \mathbb{R}; D) \oplus_s \mathfrak{X}_{\mathcal{F},c}^\sim(M; D)$ with Lie bracket $[\cdot, \cdot]^s$ defined by

$$[(g_1, [X_{f_1}]), (g_2, [X_{f_2}])]^s = (\mathcal{L}_{[X_{f_1}]}g_2 - \mathcal{L}_{[X_{f_2}]}g_1, [[X_{f_1}], [X_{f_2}]]^\sim)$$

for every $[X_{f_1}], [X_{f_2}] \in \mathfrak{X}_{\mathcal{F},c}^\sim(M; D)$ with $[[X_{f_1}], [X_{f_2}]]^\sim \in \mathfrak{X}_{\mathcal{F},c}^\sim(M; D)$ and all $g_1, g_2 \in \mathcal{F}_K(M, \mathbb{R}; D)$. The pair $(\mathcal{S}_{GQ}(M; D), [\cdot, \cdot]^s)$ is called *GQ-consistent kinematical algebra* in (M, ω, P) .

Define a mapping, denoted by $\Xi \oplus_s \pi_{D*}^c$, from the algebra $\mathcal{S}_{GQ}(M; D)$ into the kinematical algebra $\mathcal{S}(M/D) = C^\infty(M/D, \mathbb{R}) \oplus_s \mathfrak{X}_c(M/D)$ through

$$\mathcal{S}_{GQ}(M; D) \ni (g, [X_f]) \mapsto (\Xi(g), \pi_{D*}^c[X_f]) \in \mathcal{S}(M/D).$$

The mapping $\Xi \oplus_s \pi_{D*}^c$ is clearly a partial Lie homomorphism. The GQ-consistent kinematical algebra $\mathcal{S}_{GQ}(M; D)$ is said to be *almost complete* whenever the mapping $\Xi \oplus_s \pi_{D*}^c$ is a partial Lie isomorphism. Furthermore, $\mathcal{S}_{GQ}(M; D)$ is said to be *complete* if it is almost complete and $\mathfrak{X}_{\mathcal{F},c}^\sim(M; D)$ is equal to $\mathfrak{X}_{\mathcal{F}}^\sim(M; D)$.

Let Φ be a closed integral real 2-form on M/D and $(\mathfrak{H}, \mathbb{Q}, \mathbb{P}, \theta^\infty)$ an elementary differentiable Φ -compactible quantum Borel kinematic on M/D [5]. Let $\mathbb{Q}_{GQ}$ be the restriction $\mathbb{Q}|_{\Xi(\mathcal{F}_K(M, \mathbb{R}; D))}$ and $\mathbb{P}_{GQ}$ the restriction $\mathbb{P}|_{\pi_{D*}^c(\mathfrak{X}_{\mathcal{F},c}^\sim(M; D))}$. The *GQ-consistent quantum Borel kinematics subordinated to the quantum Borel kinematic* $(\mathfrak{H}, \mathbb{Q}, \mathbb{P}, \theta^\infty)$ for the symplectic manifold with polarization (M, ω, P) is the quadruple $(\mathfrak{H}, \mathbb{Q}_{GQ}, \mathbb{P}_{GQ}, \vartheta^\infty)$. A GQ-consistent quantum Borel kinematics is said to be almost complete whenever the associated GQ-consistent kinematical algebra is almost complete. A GQ-consistent quantum Borel kinematics is said to be complete whenever the associated GQ-consistent kinematical algebra is complete.

If $(\mathfrak{H}, \mathbb{Q}, \mathbb{P}, \vartheta^\infty)$ and $(\mathfrak{H}', \mathbb{Q}', \mathbb{P}', \vartheta'^\infty)$ are two differentiable quantum Borel kinematics on M/D which are equivalent, then the GQ-consistent quantum Borel kinematics $(\mathfrak{H}, \mathbb{Q}_{GQ}, \mathbb{P}_{GQ}, \vartheta^\infty)$ and $(\mathfrak{H}', \mathbb{Q}'_{GQ}, \mathbb{P}'_{GQ}, \vartheta'^\infty)$ are said to be equivalent.

2. MAIN RESULTS

Let (M_i, ω_i, P_i) , with $i = 1, 2$, be two symplectic manifolds with polarizations and $\rho : M_1 \rightarrow M_2$ a symplectomorphism from (M_1, ω_1, P_1) onto (M_2, ω_2, P_2) . The symplectomorphism $\rho : M_1 \rightarrow M_2$ is said to be $P_1 P_2$ -preserving² whenever for

² $X_1 \in P_1$ means, $X_1(m_1) \in P(m_1)$ for every $m_1 \in M_1$.

every $X_1 \in P_1$ there is exactly a vector field $X_2 \in P_2$ which is ρ -related to X_1 , i.e. $X_2 = \rho_* X_1$. The symplectomorphism $\rho : M_1 \rightarrow M_2$ is called *polarizations preserving symplectomorphism* when ρ is $P_1 P_2$ -preserving and in the same time is also $P_2 P_1$ -preserving. From now on we assume that all symplectomorphisms involved in the sequel are polarizations preserving symplectomorphisms.

Let $D_i = P_i \cap \overline{P_i} \cap TM_i$ ($i = 1, 2$). The symplectomorphism ρ induces then a diffeomorphism $\varphi : M_1/D_1 \rightarrow M_2/D_2$ so that the diagram

$$\begin{array}{ccc} M_1 & \xrightarrow{\rho} & M_2 \\ \pi_{D_1} \downarrow & & \downarrow \pi_{D_2} \\ M_1/D_1 & \xrightarrow{\varphi} & M_2/D_2 \end{array} \quad (1)$$

is commutative. Therefore, ρ maps every leaf of D_1 in exactly a leaf of D_2 .

Let $\mathfrak{R}_1$ denote the assignment $X \mapsto (\rho^{-1})_* X$ from the Lie algebra $\mathfrak{X}(M_2; D_2)$ of all vector fields on M_2 which preserve D_2 into the Lie algebra $\mathfrak{X}(M_1; D_1)$ of all vector fields defined on M_1 which preserve D_1 and $\mathfrak{R}_2$ be the mapping $\mathfrak{X}_{\mathcal{F}}(M_2; D_2) \ni X_f \mapsto X_{f \circ \rho} \in \mathfrak{X}_{\mathcal{F}}(M_1; D_1)$. Then, it is clear that $\mathfrak{R}_1$ and $\mathfrak{R}_2$ are Lie isomorphisms.

Furthermore, let $\mathfrak{R}_3$ be the mapping defined by

$$\mathfrak{R}_3 : \mathfrak{X}_{\mathcal{F}}(M_2; D_2) \ni [X_f] \mapsto [\mathfrak{R}_2(X_f)] = [X_{f \circ \rho}] \in \mathfrak{X}_{\mathcal{F}}(M_1; D_1),$$

and $\mathfrak{R}_4$ be defined as the restriction $\mathfrak{R}_3|_{\mathfrak{X}_{\mathcal{F},c}(M_2; D_2)}$ of $\mathfrak{R}_3$ to $\mathfrak{X}_{\mathcal{F},c}(M_2; D_2)$. Then, $\mathfrak{R}_3$ is clearly a Lie isomorphism and $\mathfrak{R}_4$ is a partial Lie isomorphism.

It is straightforward to see that the diagram depicted in Fig.1 is commutative. In the diagram, i means the inclusion and π_i ($i = 1, 2$) the natural projection.

We consider now the subalgebra $\mathcal{F}_{\mathcal{K}}(M_i, \mathbb{R}, D_i)$. When f_2 is contained in $\mathcal{F}_{\mathcal{K}}(M_2, \mathbb{R}, D_2)$, then $X_{f_2} \in D_2$. Since ρ^{-1} is $P_2 P_1$ -preserving, we have $(\rho^{-1})_* X_{f_2} \in D_1$. Hence, it follows that

$$\pi_{D_1*}(\rho^{-1})_* X_{f_2} = 0.$$

Since $(\rho^{-1})_* X_{f_2} \lrcorner \omega = -d(f_2 \circ \rho)$, $(\rho^{-1})_* X_{f_2}$ is the Hamiltonian vector field generated by $f_2 \circ \rho$. Therefore, $f_2 \circ \rho$ belongs to the algebra $\mathcal{F}(M_1, \mathbb{R}, D_1)$. and $f_2 \circ \rho$ is even contained in the set $\mathcal{F}_{\mathcal{K}}(M_1, \mathbb{R}, D_1)$. Conversely, one can show that $f_1 \circ \rho^{-1}$ belongs to $\mathcal{F}_{\mathcal{K}}(M_2, \mathbb{R}, D_2)$ for every function $f_1 \in \mathcal{F}_{\mathcal{K}}(M_1, \mathbb{R}, D_1)$.

Let $\mathfrak{r}$ denote the map $\mathcal{F}_{\mathcal{K}}(M_2, \mathbb{R}, D_2) \ni f \mapsto f \circ \rho \in \mathcal{F}_{\mathcal{K}}(M_1, \mathbb{R}, D_1)$. It is clear that $\mathfrak{r}$ is bijective. The diagram Eq.(1) leads also to the following commutative diagram

$$\begin{array}{ccc} \mathcal{F}_{\mathcal{K}}(M_1, \mathbb{R}, D_1) & \xleftarrow{\mathfrak{r}} & \mathcal{F}_{\mathcal{K}}(M_2, \mathbb{R}, D_2) \\ \Xi_1 \downarrow & & \downarrow \Xi_2 \\ C^\infty(M_1/D_1, \mathbb{R}) & \xleftarrow{\mathfrak{r}_\varphi} & C^\infty(M_2/D_2, \mathbb{R}), \end{array} \quad (2)$$

where Ξ_i ($i = 1$ or 2) is the assignment $\mathcal{F}_{\mathcal{K}}(M_i, \mathbb{R}, D_i) \ni f_i \mapsto \zeta_{f_i} \in C^\infty(M_i/D_i, \mathbb{R})$ with $f_i = \zeta_{f_i} \circ \pi_{D_i}$ and $\mathfrak{r}_\varphi$ is the assignment $C^\infty(M_2/D_2, \mathbb{R}) \ni \zeta \mapsto \zeta \circ \varphi \in C^\infty(M_1/D_1, \mathbb{R})$.

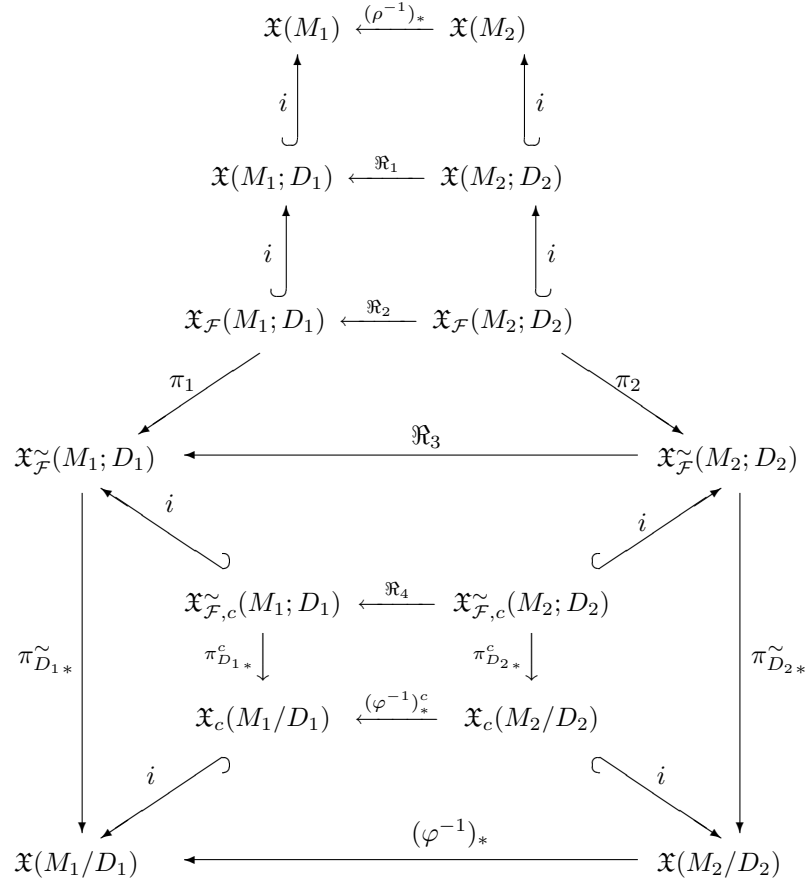


Figure 1

From diagram Fig.1 and the diagram Eq.(2), it follows then the commutative diagram

$$\begin{array}{ccc}
 \mathcal{F}_{\mathcal{K}}(M_1, \mathbb{R}, D_1) \oplus_s \mathfrak{X}_{\mathcal{F},c}^{\sim}(M_1; D_1) & \xleftarrow{\mathfrak{r} \oplus_s \mathfrak{R}_4} & \mathcal{F}_{\mathcal{K}}(M_2, \mathbb{R}, D_2) \oplus_s \mathfrak{X}_{\mathcal{F},c}^{\sim}(M_2; D_2) \\
 \Xi_1 \oplus_s \pi_{D_1*}^c \downarrow & & \downarrow \Xi_2 \oplus_s \pi_{D_2*}^c \\
 C^\infty(M_1/D_1, \mathbb{R}) \oplus_s \mathfrak{X}_c(M_1/D_1) & \xleftarrow{\mathfrak{r}_\varphi \oplus_s (\varphi^{-1})_*^c} & C^\infty(M_2/D_2, \mathbb{R}) \oplus_s \mathfrak{X}_c(M_2/D_2),
 \end{array} \tag{3}$$

where $\mathfrak{r} \oplus_s \mathfrak{R}_4$ is the bijection $(g_2, [X_{f_2}]) \mapsto (\mathfrak{r}(g_1), \mathfrak{R}_4([X_{f_1}]))$ and $\mathfrak{r}_\varphi \oplus_s (\varphi^{-1})_*^c$ is the bijection $(\zeta, X) \mapsto (\mathfrak{r}_\varphi(\zeta), (\varphi^{-1})_*^c(X))$.

Lemma 2.1 *Let (M_i, ω_i) ($i = 1, 2$) be symplectic manifolds with reducible polariza-*

tions P_i and $\rho : M_1 \rightarrow M_2$ a polarizations preserving symplectomorphism. Let D_i ($i = 1, 2$) denote the isotropic distribution associated to P_i , i.e. $D_i^c = P_i \cap \overline{P_i}$. The GQ-consistent kinematical algebra $\mathcal{S}_{GQ}(M_1; D_1)$ is almost complete if and only if the GQ-consistent kinematical algebra $\mathcal{S}_{GQ}(M_2; D_2)$ is almost complete. The GQ-consistent kinematical algebra $\mathcal{S}_{GQ}(M_1; D_1)$ is complete if and only if the GQ-consistent kinematical algebra $\mathcal{S}_{GQ}(M_2; D_2)$ is complete.

Proof : The first statement means, $\Xi_1 \oplus_s \pi_{D_1*}^c$ is a partial Lie isomorphism if and only if $\Xi_2 \oplus_s \pi_{D_2*}^c$ is so. The last statement follows however from the commutative diagram Eq.(3).

According to the definition of the completeness of GQ-consistent kinematical algebras, a GQ-consistent kinematical algebra $\mathcal{S}_{GQ}(M; D)$ for a symplectic manifold (M, ω) with reducible polarization P is complete when $\mathcal{S}_{GQ}(M; D)$ is almost complete and in the same time $\mathfrak{X}_{\mathcal{F},c}^\sim(M; D)$ equals $\mathfrak{X}_{\mathcal{F}}^\sim(M; D)$. The second statement of the lemma means therefore, $\pi_{D_1*}^c$ is a partial Lie isomorphism and $\mathfrak{X}_{\mathcal{F},c}^\sim(M_1; D_1) = \mathfrak{X}_{\mathcal{F}}^\sim(M_1; D_1)$ if and only if $\pi_{D_2*}^c$ is a partial Lie isomorphism and $\mathfrak{X}_{\mathcal{F},c}^\sim(M_1; D_1) = \mathfrak{X}_{\mathcal{F}}^\sim(M_2; D_2)$. According to the first statement of the lemma, this is equivalent to

$$\mathfrak{X}_{\mathcal{F},c}^\sim(M_1; D_1) = \mathfrak{X}_{\mathcal{F}}^\sim(M_1; D_1) \iff \mathfrak{X}_{\mathcal{F},c}^\sim(M_2; D_2) = \mathfrak{X}_{\mathcal{F}}^\sim(M_2; D_2)$$

which follows from the diagram in Fig.1, where the maps i should be identity. $\square$

2.2 The Behavior of GQ-consistent Quantum Borel Kinematics

Lemma 2.2 *Let Q_1 and Q_2 be smooth manifolds, $\varphi : Q_1 \rightarrow Q_2$ a diffeomorphism from Q_1 onto Q_2 , ν_2 a smooth Borel measure on Q_2 , and X a complete smooth vector field on Q_1 . Then, $\text{div}_{\nu_2}(\varphi_* X) = [\text{div}_{\nu_2 \circ \varphi}(X)] \circ \varphi^{-1}$ holds.*

Proof : It is clear that $\varphi_* X$ is a complete vector field on Q_2 . Let $\phi^{\varphi_* X}$ be the flow of $\varphi_* X$. Note that

$$\text{div}_{\nu_2}(\varphi_* X) = -\frac{\partial}{\partial t} \Big|_{t=0} \varrho^{\varphi_* X}$$

holds (see for instance in [1]), where $\varrho^{\varphi_* X}$ is a smooth function

$$\begin{aligned} \varrho^{\varphi_* X} &: \mathbb{R} \times Q_2 \rightarrow \mathbb{R} \\ &: (t, q_2) \mapsto \varrho_t^{\varphi_* X}(q_2) \end{aligned}$$

with $\nu_2 \circ \phi_{-t}^{\varphi_* X} = \varrho_t^{\varphi_* X} \cdot \nu_2$. Let ν_1 denote $\nu_2 \circ \varphi$. Then, we obtain by the use of Proposition 1.7 in [3]

$$\begin{aligned} \nu_1 \circ \varphi^{-1} \circ \phi_{-t}^{\varphi_* X} &= \varrho_t^{\varphi_* X} \cdot (\nu_1 \circ \varphi^{-1}) \\ \iff \nu_1 \circ \varphi^{-1} \circ \phi_{-t}^{\varphi_* X} \circ \varphi &= (\varrho_t^{\varphi_* X} \circ \varphi) \cdot \nu_1 \\ \iff \nu_1 \circ \phi_{-t}^X &= \varrho_t^X \cdot \nu_1, \end{aligned}$$

where $\varrho^X := \varrho^{\varphi_* X} \circ \varphi$ and ϕ_t^X is the flow of X . Finally, it follows that

$$\text{div}_{\nu_2}(\varphi_* X) = -\frac{\partial}{\partial t} \Big|_{t=0} \varrho^{\varphi_* X} = -\frac{\partial}{\partial t} \Big|_{t=0} \varrho^X \circ \varphi^{-1} = [\text{div}_{\nu_1}(X)] \circ \varphi^{-1},$$

with $\text{div}_{\nu_1}(X) = -\frac{\partial}{\partial t}|_{t=0}\varrho^X$. $\square$

Now let Φ be an integral closed 2-Form on M_1/D_1 and $(\mathfrak{H}, \mathbb{Q}, \mathbb{P}, \theta^\infty)$ a elementary differentiable Φ -compatible quantum Borel kinematics on M_1/D_1 .

Furthermore we define the quadrupel $((\varphi^{-1})^*\mathfrak{H}, \mathfrak{r}_\varphi^*\mathbb{Q}, (\varphi^{-1})^*\mathbb{P}, (\varphi^{-1})^*\theta^\infty)$ via

1. $(\varphi^{-1})^*\mathfrak{H} = L^2((\varphi^{-1})^*\eta, \langle \cdot, \cdot \rangle_2, (\varphi^{-1})^*\nu)$, where $(\varphi^{-1})^*\eta$ is the line bundle over M_2/D_2 induced from η under (φ^{-1}) . $(\varphi^{-1})^*\eta$ is given by

$$(\varphi^{-1})^*\eta = \bigcup_{x_2 \in M_2/D_2} (\eta_{\varphi^{-1}(x_2)}, x_2)$$

2. $\mathfrak{r}_\varphi^*\mathbb{Q}(\zeta_2) := [\mathbb{Q} \circ \mathfrak{r}_\varphi(\zeta_2)] \circ \varphi^{-1}$ for every $\zeta_2 \in C^\infty(M_2/D_2, \mathbb{R})$, where

$$\mathfrak{r}_\varphi : C^\infty(M_2/D_2, \mathbb{R}) \ni \zeta_2 \mapsto \zeta_2 \circ \varphi \in C^\infty(M_1/D_1, \mathbb{R})$$

3. $(\varphi^{-1})^*\theta^\infty = \{\sigma_1 \circ \varphi^{-1} \in \text{Sec}_0^\infty((\varphi^{-1})^*\eta) | \sigma_1 \in \text{Sec}_0^\infty(\eta)\}$
4. $(\varphi^{-1})^*\mathbb{P}(X_2)\sigma_2 := [\mathbb{P}((\varphi^{-1})^c X_2)\sigma_1] \circ \varphi^{-1}$ for every $\sigma_2 \in (\varphi^{-1})^*\theta^\infty$ and $X_2 \in \mathfrak{X}_c(M_2/D_2)$, where $(\varphi^{-1})^c$ is the restriction $(\varphi^{-1})^*|_{\mathfrak{X}_c(M_2/D_2)}$ of $(\varphi^{-1})^*$ to $\mathfrak{X}_c(M_2/D_2)$ and $\sigma_1 \in \theta^\infty$ so that $\sigma_1 = \sigma_2 \circ \varphi$.

Lemma 2.3 *The quadrupel $((\varphi^{-1})^*\mathfrak{H}, \mathfrak{r}_\varphi^*\mathbb{Q}, (\varphi^{-1})^*\mathbb{P}, (\varphi^{-1})^*\theta^\infty)$ is an elementary differentiable $(\varphi^{-1})^*\Phi$ -compatible quantum Borel kinematik on M_2/D_2 so that*

$$\text{Dom}(\mathfrak{r}_\varphi^*\mathbb{Q}(\zeta_2)) = \{\psi_1 \circ \varphi^{-1} | \psi_1 \in \text{Dom}(\mathbb{Q}(\zeta_2 \circ \varphi))\}$$

for every $\zeta_2 \in C^\infty(M_2/D_2, \mathbb{R})$. Furthermore the assignment

$$(\mathfrak{H}, \mathbb{Q}, \mathbb{P}, \theta^\infty) \mapsto ((\varphi^{-1})^*\mathfrak{H}, \mathfrak{r}_\varphi^*\mathbb{Q}, (\varphi^{-1})^*\mathbb{P}, (\varphi^{-1})^*\theta^\infty)$$

defines a bijection $\mathfrak{M}_\rho$ from the set $\mathfrak{Q}\mathfrak{B}\mathfrak{K}_1$ of all elementary differentiable quantum Borel kinematics on M_1/D_1 into the set $\mathfrak{Q}\mathfrak{B}\mathfrak{K}_2$ of all elementary differentiable quantum Borel kinematics on M_2/D_2 . The map $\mathfrak{M}_\rho$ preserves equivalent classes.

Proof :

1. Let $\tau_{(\varphi^{-1})}$ denote the map $((\varphi^{-1})^*\eta)_{x_2} \ni (z, x_2) \mapsto z \in \eta_{\varphi^{-1}(x_2)}$. Then, $\tau_{(\varphi^{-1})}$ is a map of line bundles so that $\tau_{(\varphi^{-1})}$ is the lift of φ^{-1} . The map $\tau_{(\varphi^{-1})}$ is even a map of line bundles with connections and $((\varphi^{-1})^*\eta, \tau_{(\varphi^{-1})}^*\theta)$ a line bundles over M_2/D_2 with connection $\tau_{(\varphi^{-1})}^*\theta$. Let $\{(V_\alpha, s_\alpha)\}_{\alpha \in I}$ be a local system of η with transition functions $c_{\alpha\beta}$ so that the connection ∇ can be written locally as $\nabla s_\alpha = -i\theta_\alpha s_\alpha$, where $\theta_\alpha = s_\alpha^*\theta$. This induces (see [7]) a local system $\{(\varphi(V_\alpha), \mathfrak{s}_\alpha)\}_{\alpha \in I}$ of the bundle $(\varphi^{-1})^*\eta$ with transition functions $\mathfrak{c}_{\alpha\beta}$ so that the induced connection $(\varphi^{-1})^*\nabla$ in $(\varphi^{-1})^*\eta$ is given by

$$(\varphi^{-1})^*\nabla \mathfrak{s}_\alpha = -i(\varphi^{-1})^*\theta_\alpha \mathfrak{s}_\alpha.$$

Since the connection ∇ is hermitean, it follows according to Proposition 1.9.1 of [4] that the real 1-form $i(\theta - \bar{\theta})$ is exact. Therefore, the real form $i\tau_{(\varphi^{-1})}^*(\theta - \bar{\theta})$ is

also exact. The bundle $(\varphi^{-1})^*\eta$ admits a hermitean structure which is compatible with the connection $(\varphi^{-1})^*\nabla$. Let $\langle \cdot, \cdot \rangle_2$ denote this structure.

Furthermore, it is clear that $\nu \circ \varphi^{-1}$ is also a differentiable Borel measure defined on the σ -algebra $\mathcal{B}(M_2/D_2)$ of Borel subsets of M_2/D_2 .

2. Define $\mathbb{Q}_2(\zeta_2)$ for every $\zeta_2 \in C^\infty(M_2/D_2, \mathbb{R})$ through $\mathbb{Q}_2(\zeta_2)\psi_2 = \zeta_2\psi_2$ for every ψ_2 in

$$\text{Dom}(\mathbb{Q}_2(\zeta_2)) = \{\psi_2 \in (\varphi^{-1})^*\mathfrak{H} \mid \int_{M_2/D_2} |\zeta_2|^2 \|\psi_2\|^2 d(\nu \circ \varphi^{-1}) < \infty\}.$$

Then it follows that $\mathbb{Q}_2(\zeta_2)\psi_2 = \zeta_2\psi_2 = (\zeta_1 \circ \varphi^{-1})\psi_2 = \mathfrak{r}_\varphi^*\mathbb{Q}(\zeta_2)\psi_2$, where $\zeta_1 \in C^\infty(M_1/D_1, \mathbb{R})$ with $\zeta_1 \circ \varphi^{-1} = \zeta_2$. It can be pointed out that

$$\text{Dom}(\mathfrak{r}_\varphi^*\mathbb{Q}(\zeta_2)) = \{\psi_1 \circ \varphi^{-1} \mid \psi_1 \in \text{Dom}(\mathbb{Q}(\zeta_2 \circ \varphi))\}$$

for every $\zeta_2 \in C^\infty(M_2/D_2, \mathbb{R})$.

3. It is clear that $(\varphi^{-1})^*\theta^\infty$ is the set of all smooth sections of $(\varphi^{-1})^*\eta$ which have compact supports.

4. For every $X_2 \in \mathfrak{X}_c(M_2/D_2)$ there exists uniquely a complete vector field X_1 on M_1/D_1 so that $X_1 = (\varphi^{-1})_*^c X_2$. Therefore, one has

$$\mathbb{P}((\varphi^{-1})_*^c X_2)\sigma_1 = -i \nabla_{(\varphi^{-1})_*^c X_2} \sigma_1 + \left(\frac{1}{2i} + \gamma\right) \text{div}_\nu((\varphi^{-1})_*^c X_2)\sigma_1,$$

for every $\sigma_1 \in \theta^\infty$. From the definition of $(\varphi^{-1})^*\mathbb{P}$ and by the use of Lemma 2.2 it can be shown that

$$[(\varphi^{-1})^*\mathbb{P}(X_2)\sigma_2] \circ \varphi = [-i(\varphi^{-1})^* \nabla_{X_2} \sigma_2 + \left(\frac{1}{2i} + \gamma\right) \text{div}_{\nu_2}(X_2)\sigma_2] \circ \varphi$$

where $\sigma_2 \in \text{Sec}_0^\infty((\varphi^{-1})^*\eta)$ with $\sigma_1 = \sigma_2 \circ \varphi$ and $\nu_2 := \nu \circ \varphi^{-1}$. This is equivalent however to

$$(\varphi^{-1})^*\mathbb{P}(X_2)\sigma_2 = -i(\varphi^{-1})^* \nabla_{X_2} \sigma_2 + \left(\frac{1}{2i} + \gamma\right) \text{div}_{\nu_2}(X_2)\sigma_2,$$

for every $X_2 \in \mathfrak{X}_c(M_2/D_2)$ and $\sigma_2 \in (\varphi^{-1})^*\theta^\infty$.

That the map $\mathfrak{M}_\rho$ is bijective is already clear, since φ is a diffeomorphism. That $\mathfrak{M}_\rho$ preserves equivalent classes follows from the fact that $\tau_{(\varphi^{-1})}$ is a mapping of line bundles with connections : Let $(\mathfrak{H}', \mathbb{Q}', \mathbb{P}', \theta'^\infty)$ be an elementary differentiable Φ -compatible quantum Borel kinematics on M_1/D_1 which is equivalent to $(\mathfrak{H}, \mathbb{Q}, \mathbb{P}, \theta^\infty)$. The quadruple

$$\mathfrak{M}_\rho((\mathfrak{H}', \mathbb{Q}', \mathbb{P}', \theta'^\infty)) = ((\varphi^{-1})^*\mathfrak{H}', \mathfrak{r}_\varphi^*\mathbb{Q}', (\varphi^{-1})^*\mathbb{P}', (\varphi^{-1})^*\theta'^\infty)$$

is therefore an elementary differentiable $(\varphi^{-1})^*\Phi$ -compatible quantum Borel kinematics on M_2/D_2 . Hence, it is the equivalence of line bundles with connection

which remains to be shown. Let (η', θ') be the hermitean line bundles with connections θ' so that $\mathfrak{H}' = L^2(\eta', \langle \cdot, \cdot \rangle_1, \nu')$; further let $\tau'_{(\varphi^{-1})}$ be the natural map $(\varphi^{-1})^*\eta' \rightarrow \eta'$ of line bundles with connections which is defined analogous to the map $\tau_{(\varphi^{-1})}$. Since (η', θ') and (η, θ) are equivalent, there is a diffeomorphism $\tau_1 : \eta' \rightarrow \eta$ so that τ_1 induces an isomorphism $\eta'_{x_1} \rightarrow \eta_{x_1}$ for every $x_1 \in M_1/D_1$ and $\tau_1^*\theta = \theta'$. Define τ_2 a composition $\tau_\varphi \circ \tau_1 \circ \tau'_{(\varphi^{-1})}$. Then, τ_2 induces an isomorphism $((\varphi^{-1})^*\eta')_{x_2} \rightarrow ((\varphi^{-1})^*\eta)_{x_2}$ for every $x_2 \in M_2/D_2$ and $\tau_2^*(\tau_{(\varphi^{-1})}^*\theta) = \tau_{(\varphi^{-1})}^*\theta'$ holds also, i.e. $((\varphi^{-1})^*\eta', \tau_{(\varphi^{-1})}^*\theta')$ is equivalent to $((\varphi^{-1})^*\eta, \tau_{(\varphi^{-1})}^*\theta)$. The diagram

$$\begin{array}{ccc} ((\varphi^{-1})^*\eta', \tau_{(\varphi^{-1})}^*\theta') & \xrightarrow{\tau_2} & ((\varphi^{-1})^*\eta, \tau_{(\varphi^{-1})}^*\theta) \\ \tau'_{(\varphi^{-1})} \downarrow & & \tau_{(\varphi^{-1})} \downarrow \\ (\eta', \theta') & \xrightarrow{\tau_1} & (\eta, \theta). \end{array}$$

describes the equivalence of the line bundles with connections. $\square$

Now we state our first main result as

Theorem 2.1 *Let $(\mathfrak{H}, \mathbb{Q}_{GQ}, \mathbb{P}_{GQ}, \theta^\infty)$ be a smooth GQ-consistent Φ -comptible quantum Borel kinematics for (M_1, ω_1, P_1) which is subordinated by an elementary differentiable Φ -comptible quantum Borel kinematics $(\mathfrak{H}, \mathbb{Q}, \mathbb{P}, \theta^\infty)$ in $\mathfrak{QBK}_1$. Then $((\varphi^{-1})^*\mathfrak{H}, \mathfrak{r}_\varphi^*\mathbb{Q}_{GQ}, (\varphi^{-1})^*\mathbb{P}_{GQ}, (\varphi^{-1})^*\theta^\infty)$ is an elementary smooth GQ-consistent Φ -compatible quantum Borel kinematics for (M_2, ω_2, P_2) which is subordinated by $((\varphi^{-1})^*\mathfrak{H}, \mathfrak{r}_\varphi^*\mathbb{Q}, (\varphi^{-1})^*\mathbb{P}, (\varphi^{-1})^*\theta^\infty)$. The bijection $\mathfrak{M}_\rho$ induces a map*

$$\mathfrak{M}_\rho^{GQ} : \mathfrak{QBK}_1^{GQ} \rightarrow \mathfrak{QBK}_2^{GQ}$$

which is also bijective and equivalence classes preserving, where $\mathfrak{QBK}_1^{GQ}$ is the set of all elementary differentiable GQ-consistent quantum Borel kinematics for (M_1, ω_1, P_1) and $\mathfrak{QBK}_2^{GQ}$ the set of all elementary differentiable GQ-consistent quantum Borel kinematics for (M_2, ω_2, P_2) .

Proof : For the first statement, it is sufficient to show that

$$[(\varphi^{-1})^*\mathbb{P}]_{GQ} \equiv (\varphi^{-1})^*\mathbb{P}|_{\pi_{D_2*}^c(\mathfrak{X}_{\mathcal{F},c}^\sim(M_2; D_2))} = (\varphi^{-1})^*\mathbb{P}_{GQ}$$

and

$$(\mathfrak{r}_\varphi^*\mathbb{Q})_{GQ} \equiv \mathfrak{r}_\varphi^*\mathbb{Q}|_{\Xi_2(\mathcal{F}_K(M_2, \mathbb{R}; D_2))} = \mathfrak{r}_\varphi^*\mathbb{Q}_{GQ}. \quad (4)$$

From the following commutative diagram

$$\begin{array}{ccc} \mathfrak{X}_{\mathcal{F},c}^\sim(M_1; D_1) & \xleftarrow{\mathfrak{R}_4} & \mathfrak{X}_{\mathcal{F},c}^\sim(M_2; D_2) \\ \pi_{D_1*}^c \downarrow & & \pi_{D_2*}^c \downarrow \\ \pi_{D_1*}^c(\mathfrak{X}_{\mathcal{F},c}^\sim(M_1; D_1)) & \xleftarrow{(\varphi^{-1})_*^c} & \pi_{D_2*}^c(\mathfrak{X}_{\mathcal{F},c}^\sim(M_2; D_2)), \end{array}$$

we obtain then $(\varphi^{-1})_*^c(\pi_{D_2*}^c(\mathfrak{X}_{\mathcal{F},c}^\sim(M_2; D_2))) = \pi_{D_1*}^c(\mathfrak{X}_{\mathcal{F},c}^\sim(M_1; D_1))$. Since

$$(\varphi^{-1})^*\mathbb{P}(X_2)\sigma_2 = [\mathbb{P}((\varphi^{-1})_*^c X_2)\sigma_1] \circ \varphi^{-1},$$

where $\sigma_1 = \sigma_2 \circ \varphi$, the restriction of $(\varphi^{-1})^*\mathbb{P}$ to $\pi_{D_2*}^c(\mathfrak{X}_{\mathcal{F},c}^\sim(M_2; D_2))$ means also the restriction of $\mathbb{P}$ to $\pi_{D_1*}^c(\mathfrak{X}_{\mathcal{F},c}^\sim(M_1; D_1))$, i.e. $\mathbb{P}_{GQ}$. Hence, we have

$$[(\varphi^{-1})^*\mathbb{P}]_{GQ}(\pi_{D_2*}^c(\mathfrak{X}_{\mathcal{F},c}^\sim(M_2; D_2)))\sigma_2 = (\varphi^{-1})^*\mathbb{P}_{GQ}(\pi_{D_2*}^c(\mathfrak{X}_{\mathcal{F},c}^\sim(M_2; D_2)))\sigma_2.$$

Then Eq.(4) is proven by the following commutative diagram

$$\begin{array}{ccc} \mathcal{F}_{\mathcal{K}}(M_1, \mathbb{R}, D_1) & \xleftarrow{\tau} & \mathcal{F}_{\mathcal{K}}(M_2, \mathbb{R}, D_2) \\ \Xi_1 \downarrow & & \Xi_2 \downarrow \\ C^\infty(M_1/D_1, \mathbb{R}) & \xleftarrow{\tau_\varphi} & C^\infty(M_2/D_2, \mathbb{R}). \end{array}$$

That the mapping $\mathfrak{M}_\rho^{GQ}$ is bijective and equivalence classes preserving follows naturally from the fact that $\mathfrak{M}_\rho$ is bijective and equivalence classes preserving. The proof is therefore complete. $\square$

2.2 Unitary Representations of Canonical Transformations of Cotangent Bundles

In the preceeding work [8], it has been shown that the GQ-consistent kinematical algebra $\mathcal{S}_{GQ}(T^*Q; D)$ of the cotangent bundle (T^*Q, ω_c) with the vertical polarization D^v is partially Lie isomorphic to the kinematical algebra $\mathcal{S}(Q) = C^\infty(Q, \mathbb{R}) \oplus_s \mathfrak{X}_c(Q)$ of Q . This means therefore that (T^*Q, ω_c, D^v) admits elementary almost complete GQ-consistent quantum Borel kinematics. Furthermore, when Q is compact, then (T^*Q, ω_c, D^v) admits elementary complete GQ-consistent quantum Borel kinematics. Let ρ be a canonical transformation from (T^*Q, ω_c) into itself. We can build a real reducible polarization D^ρ from the vertical polarization D^v according to

$$D^\rho = \bigcup_{m \in T^*Q} \{\rho_*Z \in T_m(T^*Q) | Z \in D^v(\rho^{-1}(m))\}.$$

The transformation $\rho : (T^*Q, \omega_c, D^v) \rightarrow (T^*Q, \omega_c, D^\rho)$ is clearly a polarization preserving symplectomorphism. It has also been shown in [8] that the GQ-consistent kinematical algebra $\mathcal{S}_{GQ}(T^*Q; D^\rho)$ is almost complete and admits an almost complete GQ-consistent quantum Borel kinematics.

Let Φ be an integral closed 2-form on Q and $(\mathfrak{H}, \mathbb{Q}_{GQ}, \mathbb{P}_{GQ}, \theta^\infty)$ is an elementary differentiable Φ -compatible GQ-consistent quantum Borel kinematics for (T^*Q, ω, D^v) .

Theorem 2.2 *For every canonical transformation ρ of (T^*Q, ω) into itself The quantum Borel kinematics $(\mathfrak{H}, \mathbb{Q}_{GQ}, \mathbb{P}_{GQ}, \theta^\infty)$ and the quantum Borel kinematics*

$$\mathfrak{M}_\rho^{GQ}((\mathfrak{H}, \mathbb{Q}_{GQ}, \mathbb{P}_{GQ}, \theta^\infty)) = ((\varphi_\rho^{-1})^*\mathfrak{H}, \mathfrak{r}_{\varphi_\rho}^*\mathbb{Q}_{GQ}, (\varphi_\rho^{-1})^*\mathbb{P}_{GQ}, (\varphi_\rho^{-1})^*\theta^\infty)$$

for (T^*Q, ω, D^ρ) are unitarily related, i.e. there is a unitary mapping $\mathbb{U}_\rho : \mathfrak{H} \rightarrow (\varphi_\rho^{-1})^*\mathfrak{H}$ so that $\mathbb{U}_\rho \mathbb{Q}_{GQ}(\zeta) \mathbb{U}_\rho^{-1} = \mathfrak{r}_{\varphi_\rho}^* \mathbb{Q}_{GQ}(\zeta)$ and $\mathbb{U}_\rho \mathbb{P}_{GQ}(X) \mathbb{U}_\rho^{-1} = (\varphi_\rho^{-1})^* \mathbb{P}_{GQ}(X)$,

for every $\zeta \in C^\infty(Q, \mathbb{R})$ and every $X \in \mathfrak{X}_c(Q)$. The unitary mapping $\mathbb{U}_\rho$ is uniquely determined up to a scalar multiplier in the unit circle $\mathbb{T} \subset \mathbb{C}^*$.

Proof : For every $(p, q) := (p_1, \dots, p_n, q^1, \dots, q^n) \in T^*Q$ the point $\rho((p, q))$ can be written as $(\rho_j(p, q), \rho_{n+j}(p, q))$. Let $(0, q_a) := (0, \dots, 0, q_a^1, \dots, q_a^n)$ be a fixed point in T^*Q . When $B_{q_a} := \{(p_1, \dots, p_n, q_a) \in T^*Q\} \cong \mathbb{R}^n$ denotes the leaf of D^v containing the point $(0, q_a)$, then

$$\rho(B_{q_a}) = \{(\rho_j(p, q_a), \rho_{n+j}(p, q_a)) \in T^*Q \mid (p, q_a) \in B_{q_a}\} \cong \mathbb{R}^n$$

the leaf of D^ρ containing the point $(0, \rho_{n+j}(p'_1, \dots, p'_n, q_a))$, where $p'_1, \dots, p'_n$ are the solutions of

$$\begin{aligned} \rho_1(p'_1, \dots, p'_n, q_a) &= 0 \\ \vdots &\quad \quad \quad \vdots \\ \rho_n(p'_1, \dots, p'_n, q_a) &= 0. \end{aligned} \tag{5}$$

The existence and uniqueness of the solutions $p'_1, \dots, p'_n$ of Eq.(5) are clearly assured : Define a mapping $\mathbf{pr} : T^*Q \rightarrow \mathbb{R}^n$ through $\mathbf{pr} = \pi^p \circ \rho$, where π^p is a projection $T^*Q \rightarrow \mathbb{R}^n$ which is defined as $\pi^p(p, q) = p$ for every $(p, q) \in T^*Q$. $\mathbf{pr}$ is therefore as submersion. According to Theorem 1.38 of [9], $\mathbf{pr}^{-1}(0)$ is a submanifold of T^*Q . This means, there are uniquely $p'_1, \dots, p'_n$ so that $\rho_j((p'_1, \dots, p'_n, q_a)) = 0$ ($j = 1, \dots, n$). It is straightforward to show that $\mathbf{pr}^{-1}(0) = \rho^{-1}((0, Q)) \cong Q$ holds, i.e. $(p', q_a) = B_{q_a} \cap \rho^{-1}((0, Q))$.

Hence $(0, q_a) \mapsto (0, \rho_{n+j}(p'_1, \dots, p'_n, q_a))$ defines a smooth transformation φ_ρ of the zero sections of the cotangent bundle T^*Q . The differentiability of φ_ρ follows naturally from the differentiability of ρ . If Q is identified as the zero section of T^*Q , then we obtain the following commutative diagram

$$\begin{array}{ccc} (T^*Q, \omega, D^v) & \xrightarrow{\rho} & (T^*Q, \omega, D^\rho) \\ \pi_{D^v} \downarrow & & \downarrow \pi_{D^\rho} \\ Q & \xrightarrow{\varphi_\rho} & Q. \end{array}$$

Let τ_{φ_ρ} be defined as the inverse mapping of $\tau_{\varphi_\rho^{-1}}$ by

$$\tau_{\varphi_\rho} : \eta_x \ni z \mapsto (z, \varphi_\rho(x)) \in ((\varphi_\rho^{-1})^* \eta)_{\varphi_\rho(x)}.$$

It is clear that the mapping τ_{φ_ρ} is a mapping of line bundles with connections and τ_{φ_ρ} the lift of φ_ρ so that $\tau_{\varphi_\rho}^*(\tau_{\varphi_\rho^{-1}}^* \theta) = \theta$ holds.

For $c \in \mathbb{C}^*$, define μ^c as multiplication mapping of η through $\eta \ni z \mapsto \mu^c(z) = cz \in \eta$. According to Proposition 1.12.2 in [4] and its proof, we can choose uniquely a positive number a so that for arbitrary $\alpha \in \mathbb{T}$, $\bar{\tau}_{\varphi_\rho} := \tau_{\varphi_\rho} \circ \mu^{a\alpha}$ is an isometric bundles isomorphism with $\bar{\tau}_{\varphi_\rho}^*(\tau_{\varphi_\rho^{-1}}^* \theta) = \theta$. Therefore, according to remark 1.10.1. of [4], $(\varphi_\rho^{-1})^* \nabla = \bar{\tau}_{\varphi_\rho} \circ \nabla \circ \bar{\tau}_{\varphi_\rho}^{-1}$ holds. Furthermore, define a mapping $\mathbb{U}_\rho : \mathfrak{H} \rightarrow (\varphi_\rho^{-1})^* \mathfrak{H}$ as

$$\mathbb{U}_\rho \sigma = e^{-i\gamma \ln \varrho} \varrho^{1/2} \bar{\tau}_{\varphi_\rho} \sigma, \quad \forall \sigma \in \mathfrak{H},$$

where $\varrho : Q \rightarrow \mathbb{R}^+$ is the strict positive smooth function which satisfies $\nu = \varrho \cdot (\nu \circ \varphi_\rho^{-1})$. It must be shown that $\mathbb{U}_\rho$ is unitar with $\mathbb{U}_\rho^{-1}\sigma = \bar{\tau}_{\varphi_\rho}^{-1}(e^{i\gamma \ln \varrho} \varrho^{-1/2} \sigma)$. However, it proceeds straightforwardly (in spite of somewhat long computation) so that $\mathbb{U}_\rho \mathbb{Q}_{GQ}(\zeta) \mathbb{U}_\rho^{-1} = \mathfrak{r}_{\varphi_\rho}^* \mathbb{Q}_{GQ}(\zeta)$ and $\mathbb{U}_\rho \mathbb{P}_{GQ}(X) \mathbb{U}_\rho^{-1} = (\varphi_\rho^{-1})^* \mathbb{P}_{GQ}(X)$, for every $\zeta \in C^\infty(Q, \mathbb{R})$ and every $X \in \mathfrak{X}_c(Q)$. $\square$

Acknowledgement. The author would like to thank H.D. Doebner and J.D. Hennig for interesting discussion and to the Laboratory of Atomic and Nuclear Physics, Department of Physics, Gadjah Mada University, for wonderful atmosphere.

REFERENCES

1. B. ANGERMANN, *Über Quantisierungen lokalisierter Systeme - Physikalisch interpretierbare mathematische Modelle*, PhD thesis, Technische Universität Clausthal, 1983.
2. H.D. DOEBNER AND P. NATTERMANN, Borel Quantization : Kinematics and Dynamics, *Acta Physica Polonica B*, **27**(10), 2327–2340, 1996.
3. S. KOBAYASHI AND K. NOMIZU, *Foundation of Differential Geometry*, 1st volume, Interscience Publishers, New York, 14–15, 1963.
4. B. KOSTANT, Quantization and Unitary Representation, in C. T. Taam (editor), *Lectures in Modern Analysis III*, **170**, Springer-Verlag, Berlin, 87–208, 1970.
5. P. NATTERMANN, *Dynamics in Borel-Quantization : Non-linear Schrödinger Equations vs. Master Equations*, PhD thesis, Technische Universität Clausthal, 1997.
6. M. F. ROSYID, On the Relation between Configuration and Phase Space Quantization I, *Journal of the Indonesian Mathematical Society*, **9**(1), 2003.
7. M. F. ROSYID, On the Relation between Configuration and Phase Space Quantization II, presented also in this conference.
8. M.F. ROSYID, GQ-Consistent Quantum Borel Kinematics, in *Proceeding of International Conference on Applied Mathematics 2005 (ICAM05)*, Bandung, 2005.
9. F.W. WARNER, *Foundations of Differentiable Manifolds and Lie Groups*, Graduate Texts in Mathematics 94, Springer-Verlag, Berlin, 1983.
10. N.M.J. WOODHOUSE, *Geometric Quantization*, 2nd Edition, Clarendon Press, Oxford, 1992.

M.F. ROSYID: Department of Physics, Gadjah Mada University and Institut for Science in Yogyakarta, Yogyakarta, Indonesia.
E-mail: farchani@ugm.ac.id.

N-PLAYER LINEAR QUADRATIC DYNAMIC GAME FOR DESCRIPTOR SYSTEM

SALMAH

Abstract. Linear quadratic continuous non-zero-sum N player dynamic game with descriptor system will be defined. The assumption for the game is open-loop. Optimization problem in the Nash sense will be constructed. With basic approach, which is Hamiltonian, sufficiency for the existence of optimal solution in the Nash sense will be derived. Riccati type differential equation also will be considered. The connection of game problem solution and N couple Riccati equation will be studied. For infinite horizon case, non-recursive formulae for solution of algebraic Riccati differential equation will be derived.

1. Introduction

Linear quadratic dynamic game with N parties (it called players) is a problem that the players satisfy one state space system and try to minimize their individual quadratic objective function and give control to ordinary state space system. Open-loop game is based on assumption that the parties act non-cooperatively and the only information they have is it present state and the model structure. For non-cooperative game, combinations of pairs of strategies of players which are secure against any attempts by one player to unilaterally alter his strategy will be considered. This strategy is called Nash equilibrium of the game ([1], [3] and [4]).

Although it has many applications, ordinary linear quadratic optimal control, often does not provide a physical meaning in controlling, because the state variable does not corresponds with variable that we want to control. Descriptor systems have a great capacity for system modeling because they can preserve structure of physical system and can include nondynamic mode and impulsive mode. Therefore they have a potential applicability for a wide class of systems ([2], [5], [6], [7],[8] and [10]).

Linear quadratic dynamic game with descriptor system is a problem of players who try to minimize their individual quadratic performance criterion and give control to linear descriptor system. This paper will give necessary and sufficient condition for the existence of Nash equilibrium of linear quadratic game with descriptor with

Hamiltonian method. This paper will consider a coupled of generalized Riccati differential equation for linear quadratic game with finite horizon problem, extend result in [9], for game with two players.

In [9] to find Nash solution for game with infinite horizon case, it has been done directly solving generalized algebraic equation. In this paper particular consideration will be given to computational aspects of the problem. Non-recursive formulae to solve N couple algebraic Riccati equation for the infinite horizon problem will be derived. The formulae is based on a generalized eigenvector of a particular matrix. With the formulae, Nash solution of the game will can be easily found. Furthermore, numerical example with MATLAB program can be given by using the formulae.

2. DYNAMIC GAME DESCRIPTOR SYSTEM WITH N PLAYERS

Linear quadratic dynamic game with N players is defined as an open-loop game with the players give control vector to the system

$$E\dot{x} = Ax + B_1u_1 + B_2u_2 + \dots + B_Nu_N, \quad (1)$$

$$Ex(0) = Ex_0. \quad (2)$$

with $E \in \mathbb{R}^{n \times n}$, $A \in \mathbb{R}^{n \times n}$, $B_1 \in \mathbb{R}^{n \times m_1}$, $B_2 \in \mathbb{R}^{n \times m_2}$, ..., $B_N \in \mathbb{R}^{n \times m_N}$, $x(t)$ descriptor vector n dimension. Vector $u_i(t)$, $i=1, \dots, N$ are control vector m_i dimension which is done by

i -th player, $i=1, \dots, N$. Matrix E generally singular with rank $E = r < n$.

The players minimizing objective functions in the Nash sense of the form

$$J_i(u_1, u_2, \dots, u_N) = \frac{1}{2}x(T)^T E^T K_{iT} Ex(T) + \frac{1}{2} \int_0^T \left(x(t)^T Q_i x(t) + \sum_{j=1}^N u_j(t)^T R_{ij} u_j(t) \right) dt, \quad (3)$$

with all matrices symmetric. Furthermore Q_i and K_{iT} semi positive definite and R_{ij} positive definite. Assumption which is needed will be given.

Assumption 1: Descriptor system (1) regular, impulse controllable and finite dynamic stabilizable which satisfy

$$i) \quad |sE - A| \neq 0, \forall s \neq 0, \text{ except for a finite number of } s \in \mathbb{R},$$

$$ii) \quad \text{Im } E + \text{Im } A(\ker E) + \text{Im}(B_1 | B_2 | \dots | B_N) = \mathbb{R}^n,$$

$$iii) \quad \text{rank}(sE - A \quad B_1 | B_2 | \dots | B_N) = n \quad \forall s, \text{Re}[s] \geq 0.$$

To derive necessary condition of optimal Nash solution we need the Hamiltonian functions as follow.

$$\begin{aligned}
H_1(t) &= \frac{1}{2} \left[x^T Q_1 x + u_1^T R_{11} u_1 + u_2^T R_{12} u_2 + \dots + u_N^T R_{1N} u_N \right] + \gamma_1^T (Ax + B_1 u_1 + B_2 u_2 + \dots + B_N u_N), \\
H_2(t) &= \frac{1}{2} \left[x^T Q_2 x + u_1^T R_{21} u_1 + u_2^T R_{22} u_2 + \dots + u_N^T R_{2N} u_N \right] + \gamma_2^T (Ax + B_1 u_1 + B_2 u_2 + \dots + B_N u_N), \\
&\vdots \\
H_N(t) &= \frac{1}{2} \left[x^T Q_N x + u_1^T R_{N1} u_1 + u_2^T R_{N2} u_2 + \dots + u_N^T R_{NN} u_N \right] + \gamma_N^T (Ax + B_1 u_1 + B_2 u_2 + \dots + B_N u_N).
\end{aligned}$$

With Lagrange multiplier method as in [9] we get necessary conditions for objective function to be optimal in the Nash sense are

$$E^T \frac{\partial \gamma_i}{\partial t} = -\frac{\partial H_i}{\partial x}, \quad \frac{\partial H_i}{\partial u} = 0, \quad E\dot{x} = \frac{\partial H_i}{\partial \gamma_i}, \quad i=1,2,\dots,N.$$

Substitute these equations to (1) yields $E^T \dot{\gamma}_i(t) = -Q_i x(t) - A^T \gamma_i(t)$, and $R_{ii} u_i + B_i^T \gamma_i = 0$ with $i=1,2,\dots,N$. The boundary conditions are $Ex(0) = Ex_0$ and $E^T \gamma_i(T) = E^T K_{iT} Ex(T)$. From necessary condition of optimal Nash solution we get optimal strategies for n players dynamic game are $u_i = -R_{ii}^{-1} B_i^T \gamma_i(t)$, with $\gamma_i(t)$, satisfy $E^T \dot{\gamma}_i(t) = -Q_i x(t) - A^T \gamma_i(t)$, $i=1,2,\dots,N$, and boundary conditions $E^T \gamma_i(T) = E^T K_{iT} Ex(T)$, $i=1,2,\dots,N$

We can write in the matrix form and get

$$\begin{pmatrix} E & 0 & \dots & \dots & 0 \\ 0 & E^T & \dots & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & \dots & E^T \end{pmatrix} \begin{pmatrix} \dot{x}(t) \\ \dot{\gamma}_1(t) \\ \vdots \\ \dot{\gamma}_N(t) \end{pmatrix} = \begin{pmatrix} A & -B_1 R_{11}^{-1} B_1^T & -B_2 R_{22}^{-1} B_2^T & \dots & -B_N R_{NN}^{-1} B_N^T \\ -Q_1 & -A^T & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -Q_N & 0 & 0 & \dots & -A^T \end{pmatrix} \begin{pmatrix} x(t) \\ \gamma_1(t) \\ \vdots \\ \gamma_N(t) \end{pmatrix}, \quad (4)$$

with boundary conditions $Ex(0) = Ex_0$ and $E^T \dot{\gamma}_i(T) = E^T K_{iT} Ex(T)$. System (4) can be written in descriptor form as $\tilde{E}\dot{\tilde{x}} = \tilde{A}\tilde{x}$. If $(s\tilde{E} - \tilde{A})$ regular, system (4) will have solution ([9]). We need the following assumptions for equation (4).

Assumption 2: Descriptor system (1) is regular and impulse free i.e

$$\text{Im } \tilde{E} + \text{Im } \tilde{A}(\ker \tilde{E}) = \mathfrak{R}^{2n+m}.$$

If Assumption 2 is satisfied then system (4) will be regular and impulse controllable.

For N players linear quadratic dynamic game define N generalized differential Riccati equation as follow

$$E^T \dot{K}_i + A^T K_i + L_i A + Q_i - L_i \sum_{j=1}^N B_j R_{jj}^{-1} B_j^T K_i = 0, \quad i = 1, \dots, N. \quad (5)$$

with $L_i E = E^T K_i$, $i = 1, 2, \dots, N$.

The following theorem concern with relationship between the existence of dynamic game solution and generalized Riccati differential equation (5).

Theorem 1: *If generalized differential Riccati equation (5) has solution, then system (1) will have solution.*

For N players infinite time linear quadratic dynamic game the players satisfy system (1). Objective function to be minimized are in the form

$$J_i(u_1, u_2, \dots, u_N) = \frac{1}{2} \int_0^\infty \left(x^T(t) Q_i x(t) + \sum_{j=1}^N u_j^T(t) R_{ij} u_j(t) \right) dt, \quad i=1, 2, \dots, N, \quad (6)$$

with all matrices symmetric. Furthermore Q_i and K_{i_0} semi positive definite and R_{ij} positive definite.

Generalized algebraic Riccati equation for N players infinite time problem are

$$A^T K_i + L_i A + Q_i - L_i \sum_{j=1}^N B_j R_{jj}^{-1} B_j^T K_i = 0, \quad i = 1, \dots, N, \quad (7)$$

with $L_i E = E^T K_i$, $i = 1, 2, \dots, N$.

It can be proved that Theorem 1 will also be satisfied for optimal control for infinite time problem, therefore we get the optimal Nash have form $u_i(t) = -R_{ii}^{-1} B_i^T K_i x(t)$, with K_i are constant matrices, solution of (7).

3. GENERALIZED EIGENVALUE PROBLEM

To get numerical formulae of solution of algebraic Riccati equation we need to consider the generalized eigenvalue problem. Let now consider equation (4). We will study the generalized eigenvalue problem

$$\tilde{A}Z = \lambda \tilde{E}Z. \quad (8)$$

Let U is the generalized eigen vectors and generalized principal vector corresponding to stable eigenvalues eigenvalues. For the generalized eigenvalue problem we will have

$$\tilde{A}U = \tilde{E}UW,$$

Partition U as $[U_1 \ U_2 \ \cdots \ U_{N+1} \ U_{N+2} \ \cdots \ U_{2N+1}]$ we can write

$$\begin{pmatrix} A & 0 & \cdots & 0 & B_1 & \cdots & B_N \\ -Q_1 & -A^T & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ -Q_N & 0 & \cdots & -A^T & 0 & \cdots & 0 \\ 0 & B_1^T & \cdots & 0 & R_{11} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & B_N^T & 0 & \cdots & R_{NN} \end{pmatrix} \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_{N+1} \\ U_{N+2} \\ \vdots \\ U_{2N+1} \end{pmatrix} = \begin{pmatrix} E & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & E^T & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & E^T & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_{N+1} \\ U_{N+2} \\ \vdots \\ U_{2N+1} \end{pmatrix} W$$

Therefore we will have the equations

$$AU_1 + B_1U_{N+2} + \cdots + B_NU_{2N+2} = EU_1W \quad (9)$$

$$-Q_1U_1 - A^TU_2 = E^TU_2W,$$

$$-Q_2U_1 - A^TU_3 = E^TU_3W,$$

$\vdots$

$$-Q_NU_1 - A^TU_{N+1} = E^TU_{N+1}W. \quad (10)$$

$$B_1^TU_2 + R_{11}U_{N+2} = 0,$$

$$B_2^TU_3 + R_{22}U_{N+3} = 0.$$

$\vdots$

$$B_N^TU_{N+1} + R_{22}U_{2N+1} = 0. \quad (11)$$

If $\text{rank}E = r \leq n$ then E will has $(n-r)$ zero eigenvalues. Let V_1 become the eigenvectors corresponding to the zero eigenvalues. Then we have

$$EV_1 = 0. \quad (12)$$

Let V_2 be a matrix of the eigenvectors corresponding to the $(n-r)$ zero eigenvectors of E^T . Then we have

$$E^TV_2 = 0. \quad (13)$$

For existence of solution of Riccati equation, we need the following Assumption.

Assumption 3: Assume that $\begin{bmatrix} U_1 & V_1 \end{bmatrix}$ is nonsingular.

If Assumption 1, 2 and 3 are satisfied we get the following Lemma.

Lemma 1: For matrix K satisfying $KU_1 = U_3$, we will have that $(E, A + B_1K_1 + B_2K_2)$ is admissible if and only if $(EU_1 \quad (A + B_1K_1 + B_2K_2)V_1)$ is non-singular.

4. OPTIMALITY FOR FINITE TIME AND INFINITE TIME PROBLEM

In this section the value of cost function for optimal solution will be considered. This is the combination of research which is done by [3] and [5].

Now we will discussed about optimality theorem for finite time problem.

Let assumptions 1, 2 and 3 are satisfied, it can be derived that there exist solutions for the following n differential equations.

$$\begin{aligned}
 & E^T \dot{Z}_{01}(t) + Z_{01}^T(t)A + A^T Z_{01}^T(t) + Q_1 - Z_{01}^T(t)B_1R_{11}^{-1}B_1^T Z_{01}(t) - Z_{01}^T(t)B_2R_{22}^{-1}B_2^T Z_{01}(t) - \\
 & \cdots - Z_{01}^T(t)B_NR_{NN}^{-1}B_N^T Z_{01}(t) = 0, \\
 & E^T \dot{Z}_{02}(t) + Z_{02}^T(t)A + A^T Z_{02}^T(t) + Q_2 - Z_{02}^T(t)B_1R_{11}^{-1}B_1^T Z_{02}(t) - Z_{02}^T(t)B_2R_{22}^{-1}B_2^T Z_{02}(t) - \\
 & \cdots - Z_{02}^T(t)B_NR_{NN}^{-1}B_N^T Z_{02}(t) = 0 \\
 & \vdots \\
 & E^T \dot{Z}_{0N}(t) + Z_{0N}^T(t)A + A^T Z_{0N}^T(t) + Q_N - Z_{0N}^T(t)B_1R_{11}^{-1}B_1^T Z_{0N}(t) - Z_{0N}^T(t)B_2R_{22}^{-1}B_2^T Z_{0N}(t) - \\
 & \cdots - Z_{0N}^T(t)B_NR_{NN}^{-1}B_N^T Z_{0N}(t) = 0
 \end{aligned} \tag{12}$$

where

$$Z_{0i}^T(t)E = E^T Z_{0i}(t), \quad i=1,2,\dots,N \tag{13}$$

with final conditions

$$E^T Z_{0i}(t_a) = E^T K_{it_a} E \quad i=1,2,\dots,N \tag{14}$$

Optimality theorem for finite time problem will be discussed.

Theorem 2. Suppose that $T < \infty$ in (3). If Assumptions 1,2,3 and 4 are satisfied, then there exist solutions for generalized differential Riccati equations (5), with boundary conditions

$$E^T K_i(T) = E^T K_{iT} E, \quad i=1,2,\dots,N \tag{15}$$

Optimal solutions are in the form

$$u_i = -R_{ii}^{-1} B_i^T K_i(t) x(t), \quad i=1,2,\dots,N \quad (16)$$

Furthermore optimal values for J_1, J_2 are

$$J_{1\min} = \frac{1}{2} x_0^T E^T K_1(0) x_0, \quad J_{2\min} = \frac{1}{2} x_0^T E^T K_2(0) x_0. \quad \dots$$

$$J_{N\min} = \frac{1}{2} x_0^T E^T K_N(0) x_0 \quad (17)$$

Now optimal solutions for open-loop two players linear quadratic of dynamic game descriptor system for T tends to infinity will be considered. Then the value of the cost functions for optimal solutions will be considered.

In [5] it have been proved that if assumptions 1,2, and 3 are satisfied, for $T \rightarrow \infty$, the equation $E^T X(t) = \pi(t, t_a)$ convergent to constant matrix. Optimality theorem for finite time problem is given below.

Theorem 3. *If assumptions 1,2, and 3 are satisfied then there exist solution for generalized algebra Riccati equations (7). Control optimal solutions are given by*

$$u_i = -R_{ii}^{-1} B_i^T K_i(t) x(t), i=1,2,\dots,N \quad (18)$$

Furthermore optimal values for J_1, J_2 are given by

$$J_{1\min} = \frac{1}{2} x_0^T E^T K_1 x_0, \\ J_{2\min} = \frac{1}{2} x_0^T E^T K_2 x_0, \dots, J_{N\min} = \frac{1}{2} x_0^T E^T K_N x_0.$$

Moreover numerical solutions of K_1 and K_2 are given by

$$\begin{pmatrix} K_1 \\ K_2 \\ \vdots \\ K_N \end{pmatrix} = \begin{pmatrix} U_2 \\ U_3 \\ \vdots \\ U_{N+1} \end{pmatrix} V_2 \begin{pmatrix} H_1 \\ h_2 H_1 \\ \vdots \\ h_{N+1} H_1 \end{pmatrix} (U_1 \quad V_1)^{-1}. \quad (19)$$

with $H_1, h_2, \dots, h_{N+1} \in \mathbb{R}^n$.

5. NUMERICAL EXAMPLE

Here we will find solution of ARE with the non-recursive formulae (19) that has been derived in Theorem 3 to system with matrix A is 3 x 3 with three players. Consider system that satisfy

$$E\dot{x} = Ax + B_1u_1 + B_2u_2 + B_3u_3$$

With $E = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ and $A = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 2 & 0 \\ 1 & 0 & 0 \end{pmatrix},$

$$B_1 = \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}, B_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, B_3 = \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}.$$

For the cost function, given

$$Q_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, Q_2 = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{pmatrix}, Q_3 = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$R_{11} = (1), R_{22} = (2), R_{33} = (3).$$

We get that gain optimal control is

$$K_1 = \begin{pmatrix} 1.0097 & -0.4272 & 0 \\ -0.1238 & 1.1596 & 0 \\ 3.1796 & -1.8178 & 3 \end{pmatrix}, K_2 = \begin{pmatrix} 3.0291 & -1.2816 & 0 \\ -0.1238 & 1.1596 & 0 \\ 5.4903 & -0.3174 & 4 \end{pmatrix},$$

$$K_3 = \begin{pmatrix} 1.0097 & -0.4272 & 0 \\ -0.3713 & 3.4787 & 0 \\ 4.1990 & -1.6723 & 5 \end{pmatrix}.$$

Substitute the gain control optimal to system we get the controlled system is

$$E\dot{x} = A_{cl}x$$

with

$$A_{cl} = \begin{pmatrix} -0.0890 & -2.1767 & 1 \\ 3.1253 & -4.3308 & 3 \\ -5.0485 & 2.1361 & -5 \end{pmatrix}.$$

We get closed loop response in the figure below.

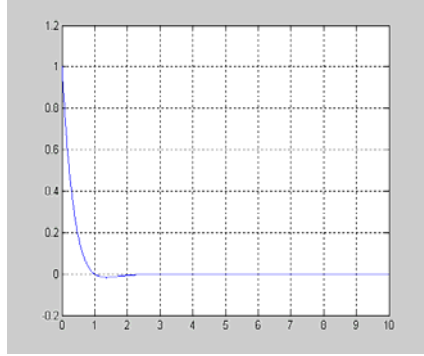


Fig. 1. Closed loop response for x_1

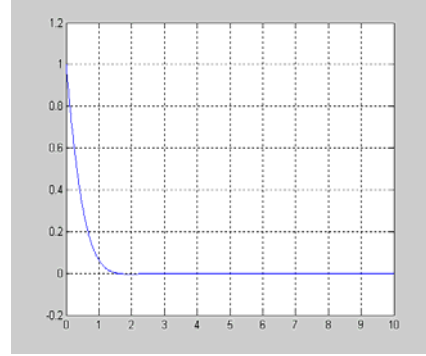


Fig. 2. Closed loop response for x_2

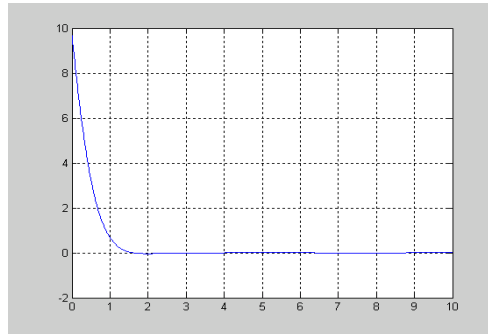


Fig. 3. Closed loop response for x_3

Figure 1, figure 2 and figure 3 give the closed loop response of the system in the numerical example. Furthermore the figure show that the closed loop system is stable, as what was desired. Theorem that ensure stability of the closed loop system with the method given need further consideration.

6. CONCLUSION

This paper consider N player non-zero-sum linear quadratic dynamic game with descriptor systems for finite horizon and infinite horizon case. Necessary condition for the existence of a Nash equilibrium have been derived with Hamiltonian method. The paper also consider N couple Riccati-type differential equation for finite horizon case

and algebraic Riccati equation for infinite horizon case. Moreover non-recursive formulae for solution of algebraic equation have been given a special attention. A numerical example and closed loop response of the system show that the formulae doing well.

REFERENCES

1. Basar, T., and Olsder, G.J., 1995, *Dynamic Noncooperative Game Theory*, second Edition, Academic Press, London, San Diego.
2. Dai, L., 1989, *Singular Control Systems*, Springer Verlag, Berlin.
3. Engwerda, J., 1998, "On the Open-loop Nash Equilibrium in LQ-games", *Journal of Economic Dynamics and Control*, Vol. 22, 729-762.
4. Engwerda, J.C. (2005). *LQ Dynamic Optimization and Differential Games*. Chichester: John Wiley & Sons.
5. Katayama, T., and Minamino, K., 1992, "Linear Quadratic Regular and Spectral Factorization for Continuous Time Descriptor Systems", *Proceedings of the 31st Conference on decision and Control*, Tucson, Arizona, 967-972.
6. Lewis, F.L., 1986, "A survey of Linear Singular Systems", *Circuits System Signal Process*, vol.5, no.1, 3-36.
7. Minamino, K., 1992, "The Linear Quadratic Optimal Regulator and Spectral Factorization for Descriptor Systems", *Master Thesis*, Department of Applied Mathematics and Physics, Faculty of Engineering, Kyoto University.
8. Mukaidani, H. and Xu, H., 2004a, "Nash Strategies for Large Scale Interconnected Systems", *43rd IEEE Conference on Decision and Control*, Bahamas, 4862-4867.
9. Salmah, Bambang, S., Nababan, S.M., and Wahyuni, S., 2002c, "Non-Zero-Sum Linear Quadratic Dynamic Game with Descriptor Systems", *Proceeding Asian Control Conference*, Singapore, 1602-1607.
10. Xu, H., and Mizukami, K., 1994, "Linear Quadratic Zero-sum Differential Games for Generalized State Space Systems", *IEEE Transactions on Automatic Control*, Vol. 39, No.1, January, 143-147.

Salmah: Department of Mathematics, Gadjah Mada University, Yogyakarta, Indonesia.

E-mail: syalmah@yahoo.com

TRACKING OPTIMAL CONTROL FOR FLEXIBLE SYSTEM

ROBERD SARAGIH

Abstract. This paper is concerned with the designing an optimal control law that forces the position of flexible system to track desired trajectory over a specified time interval. The motion of system is governed by partial differential equation, hence, to analyze dynamic and control system requires the solution of the eigenvalue problem which consists of an infinite set of eigenvalue and corresponding eigenfunction. It means that mechanically flexible systems require an infinite number of elastic modes to completely describe their behaviors. One of the fundamental problems in feedback control of flexible system is to precisely control a large dimensional system with a much smaller dimensional controller because on-board computer limitations and modeling error would restrict the control system only to a few critical modes. For designing control system, we first transform the model of flexible system to a state space model and by using the calculus variation, the control law for the tracking problem can be obtained. To verify the performance of the control system, the simulation is carried out to single-link flexible arm.

Key words and Phrases :Tracking optimal control, flexible system.

1. INTRODUCTION

The motion of flexible system is described by variables depending not only on time but also on space. As a result, the motion is governed by partial differential partial equations. The main problem for designing control for this system arises in computing the control gain, as implies infinite-dimensional gain matrices. This paper is concerned with the designing an tracking optimal control law that forces the position of flexible system to track desired trajectory over a specified time interval. Since the motion of system is governed by partial differential equation, to analyze dynamic and control system requires the solution of the eigenvalue problem which consists of an infinite set of eigenvalue and corresponding eigenfunction. It means that mechanically flexible systems require an infinite number of elastic modes to completely describe their behaviors. One of the fundamental problems in feedback control of flexible system is to precisely control a large dimensional system with a much smaller dimensional controller because on-board computer limitations and modeling error would restrict the control system only to a few critical modes. Some authors have proposed designing a active and feedback control for flexible system [1,2]. For designing control system, we first transform the model of flexible system to a state space model and by using the calculus variation, the control law for the tracking problem can be obtained. To verify the performance of the control system, the simulation is carried out. The control system is applied to the single-link flexible arm.

2. FLEXIBLE SYSTEM MODEL

The structures with general viscous damping described by partial differential equation:

$$\hbar w(x,t) + \tilde{\lambda} \frac{\partial w(x,t)}{\partial t} + m(x) \frac{\partial^2 w(x,t)}{\partial t^2} = f(x,t), 0 < x < L, \quad (1)$$

where $w(x,t)$ is the displacement of a typical point x inside the domain of the structure and at time t , $\hbar$ is a homogeneous differential stiffness operator of order $2p$, $\tilde{\lambda}$ is a homogeneous differential damping operator, $m(x)$ is the mass density and $f(x,t)$ is a distributed control force. The solution $w(x,t)$ of Eq. (1) is subject to the boundary conditions

$$B_i w(x,t) = 0, x = 0, L, i = 1, 2, \dots, p, \quad (2)$$

where B_i are differential operators of maximum order $2p-1$.

We assume an approximate solution of the Eq. (1) in the form of the finite series

$$w(x,t) = \sum_{k=1}^n \phi_k(x) q_k(t), \quad (3)$$

where $\phi_k(x)$ are comparison functions and $q_k(t)$ are generalized coordinates. Introducing the Eq. (3) into the Eq. (1), multiplying both sides of the resulting equation by $\phi_j(x)$ and integrating over the domain, we obtain

$$\sum_{k=1}^n m_{jk} \frac{d^2 q_k(t)}{dt^2} + \sum_{k=1}^n c_{jk} \frac{dq_k(t)}{dt} + \sum_{k=1}^n \kappa_{jk} q_k(t) = Q_j(t), j = 1, 2, \dots, n \quad (4)$$

where

$$m_{jk} = \int_0^L m \phi_j \phi_k dx = m_{kj}, j, k = 1, 2, \dots, n$$

$$\kappa_{jk} = \int_0^L \phi_j \hbar \phi_k dx = \kappa_{kj}, j, k = 1, 2, \dots, n$$

are symmetric mass and stiffness coefficients, respectively and

$$F_j(t) = \int_0^L \phi_j(x) f(x,t) dx, j = 1, 2, \dots, n$$

are generalized forces.

The equation (4) can be written in the matrix form

$$\bar{M} \frac{d^2 q(t)}{dt^2} + \bar{C} \frac{dq(t)}{dt} + \bar{K} q(t) = F(t), \quad (5)$$

where $\bar{M} = [m_{jk}] = \bar{M}^T$, $\bar{K} = [k_{jk}] = \bar{K}^T$, $q(t) = [q_1(t) \ q_2(t) \ \cdots \ q_n(t)]^T$, $F(t) = [F_1(t) \ F_2(t) \ \cdots \ F_n(t)]^T$.

By introducing the state vector $z(t) = \left[q^T(t) \ \frac{dq^T(t)}{dt} \right]^T$, the Eq. (5) can be written in the state space form

$$\begin{aligned} \frac{dz(t)}{dt} &= Az(t) + Bu(t) \\ y(t) &= Cz(t) \end{aligned} \quad (6)$$

where

$$A = \begin{bmatrix} 0 & I \\ -\bar{M}^{-1}\bar{K} & -\bar{M}^{-1}\bar{C} \end{bmatrix}, B = \begin{bmatrix} 0 \\ \bar{M}^{-1} \end{bmatrix}, u(t) = F(t), D = [I \ 0].$$

3. TRACKING OPTIMAL CONTROL

In this section we derive an optimal control law based on variation calculus that forces the position of flexible system to track desired trajectory over a specified time interval. The optimal control problem can be stated as follows:

Given $r(t)$ as a reference track, find $u(t)$, $t \in [0, T]$ in the Eq. (6) such that the quadratic cost index

$$J(0) = \frac{1}{2} (Cz(T) - r(T))^T P (Cz(T) - r(T)) + \frac{1}{2} \int_0^T [(Cz - r)^T Q (Cz - r) + uRu] dt \quad (7)$$

with $P \geq 0$, $Q \geq 0$, $R > 0$ is minimum.

Note that the differential equation of the flexible system acts as a constraint that we must adjoin using a Lagrange multiplier $p(t)$, as before:

$$J_a = h(z(T), T) + \int_0^T \left[g(z(t), u(t), t) + p^T(t) \left\{ a(z(t), u(t), t) \right\} - \frac{dz(t)}{dt} \right] dt \quad (8)$$

where $h(z(T), T) = \frac{1}{2} (Dz(T) - r(T))^T P (Dz(T) - r(T))$, and

$$a(z(t), u(t), t) = Az(t) + Bu(t).$$

The variation of J_a is

$$\begin{aligned} \delta J_a = & \frac{\partial h}{\partial z} \Big|_{z=z(T)} \delta z(T) + \frac{\partial h}{\partial t} \Big|_{t=T} \delta T \\ & + \int_0^T \left[\frac{\partial g}{\partial z} \delta z + \frac{\partial g}{\partial u} \delta u + \left(a - \frac{dz}{dt} \right)^T \delta p + p^T \left\{ \frac{\partial a}{\partial z} \delta z + \frac{\partial a}{\partial u} \delta u - \delta \left(\frac{dz}{dt} \right) \right\} \right] dt + \left[g + p^T \left(a - \frac{dz}{dt} \right) \right] \Big|_{t=T} \delta T \end{aligned} \quad (9)$$

By defining the Hamiltonian

$$H(z, u, p, t) = g(z(t), u(t), t) + p^T(t) a(z(t), u(t), t), \quad (10)$$

then

$$\begin{aligned} \delta J_a = & \frac{\partial h}{\partial z} \Big|_{z=z(T)} \delta z(T) + \left[\frac{\partial h}{\partial t} + g + p^T \left(a - \frac{dz}{dt} \right) \right] \Big|_{t=T} \delta T \\ & + \int_0^T \left[\frac{\partial H}{\partial z} \delta z + \frac{\partial H}{\partial u} \delta u + \left(a - \frac{dz}{dt} \right)^T \delta p - p^T \delta \left(\frac{dz}{dt} \right) \right] dt \end{aligned} \quad (11)$$

Note that :

$$-\int_0^T p^T(t) \delta \left(\frac{dz}{dt} \right) dt = -p^T(T) \left(\delta z(T) - \frac{dz(T)}{dt} \delta T \right) + \int_0^T \frac{dp^T(t)}{dt} \delta z dt. \quad (12)$$

So we now can rewrite the variation as

$$\begin{aligned} \delta J_a = & \left(\frac{\partial h}{\partial z} \Big|_{z=z(T)} - p^T(T) \right) \delta z(T) + \left[\frac{\partial h}{\partial t} + g + p^T \left(a - \frac{dz}{dt} \right) + p^T \frac{dz}{dt} \right] \Big|_{t=T} \delta T \\ & + \int_0^T \left[\left(\frac{\partial H}{\partial z} + \frac{dp^T}{dt} \right) \delta z + \frac{\partial H}{\partial u} \delta u + \left(a - \frac{dz}{dt} \right)^T \delta p \right] dt \end{aligned} \quad (13)$$

The necessary conditions for the variation to be zero are that for $t \in [0, T]$ are given by

$$\frac{dz(t)}{dt} = a(z, u, t), \quad \frac{dp(t)}{dt} = -\frac{\partial H^T}{\partial z}, \quad \text{and} \quad \frac{\partial H}{\partial u} = 0. \quad (14)$$

Based on the Eq. (14), we can obtain the input control $u(t)$.

3. SIMULATION RESULTS

The optimal control in the Eq. (14) is applied to the single-link flexible arm. Theoretically, the modeling of a slewing structure is similar to modeling a flexible single-link manipulator moving in the horizontal plane. The slewing motion is assumed to be a rotation of a structure about a point or axis. A single-link flexible arm is connected with a rigid rotation shaft, and a torque $u(t)$ is applied as

shown in Fig. 1. The properties of EI , and M are constant over the beam's length L .

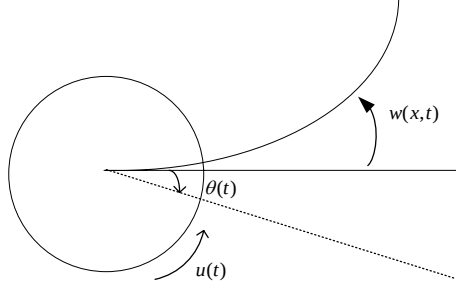


Fig.1 Single-link Flexible Arm

The equation of motion for the flexible arm rotating about an axis [3] is

$$M \left(x + R \right) \frac{d^2 \theta(t)}{dt^2} = - \left\{ M \frac{\partial^2 w(x,t)}{\partial t^2} + EI \frac{\partial^4 w(x,t)}{\partial x^4} + C_e I \frac{\partial}{\partial t} \left(\frac{\partial^4 w}{\partial x^4} \right) \right\}, \quad (15)$$

$$J \frac{d^2 \theta(t)}{dt^2} + \int_0^L \left\{ M(x+R)^2 \frac{d^2 \theta(t)}{dt^2} + M(x+R) \frac{\partial^2 w(x,t)}{\partial t^2} \right\} dx = u(t)$$

where R : radius of disk, EI : flexural rigidity, J : equivalent rotatory inertia, $w(x,t)$: deflection of beam with respect to x , $\theta(t)$: angular displacement of the rigid body motion, M : mass of beam, L : beam length, $u(t)$: torque applied to the beam. And the following boundary conditions

$$w(0,t) = \frac{\partial w(0,t)}{\partial x} = 0, \quad \frac{\partial^2 w(L,t)}{\partial x^2} = \frac{\partial^3 w(L,t)}{\partial x^3} = 0.$$

The state space equation for flexible arm as given in Eq. (6) with each entry is given by

$$\bar{M} = \begin{bmatrix} J & H_e^T \\ H_e & M_e \end{bmatrix}, \quad M_e = I, \quad H_e = \int_0^L M(x+R) \phi^T(x) dx, \quad \bar{C} = \begin{bmatrix} 0 & 0 \\ 0 & C_e \end{bmatrix}, \quad \bar{K} = \begin{bmatrix} 0 & 0 \\ 0 & K_e \end{bmatrix}$$

$$C_e = \text{diag} [2\xi_1 \omega_1 \quad 2\xi_2 \omega_2 \quad \cdots \quad 2\xi_n \omega_n], \quad K_e = \text{diag} [\omega_1^2 \quad \omega_2^2 \quad \cdots \quad \omega_n^2],$$

where ξ_i and ω_i are the frequencies of oscillation and the damping ratio, respectively.

In the simulation, the parameter of a single-link flexible beam are as follows:
 $L = 9.12 \times 10^{-1} m$, $M = 1.55 \times 10^{-1} kg/m$, $EI = 2.058 \times 10^{11} N/m$,

$$J = 1.05 \times 10^{-3} \text{ kgm}^2, R = 1.0 \times 10^{-1} \text{ m}.$$

From the simulation results, we obtain that the position of flexible arm can follow the desired trajectory only in 0.5 second as given in the following figures (Fig.2 and Fig. 3). The other result is, the vibration can be reduced by using the obtained control system as given in Fig.4.

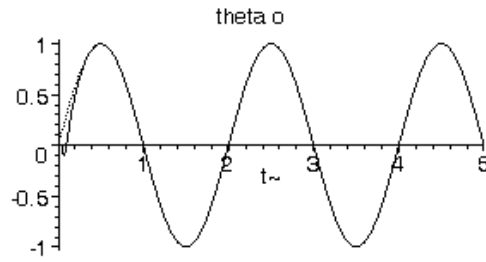


Figure 2: The desired trajectory is $\sin \pi t$

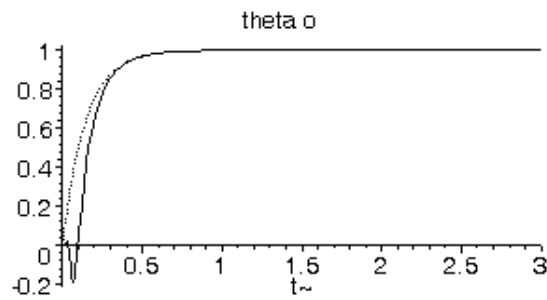


Figure 3: The desired trajectory is $1 - e^{-7t}$

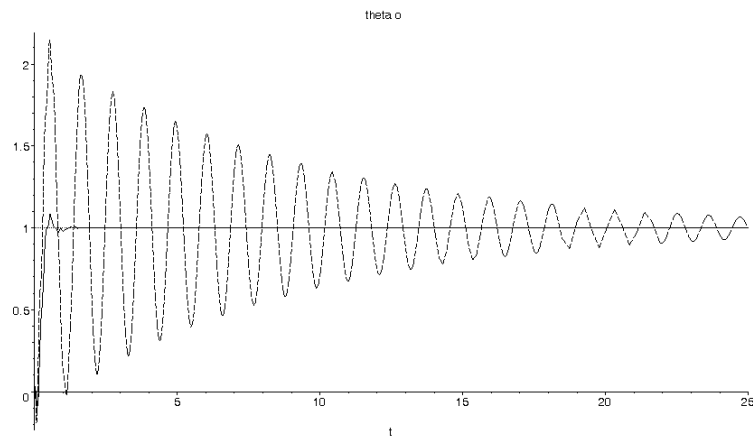


Figure 4. Performance of the control system (---: No controller; ____ controlled)

4. CONCLUSIONS

In this paper, the tracking optimal control for flexible system is derived by using the concept of the calculus variation. The obtained optimal control has applied to flexible arm to force the position of flexible system to track desired trajectory over a specified time interval. From the simulations, the position of the flexible system can follow the reference trajectory in 0.5 second. Furthermore, the control system can suppress the vibration of flexible system.

Acknowledgement. The research of this paper was supported by Research Program of ITB, and this is gratefully acknowledged.

REFERENCES

1. M.J. Balas, Active Control of Flexible System, *Journal of Optimization Theory and Applications*, Vol. 25, No. 3, 415-436, 1978.
2. M.J. Balas, Feedback Control of Flexible Systems, *IEEE Transactions on Automatic Control*, Vol. AC-23, No. 4, 673-679, 1979.
3. Roberd Saragih and Kazuo Yoshida, Order Reduction Method of Controller based on upon Modal Truncation for Flexible Structure, *Proceeding of the Third International Conference on Motion and Vibration Control*, Vol. 1, 30-35, 1996.

Roberd Saragih: Industrial and Financial Mathematics Group, Faculty of Mathematics and Natural Science, Institut Teknologi Bandung, Bandung, Indonesia.
E-mail: roberd@math.itb.ac.id

PERFORMANCE OF GENETIC ALGORITHM FOR PERMUTATION FLOWSHOP SCHEDULING PROBLEM

D. R. SILABAN, D. KUSHARYADI, Y. SATRIA

Abstract. The Flowshop Scheduling Problem (FSP) is a scheduling problem in which n jobs have to be processed on m machines. The problem is to find a sequence of jobs for each machine that minimize completion time (makespan). Our interest is solving a special case of FSP called Permutation Flow Shop Scheduling Problem (PFSP) in which the ordering sequence of jobs on all machines is identical. We solve PFSP using genetic algorithm (GA). GA is a heuristic algorithm that adopts evolutionary principle in biology to computer science. There are three basic operators in genetic algorithm: selection, crossover, and mutation. We use roulette and tournament selection, similar job order crossover, and shift mutation. To improve the performance of the GA, we apply two types of scheme those are generational scheme and restart scheme. In this paper, we test the performance of GA in solving PFSP. The algorithm performance is based on how close the approximation solution generated by GA to the best-known solution of the test problems. We use test problems from OR-library. Our simulation shows that the performance of GA in solving permutation flowshop scheduling problem is quite good with the relative error less than 0.4 %.

Key words and Phrases: flowshop scheduling problem, permutation flowshop scheduling problem, genetic algorithm, generational scheme, restart scheme.

1. INTRODUCTION

The Flowshop Scheduling Problem (FSP) is a problem of finding a sequence of a finite number of jobs that have to be processed on a finite number of machines. The objective is to minimize the completion time or makespan of the schedule. Permutation Flow Shop Scheduling Problem (PFSP) is a special case of FSP in which the jobs have identical ordering sequences on all machines.

PFSP is a combinatorial optimization. Exact method such as linear programming works only for small size of this problem [9]. For large problem, the solution can be approximated by heuristic approach. In this paper, we use one of heuristic algorithm that is a genetic algorithm (GA) to solve PFSP.

GA is a heuristic algorithm based on the theory of evolution by Charles Darwin [6]. A solution is represented by a string of values known as chromosome or genome. At first, an initial population made of a group of chromosomes is generated. Then, each chromosome is evaluated by fitness function to know how good the chromosome is. Afterwards, a new population

is generated using genetics operators until a certain stopping criterion is reached.

There are three main operators in GA: selection, crossover, and mutation. The selection operator is used to choose parent chromosomes. The crossover operator will create a child chromosome from the parent. The child is expected to be better than the parents. Mutation operator is used to maintain the variation of the chromosomes in the population. In this paper, we use roulette wheel and tournament selection, similar job order crossover (SJOX), and shift mutation as the operators. To improve the performance of the algorithm, two schemes are added those are generational and replacement scheme. Crossover and mutation operators are not always utilized. Their utilization is based on given crossover probability (P_C) and mutation probability (P_m).

The objective of this research is to see the performance of GA for solving PFSP based on the closeness of the solution generated by GA compared to the Best-Known Solution (BKS) of the test problems and to simulate the effect of the changes of crossover probability and mutation probability on GA performance. The test problems that are supplied with BKS are taken from OR Library [7].

2. PERMUTATION FLOWSHOP SCHEDULING PROBLEM

In a flowshop scheduling problem (FSP) we have n independent jobs $\{J_1, \dots, J_n\}$ that have to be processed on m different machines $\{M_1, \dots, M_m\}$. Every job is composed of m operations and every operation requires a different machine. Let O_{ij} denotes the operation on machine i of job j with processing time p_{ij} . In a flowshop, all jobs have the same processing order on machines since there is a technological order on the machine for different jobs to be followed. The objective is to find an ordering of the jobs on the machine or schedule that optimizes some given criteria. The most common criterion is to minimize total completion time of the schedule, often referred as makespan.

There are several additional conditions to this problem [2]: all operations are independent and available for processing at time 0, all m machines are continuously available, each machine can process at most one job at a time, each job can be processed only on one machine at a time, the processing of a given operation cannot be interrupted (no preemption is allowed), setup and removal times are sequence independent and are included in processing times or otherwise are negligible and can be ignored, and in-process inventory is allowed. If a given operation needs an unavailable machine then the operation joins a queue of unlimited size at that machine.

The FSP is known to be NP-complete in the strong sense when $m \geq 3$ [5]. In general, $(n!)^m$ schedules have to be considered in FSP with n jobs and m machines. In this paper we deal with simplification of FSP, that is the permutation flowshop scheduling problem or PFSP. In PFSP, the processing sequence of jobs is the same for all machines. There are $n!$ schedules to be considered in PFSP [8]. The objective of PFSP is to find sequence of the jobs on

machine that minimizes total completion time of the schedule (makespan). To count the makespan of the schedule we use Gantt-Chart.

Table 1. The example of PFSP with $n = 8$ and $m = 2$

Jobs	Processing Time	
	Machine 1	Machine 2
Job 1	4	3
Job 2	1	2
Job 3	4	2
Job 4	2	3
Job 5	2	4
Job 6	1	3
Job 7	4	1
Job 8	3	4

In Table 1 we give an example of PFSP with $n = 8$ and $m = 2$. An example of a schedule (solution) of this problem in (job, machine, processing time) is:

(1,1,4) (1,2,3) (2,1,1) (2,2,2) (3,1,4) (3,2,2) (4,1,2) (4,2,3) (5,1,2) (5,2,5)
 (6,1,1)(6,2,3) (7,1,4) (7,2,1) (8,1,3) (8,2,4)

Gantt-Chart for the above schedule shows that the makespan of the schedule is 26 (Figure 1).

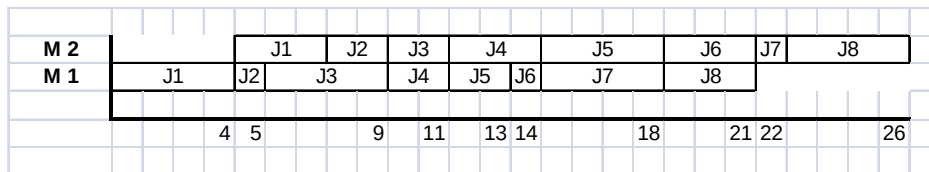


Figure 1. The example of Gantt-Chart

3. GENETIC ALGORITHM

The most important step of GA is the evolution that involves three main operators, which are reproduction/selection, crossover, and mutation. As mentioned before, we use roulette wheel and tournament reproduction, similar job order crossover (SJOX), and shift mutation operators. To improve GA performance, generational and replacement schemes are added.

Chromosome Representation and Generating Initial Population

To simplify the chromosome representation, assume machines are consecutive. Because the sequence of jobs is the same on each machine, chromosome can be represented by a string of jobs. Figure 2 gives an example of chromosome representation of the solution given before.

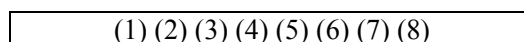


Figure 2. The example of chromosome representation

Population in GA contains *popsiz*e number of chromosomes that represent solutions of PFSP. Chromosomes in initial population are generated in three ways: with NEH procedure, modified NEH procedure, and random. The first chromosome is generated using NEH procedure. The next 25% percent of chromosome are generated using modified NEH procedure, and the rest are generated randomly. Those chromosomes form the initial population that represents the initial solution of the problem.

NEH procedure was found by Nawaz, Ensore, and Ham in 1983 and was considered as the best heuristic for PFSP in that era [11]. The steps of NEH procedure are:

1. Sort the n jobs by decreasing sums of processing time in all machine
2. Take the first two jobs and schedule them in order to minimize partial makespan of them.
3. For $i = 3$ to n do
4. Insert the i -th job in the place that minimizes the partial makespan.

Modified NEH procedure is the same with NEH except the first two jobs are chosen randomly.

Evaluation of the Fitness Function of Chromosomes

A fitness function is used to determine how good is the solution that a chromosome represented. If $T(x)$ denotes the makespan of chromosome x and max_length denotes makespan of the worst chromosome in the current population, the fitness function of a chromosome x , $F(x)$, is defined as,

$$F(x) = (max\_length - T(x)) + small\_int$$

where the *small_int* is the smallest fitness value allowed.

Reproduction

Reproduction is a process of choosing a parent population from a current population. Reproduction operator is applied by combining roulette wheel and tournament selection. Roulette wheel is developed based on the chromosome's fitness value so that each chromosome has an opportunity that proportional to chromosome's fitness value to be a parent. Parent is selected by spinning the roulette wheel. Tournament selection is an operator that chooses the best chromosome from a group of chromosomes. In the reproduction mechanism, we use roulette wheel to choose two chromosomes from the population as candidate parents and then tournament selection to select the best of them to be a parent.

Crossover

Crossover is a process of creating a child chromosome from a couple of parent chromosome. In GA, crossover is not always done. The use of crossover depends on the given crossover probability (P_c). If crossover is not carried out, then the parent will go straight to the next step. We use Similar Job Order Crossover (SJOX) operator. The SJOX mechanism can be explained as follows: First, the parents are examined on a position-by-position basis. Identical jobs at the same position are copied over to both offspring (Figure3a). Then, each offspring directly inherits all jobs from a parent up to a randomly chosen cut

point (Offspring 1 inherit directly from parent 1 and Offspring 2 inherit directly from parent 2)(Figure 3b). At last, the missing elements of each offspring are copied in the relative order to the other parent (Figure 3.c).

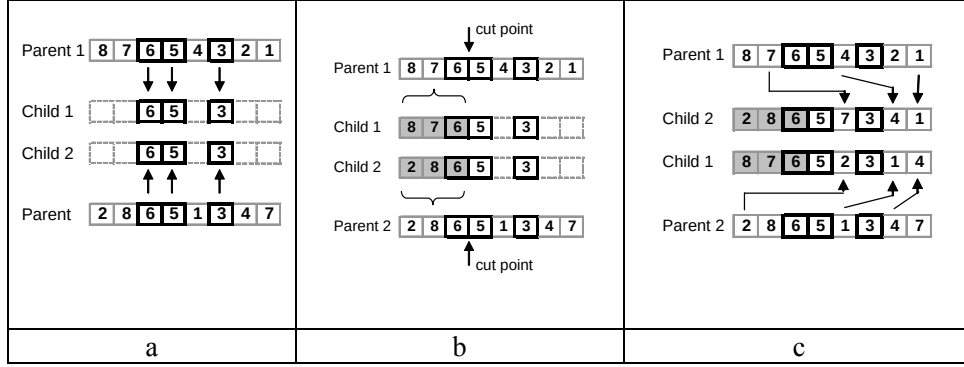


Figure 3. The illustration of SJOX crossover

Mutation

Mutation is used to maintain the diversity of the chromosomes in a population. Just like crossover, the use of mutation is depend on the mutation probability (P_m). We use shift mutation as a mutation operator. In shift mutation, the position of jobs in a part of the chromosome is shifted. This part is chosen randomly (Figure 4).

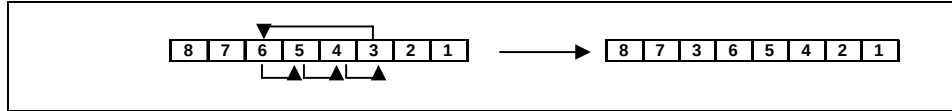


Figure 4. The illustration of shift mutation

Generational Scheme

A new population is generated by a generational scheme that can be explained as follows; let the worst chromosome in the population and its makespan are denoted by p_{worst} and c_{worst} . Several constrains are imposed to when offspring can replace p_{worst} . First, a given offspring can replace p_{worst} if its makespan is lower then c_{worst} . Also, a new individual will replace p_{worst} if its makespan is better than c_{worst} and the sequence of the individual is unique, i.e the sequence is not represented in the population.

Replacement Scheme

Replacement scheme is done if there is no improvement in the solution after some generations (iterations). Restart scheme is a process of changing chromosomes in a population. The purposes are to maintain genetically diversity in the population and to overcome the problem of the population getting stalled in local optima. Let Gr denotes the number of generation to do restart scheme, the restart scheme algorithm is as follows [10]:

1. At each generation i , store the minimum makespan, mak_i
2. If $mak_i = mak_{i-1}$ then $countmak = countmak + 1$. Otherwise $countmak = 0$

3. If $countmak > Gr$ then apply the following procedure:
 Sort the chromosomes in the population in ascending order of makespan.
 Keep the first 20% (the best chromosomes) in the population.
 For the remaining 80%, 50% of them are replaced by chromosomes created by doing simple SHIFT mutations to the first 20% best chromosomes, the next 25% are replaced by newly generated chromosomes using modified NEH procedure, and the remaining 25% are replaced by newly randomly generated chromosomes.

4. SIMULATION

GA is implemented using MATLAB 7 and run on PC with processor AMD Athlon 1533 MHz, 256 MB memory, and operating system Microsoft Windows XP Professional.

To test the performance of GA in solving PFSP, 12 test problems are used. All test problems are taken from OR-Library [7] and each of them is supplied with best-known solution (BKS). In each simulation, we do 5 times trials for each problem. Stopping criterion is the number of generation (gen) supplied by user. Population size ($popsiz$), crossover probability, mutation probability, and number of generation (Gr) to do restart scheme are also user supplied.

For the first simulation we use $popsiz = 50$, number of generation = 500, crossover probability = 0.1, mutation probability = 0.005, and number of generation (Gr) to do restart scheme = 25. The best result from 5 trials is given in Table 2 where column Test Problem, BKS, Best GA, R. Error represent the file name of the problem tested, the makespan of the best-known solution [7], the best value of the makespan based on the trials, and the relative error of Best GA compared to BKS.

Tabel 2. The result of the first simulation with $popsiz = 50$, $gen = 500$, $P_c = 0.1$, $P_m = 0.005$, and $Gr = 25$

Test Problem	# of jobs	# of machines	BKS	Best GA	R.error (%)
ta01	20	5	1278	1278	0
ta08	20	5	1206	1206	0
ta09	20	5	1230	1230	0
ta11	20	10	1582	1594	0.76
ta17	20	10	1484	1492	0.54
ta19	20	10	1593	1599	0.38
ta21	20	20	2297	2313	0.70
ta22	20	20	2099	2114	0.71
ta26	20	20	2226	2236	0.45
ta31	50	5	2724	2724	0
ta33	50	5	2621	2621	0
ta40	50	5	2782	2782	0

Table 2 shows that for 6 test problems with number of machine $n = 5$ the solution is the same as the BKS. For 6 other test problems ($n = 10$ or 20), the relative error is between 0.38 and 0.76. Actually, from this result we can conclude that performance of GA to solve PFSP is good. But, we want to do further exploration in the performance of GA to solve PFSP by doing some

simulation with different parameters.

Tabel 3. The result of the second simulation with $gen = 500$, $P_c = 0.1$, $P_m = 0.005$, $Gr = 25$, and altering value of $popsiz$

Test Problem	# of jobs	# of machines	BKS	<i>popsiz</i>				Best GA	R.error (%)
				26	50	76	100		
ta01	20	5	1278	1278	1278	1278	1278	1278	0
ta08	20	5	1206	1206	1206	1206	1206	1206	0
ta09	20	5	1230	1262	1230	1245	1230	1230	0
ta11	20	10	1582	1594	1594	1586	1594	1586	0.76
ta17	20	10	1484	1493	1492	1489	1492	1489	0.54
ta19	20	10	1593	1608	1599	1608	1608	1599	0.38
ta21	20	20	2297	2314	2313	2311	2310	2310	0.70
ta22	20	20	2099	2120	2114	2126	2140	2114	0.71
ta26	20	20	2226	2242	2236	2246	2232	2232	0.45
ta31	50	5	2724	2724	2724	2724	2724	2724	0
ta33	50	5	2621	2621	2621	2626	2622	2621	0
ta40	50	5	2782	2783	2782	2783	2788	2782	0

The second simulation used $gen = 500$, $P_c = 0.1$, $P_m = 0.005$, $Gr = 25$ with altering $popsiz$ values. The best result out of 5 trials is given in Table 3 where column $popsiz$ indicates the best makespan from 5 trials using the particular value of $popsiz$, and bold number indicates the best results of the problem for different value of $popsiz$. Table 3 explains that $popsiz = 50$ gives better solution so that for the next simulation we use the value of $popsiz = 50$.

Tabel 4. The result of the second simulation with $popsiz = 50$, $gen = 500$, $P_m = 0.005$, $Gr = 25$, and altering value of P_c

Test Problem	# of jobs	# of machines	BKS	Crossover probability					Best GA	R.error (%)
				0.0	0.1	0.2	0.3	0.4		
ta01	20	5	1278	1281	1278	1278	1278	1278	1278	0
ta08	20	5	1206	1217	1206	1206	1206	1206	1206	0
ta09	20	5	1230	1262	1230	1245	1252	1230	1230	0
ta11	20	10	1582	1638	1594	1594	1599	1594	1594	0.76
ta17	20	10	1484	1503	1492	1492	1489	1492	1489	0.34
ta19	20	10	1593	1608	1599	1608	1608	1608	1599	0.38
ta21	20	20	2297	2325	2313	2298	2298	2311	2298	0.04
ta22	20	20	2099	2149	2114	2118	2116	2108	2108	0.43
ta26	20	20	2226	2259	2236	2235	2230	2232	2230	0.18
ta31	50	5	2724	2724	2724	2724	2724	2724	2724	0
ta33	50	5	2621	2641	2621	2622	2622	2622	2621	0
ta40	50	5	2782	2783	2782	2783	2788	2783	2782	0

The third simulation used $popsiz = 50$, $gen = 500$, $P_m = 0.005$, $Gr = 25$ with altering value of crossover probability, P_c . The best result out of 5 trials is given in Table 4 where column crossover probability indicates the best makespan from 5 trials using that particular value of P_c . The table shows that there is no value of P_c that outperforms the other value for all test problems. But, for all test problems with 5 machines, $P_c = 0.1$ always give the solution with the same makespan with BKS.

The fourth simulation used $popsiz = 50$, $gen = 500$, $P_c = 0.1$, $Gr = 25$ with altering the value of mutation probability, P_m . The best result out of 5 trials is given in Table 5 where column mutation probability indicates the best makespan from 5 trials using that particular value of P_m . Table 5 explains that the best solution is more often found using $P_m = 0.005$ and $P_m = 0.02$. But, $P_m = 0.005$ gives the solution with the same makespan with BKS only for test problem with 5 machines.

Tabel 5. The result of the second simulation with $popsiz = 50$, $gen = 500$, $P_c = 0.1$, $Gr = 25$, and altering value of P_m

Test Problem	# of jobs	# of machines	BKS	mutation probability					Best GA	R.error (%)
				0.0	0.005	0.01	0.015	0.02		
ta01	20	5	1278	1278	1278	1278	1278	1278	1278	0
ta08	20	5	1206	1206	1206	1206	1206	1206	1206	0
ta09	20	5	1230	1230	1230	1230	1230	1230	1230	0
ta11	20	10	1582	1599	1594	1597	1592	1586	1586	0.25
ta17	20	10	1484	1492	1492	1492	1491	1489	1489	0.34
ta19	20	10	1593	1608	1599	1608	1608	1608	1599	0.38
ta21	20	20	2297	2316	2313	2319	2312	2318	2312	0.65
ta22	20	20	2099	2143	2114	2119	2120	2105	2105	0.29
ta26	20	20	2226	2257	2236	2239	2232	2236	2232	0.27
ta31	50	5	2724	2724	2724	2724	2724	2724	2724	0
ta33	50	5	2621	2624	2621	2622	2622	2622	2621	0
ta40	50	5	2782	2782	2782	2783	2782	2783	2782	0

The fifth simulation used $popsiz = 50$, $gen = 500$, $P_c = 0.1$, $P_m = 0.005$, with altering the value of the number of generation to do restart scheme, Gr . The best result out of 5 trials is given in Table 6 where column Gr restart scheme indicates the best makespan from 5 trials using that particular value of Gr . Table 6 shows that $Gr = 25$ is outperform the other Gr value for 10 out of 12 test problems.

Tabel 6. The result of the second simulation with $popsiz = 50$, $gen = 500$, $P_c = 0.1$, $P_m = 0.005$, and altering value of Gr

Test Problem	# of jobs	# of machines	BKS	Gr restart scheme						Best GA	R.error (%)
				15	20	25	30	50	75		
ta01	20	5	1278	1278	1278	1278	1278	1278	1278	1278	0
ta08	20	5	1206	1206	1206	1206	1214	1212	1214	1206	0

ta09	20	5	1230	1230	1230	1230	1230	1230	1252	1230	0
ta11	20	10	1582	1594	1596	1594	1592	1603	1598	1594	0.76
ta17	20	10	1484	1489	1493	1492	1493	1493	1493	1489	0.34
ta19	20	10	1593	1608	1608	1599	1608	1608	1608	1599	0.38
ta21	20	20	2297	2318	2319	2313	2319	2316	2319	2313	0.07
ta22	20	20	2099	2118	2118	2114	2143	2124	2121	2114	0.71
ta26	20	20	2226	2235	2234	2236	2232	2236	2240	2232	0.27
ta31	50	5	2724	2724	2724	2724	2724	2724	2724	2724	0
ta33	50	5	2621	2622	2622	2621	2622	2623	2622	2621	0
ta40	50	5	2783	2783	2783	2782	2783	2783	2783	2782	0

The best result for each test problem from all simulations is given in Table 7, where column Parameter indicates the condition in which the best solution is found. In this column, “all” indicates the best objective function value is found for every P_C and P_m tested, while “***” indicates the best objective function value is found on more than 2 conditions.

Tabel 7. The best result from all simulations

Test Problem	# of jobs	# of machines	BKS	Best GA	R.error (%)	Value Parameters
ta01	20	5	1278	1278	0	***

ta08	20	5	1206	1206	0	***
ta09	20	5	1230	1230	0	***
ta11	20	10	1582	1586	0.25	Popsize=76, Pc=0.1, Pm=0.005, Gr=25; Popsize=50, Pc=0.1, Pm=0.002, Gr=25
ta17	20	10	1484	1489	0.34	***
ta19	20	10	1593	1599	0.38	***
ta21	20	20	2297	2294	0.04	Popsize=50, Pc=0.2, Pm=0.005, Gr=25; Popsize=50, Pc=0.3, Pm=0.005, Gr=25
ta22	20	20	2099	2105	0.29	Popsize=50, Pc=0.1, Pm=0.02, Gr=25
ta26	20	20	2226	2230	0.18	Popsize=50, Pc=0.3, Pm=0.005, Gr=25
ta31	50	5	2724	2724	0	***
ta33	50	5	2621	2621	0	***
ta40	50	5	2782	2782	0	***

Table 7 shows that for 6 of 12 test problems the solution is the same as the BKS. For 6 other test problems, the relative error is between 0.04 and 0.38. For test problem with 5 machines, GA gives solution with the same makespan with BKS in all simulation. Unfortunately, in all simulations BKS is never attended for the test problems with the number of machines are more than 5.

5. CONCLUSION

The performance of Genetic Algorithm in solving PFSP is quite good. Relative error of makespan of the best solution generated by GA compares to BKS is less than 0.4 % for all test problems. In our simulation, the value of *popsiz*e = 50, crossover probability = 0.1, and number of generation to do restart scheme = 25 make the performance of GA better, but we do not get the best value for mutation probability.

REFERENCES

1. <http://www.metaheuristics.org>, 29 October 2004, at. 20.30.
2. Baker, K. R. *Introduction to Sequencing and Scheduling*. John Wiley & Sons, New York. 1974.
3. Ombuki, B. M., and Ventresca, M. *Meta-Heuristics for the Job Shop Scheduling Problem*. 19 December 2003.
4. Ombuki, B. M., and Ventresca, M. Local Search Genetic Algorithm for the Job Shop Scheduling Problem. *Journal of Applied Intelligence* 21 (1): 99-109, July 2004.
5. Garey, M. R., Johnson, D. S., and Sethi, R. The complexity of flowshop and job shop scheduling. *Mathematics of Operations Research*, 1(2): 117-129. 1976.
6. *Genetic Algorithms*.
http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/tcw2/report.html#Internet, 13 Juni 2006, at. 20.04.
7. *OR Library*, <http://www.brunel.ac.uk/depts/ma/research/jeb/info.html>, 4

October 2004.

8. Reeves, C. R. and Yamada T., *Genetic Algorithms, Path Relinking and the Flowshop Sequencing Problem*. 1998.
9. Ruiz, R., Maroto, C., and Alcaraz, J., *New Genetic Algorithms for the Permutation Flowshop Scheduling Problem*. August 2003.
10. Ruiz, R., Maroto, C., dan Alcaraz, J., *Two New Robust Genetic Algorithms for the Flowshop Scheduling Problem*, 7 December 2004.
11. Taillard, E. *Some Efficient Heuristic Methods for the Flow Shop Sequencing Problem*. October 1988.

D. R. Silaban, D. Kusharyadi, Y. Satria:

Department of Mathematics, University of Indonesia, Depok 16424, Indonesia.

E-mail: {denny-rs, y-satria}@ui.edu.

THE IMMERSED INTERFACE METHOD FOR TWO CHANNELS DISSIPATION MODEL WITH DISCONTINUE CONSTANTS VELOCITY

SUMARDI, SOEPARNA DARMAWIJAYA AND LINA ARYATI

Abstract

The Immersed Interface Method to find numerical solution one dimensional two channels dissipation model in heterogeneous media is presented. The equation of two channels dissipation model are written as first-order linear hyperbolic system. The idea of the standard scheme Lax-Wendroff is used, but it is applied on a set of modified values deduced from numerical values and from jump conditions on interfaces. Numerical examples show that we can compute solutions to these equations with second order accuracy.

Keywords : hyperbolic system, finite difference method, immersed interface method

1. INTRODUCTION

Many process of fundamental equation occurring in the natural and physical sciences are obtained from conservation laws. Conservation laws are just balance laws, equations expressing the fact that some quantity is balanced throughout process. Mathematically, conservation laws usually translate into differential equations. In this paper let us consider $u(x,t)$ and $v(x,t)$ that depend on single spatial variable x and time $t \geq 0$. We assume that $u(x,t)$ and $v(x,t)$ are densities or concentrations measured in amount per unit volume that flow with speeds $c_1(x)$ and $c_2(x)$ respectively. Two groups of quantity have interactions, and then van Beckum (2003) arrived at the following set of equations for the concentrations $u(x,t)$ in the first group and $v(x,t)$ in the second group:

$$\begin{aligned} \frac{\partial u(x,t)}{\partial t} + c_1(x) \frac{\partial u(x,t)}{\partial x} &= \alpha v(x,t) - \alpha u(x,t) \\ \frac{\partial v(x,t)}{\partial t} + c_2(x) \frac{\partial v(x,t)}{\partial x} &= \alpha u(x,t) - \alpha v(x,t) \end{aligned} \quad (1)$$

The constant α is the coefficient of exchanging concentrations. The initial values the system given by

$$u(x,0) = f(x), \quad v(x,0) = g(x) \quad (2)$$

The approximation of the classical solution of the above equations that have boundary conditions has been investigated Sumardi (2005) by using finite difference method in case $c_1(x) = -c_2(x) = 1$. Sumardi (2006) has solved the exact solution by Fourier transform in case $c_1(x)$ and $c_2(x)$ constant. Now we would like to present in case the heterogeneous media with interface $x = \beta$. The velocities that change at in the interface are written:

$$c_1(x) = \begin{cases} c_1^- & x < \beta \\ c_1^+ & x > \beta \end{cases} \quad c_2(x) = \begin{cases} c_2^- & x < \beta \\ c_2^+ & x > \beta \end{cases} \quad (3)$$

The immersed interface method (IIM) has been developed recently by Leveque and Li (1994) to solve interface problem in which the differential equations have discontinuities and singularities in the coefficients and solutions. Zhang (1998) also used IIM for acoustic wave equations with discontinuities coefficients.

2. SCALAR TRANSPORT-DISSIPATION MODEL

Firstly we apply the Immersed Interface Method for scalar problem that relates with two channels dissipation model. The problem is to find solution of the scalar transport dissipation model (also called plug flow reactor, piston flow, slug flow, or ideal tubular reactor):

$$U_t + c(x)U_x = \alpha U, \quad 0 \leq x \leq 1, t > 0$$

$$c(x) = \begin{cases} c^- & x < \beta \\ c^+ & x > \beta \end{cases} \quad U(x,0) = f(x) \quad (4)$$

Here U represent some physical quantity, $c(x)$ is the speed of the transport, with some jump condition specified at $x = \beta$. The jump condition will depend on the physical problem. The first we can impose the flux cU should be continuous that is, the number of item crossing the interface per unit time must be the same on both sides, and therefore we have jump condition:

$$[cU] = 0.$$

The other is that the quantity $U(x,t)$ should be continuous, so the jump condition is

$$[U] = 0.$$

To solve this problem by the Immersed Interface method, we need finite difference method approximation.

To get second order accuracy, it is used the idea of the Lax-Wendroff method that is based on the Taylor method series expansion. We discretize the $x-t$ plane by choosing a mesh width $h \equiv \Delta x$ and a time step $k \equiv \Delta t$ and define discrete mesh points (x_j, t_n) by

$$x_j = jh, \quad j = 0, 1, 2, \dots, N$$

$$t_n = nk, \quad n = 0, 1, 2, \dots$$

For simplicity we take a uniform mesh, with h and k constant, although most of the methods discussed can be extended to variable meshes. The finite difference

methods we will develop produce approximation $U_j^n \approx U(x_j, t_n)$.

Consider

$$\begin{aligned} U_t &= \alpha U - c(x)U_x \\ U_{tt} &= \alpha U_t - c(x)U_{xt} \\ &= \alpha^2 U - 2\alpha c(x)U_x + c(x)^2 U_{xx} \end{aligned}$$

and use Taylor series, we obtain

$$\begin{aligned} U(x_j, t_{n+1}) &= U(x_j, t_n) + kU_t(x_j, t_n) + \frac{1}{2}k^2 U_{tt}(x_j, t_n) + \dots \\ &= U(x_j, t_n) + k(\alpha U(x_j, t_n) - c(x)U_x(x_j, t_n)) + \\ &\quad \frac{1}{2}k^2(\alpha^2 U(x_j, t_n) - 2\alpha c(x)U_x(x_j, t_n) + c(x)^2 U_{xx}(x_j, t_n)) + \dots \end{aligned} \quad (5)$$

To get second order accuracy, the approximate are

$$\begin{aligned} U_x(x_j, t_n) &\approx \frac{U_{j+1}^n - U_{j-1}^n}{2h} \\ U_{xx}(x_j, t_n) &\approx \frac{U_{j+1}^n - 2U_j^n + U_{j-1}^n}{h^2} \end{aligned} \quad (6)$$

The method gives

$$\begin{aligned} U_j^{n+1} &= \left(1 + k\alpha + \frac{k^2\alpha^2}{2} - \frac{k^2c(x)^2}{h^2}\right)U_j^n + \\ &\quad \left(\frac{kc(x)}{2h} + \frac{k^2\alpha c(x)}{2h} + \frac{k^2c(x)^2}{2h^2}\right)U_{j-1}^n + \left(-\frac{kc(x)}{2h} - \frac{k^2\alpha c(x)}{2h} + \frac{k^2c(x)^2}{2h^2}\right)U_{j+1}^n \end{aligned} \quad (7)$$

This valid for all the regular points, i.e. those for which a discontinuous point do not lie the interval of the 3-points stencil. Now consider the two irregular point J and $J+1$, with

$$x_J < \beta < x_{J+1}$$

i.e., β is the interface location in between of points x_J and x_{J+1} .

We will assume that β is not exactly green point, but lies strictly between points x_J and x_{J+1} . Then the Taylor series expansion is still valid, the solution is smooth in both x and t at every grid point, every those near the interface. The problem comes in approximating U_x and U_{xx} by difference formulas that uses values of U lying both sides of the interface. However, by taking into account the jumps in U and its derivatives at β , it is possible to find linear combinations of grid values that do give accurate approximation to the derivatives at grid points.

Rather than deriving expressions for U_x and U_{xx} separately, it is easiest to define three points schema of the form

$$\begin{aligned}
 U_J^{n+1} &= U_J^n + \frac{k}{h} (\gamma_{J,1} U_{J-1}^n + \gamma_{J,2} U_J^n + \gamma_{J,3} U_{J+1}^n) \\
 U_{J+1}^{n+1} &= U_{J+1}^n + \frac{k}{h} (\gamma_{J+1,1} U_{J+2}^n + \gamma_{J+1,2} U_{J+1}^n + \gamma_{J+1,3} U_J^n)
 \end{aligned} \tag{8}$$

and then compute the proper Γ coefficients so that the local truncation error (difference between exact and approximate solution) gives second order accuracy.

Consider the local truncation error at x_J

$$\begin{aligned}
 L &= \frac{1}{h} (\gamma_{J,1} U_{J-1}^n + \gamma_{J,2} U_J^n + \gamma_{J,3} U_{J+1}^n) \\
 &\quad + (-\alpha U + c(x) U_x) - \frac{1}{2} k (\alpha^2 U - 2\alpha c(x) U_x + c(x)^2 U_{xx}) + O(k^2)
 \end{aligned} \tag{9}$$

Taylor series expansion at $x = \beta$ and $h = O(k)$ gives

$$\begin{aligned}
 L &= \frac{1}{h} \gamma_{J,1} (U^- + (x_{J-1} - \beta) U_x^- + \frac{1}{2} (x_{J-1} - \beta)^2 U_{xx}^-) \\
 &\quad + \frac{1}{h} \gamma_{J,2} (U^- + (x_J - \beta) U_x^- + \frac{1}{2} (x_J - \beta)^2 U_{xx}^-) \\
 &\quad + \frac{1}{h} \gamma_{J,3} (U^+ + (x_{J+1} - \beta) U_x^+ + \frac{1}{2} (x_{J+1} - \beta)^2 U_{xx}^+) \\
 &\quad + (-\alpha (U^- + (x_J - \beta) U_x^-) + c^- (U_J^- + (x_J - \beta) U_{xx}^-)) \\
 &\quad - \frac{1}{2} k (\alpha^2 U^- - 2\alpha c^- U_x^- + (c^-)^2 U_{xx}^-) + O(k^2)
 \end{aligned} \tag{10}$$

We can eliminate U^+, U_x^+ and U_{xx}^+ by jump conditions. Given jump condition $[cU] = 0$, we can derive the jump condition for U_x and U_{xx} by using the scalar transport dissipation model:

$$\begin{aligned}
 [cU] = 0 &\Rightarrow [cU_t] = 0 \Rightarrow [c^2 U_x] = 0 \\
 [cU] = 0 &\Rightarrow [cU_{tt}] = 0 \Rightarrow [c^3 U_{xx}] = 0
 \end{aligned} \tag{11}$$

so we get equations as follows:

$$\begin{aligned}
 U^+ &= d_1 U^- \\
 U_x^+ &= d_2 U_x^- \\
 U_{xx}^+ &= d_3 U_{xx}^-
 \end{aligned} \tag{12}$$

where,

$$d_j = \left(\frac{c^-}{c^+} \right)^j, \quad (j = 1, 2, 3). \tag{13}$$

Plugging (12) into (10) yields

$$\begin{aligned}
 L = & \frac{1}{h} \gamma_{J,1} \left(U^- + (x_{J-1} - \beta) U_x^- + \frac{1}{2} (x_{J-1} - \beta)^2 U_{xx}^- \right) \\
 & + \frac{1}{h} \gamma_{J,2} \left(U^- + (x_J - \beta) U_x^- + \frac{1}{2} (x_J - \beta)^2 U_{xx}^- \right) \\
 & + \frac{1}{h} \gamma_{J,3} \left(d_1 U^- + (x_{J+1} - \beta) d_2 U_x^- + \frac{1}{2} d_3 (x_{J+1} - \beta)^2 U_{xx}^- \right) \\
 & + \left(-\alpha (U^- + (x_J - \beta) U_x^-) + c^- (U_x^- + (x_J - \beta) U_{xx}^-) \right) \\
 & - \frac{1}{2} k \left(\alpha^2 U^- - 2\alpha c^- U_x^- + (c^-)^2 U_{xx}^- \right) + O(k^2)
 \end{aligned} \tag{14}$$

or

$$\begin{aligned}
 L = & \frac{1}{h} (\gamma_{J,1} + \gamma_{J,2} + d_1 \gamma_{J,3} - \alpha h - \frac{1}{2} h k \alpha^2) U^- \\
 & + \frac{1}{h} \left((x_{J-1} - \beta) \gamma_{J,1} + (x_J - \beta) \gamma_{J,2} + d_2 (x_{J+1} - \beta) \gamma_{J,3} - \alpha h (x_J - \beta) + c^- h + k h \alpha c^- \right) U_x^- \\
 & + \frac{1}{2h} \left((x_{J-1} - \beta)^2 \gamma_{J,1} + (x_J - \beta)^2 \gamma_{J,2} + d_3 (x_{J+1} - \beta)^2 \gamma_{J,3} + 2c^- h (x_J - \beta) + (c^-)^2 k h \right) U_{xx}^- \\
 & + O(k^2)
 \end{aligned} \tag{15}$$

It will suffice for second order accuracy if we force coefficients of U^- , U_x^- and U_{xx}^- to be zero, so we obtain linear system:

$$\begin{aligned}
 \beta_{11} \gamma_{J,1} + \beta_{12} \gamma_{J,2} + \beta_{13} d_1 \gamma_{J,3} &= \alpha h + \frac{1}{2} h k \alpha^2 \\
 \beta_{21} \gamma_{J,1} + \beta_{22} \gamma_{J,2} + \beta_{23} d_2 \gamma_{J,3} &= \alpha (x_J - \beta) - c^- - k \alpha c^- \\
 \beta_{31} \gamma_{J,1} + \beta_{32} \gamma_{J,2} + \beta_{33} d_3 \gamma_{J,3} &= -2c^- \frac{(x_J - \beta)}{h} + \frac{k}{h} (c^-)^2
 \end{aligned} \tag{16}$$

where

$$\beta_{ij} = \left(\frac{(x_{J+(j-2)} - \beta)}{h} \right)^{i-1}, \quad i, j = 1, 2, 3 \tag{17}$$

Similarly at the point x_{J+1} , we obtain the linear equations

$$\begin{aligned}
 \hat{\beta}_{11} \gamma_{J+1,1} + \hat{\beta}_{12} \gamma_{J+1,2} + \hat{\beta}_{13} d_1 \gamma_{J+1,3} &= \alpha h + \frac{1}{2} h k \alpha^2 \\
 \hat{\beta}_{21} \gamma_{J+1,1} + \hat{\beta}_{22} \gamma_{J+1,2} + \hat{\beta}_{23} d_1 \gamma_{J+1,3} &= \alpha (x_J - \beta) - c^+ - k \alpha c^+ \\
 \hat{\beta}_{31} \gamma_{J+1,1} + \hat{\beta}_{32} \gamma_{J+1,2} + \hat{\beta}_{23} d_1 \gamma_{J+1,3} &= -2c^+ \frac{(x_J - \beta)}{h} + \frac{k}{h} (c^+)^2
 \end{aligned} \tag{18}$$

where

$$\hat{\beta}_{ij} = \left(\frac{(x_{J-(j-2)} - \beta)}{h} \right)^{i-1}, \quad i, j = 1, 2, 3 \tag{19}$$

For different jump conditions, we only need to make slight modification. Consider now the jump condition $[U] = 0$. We have

$$\begin{aligned} [U] = 0 &\Rightarrow [U_t] = 0 \Rightarrow [cU_x] = 0 \\ [U] = 0 &\Rightarrow [U_{tt}] = 0 \Rightarrow [c^2 U_{xx}] = 0 \end{aligned} \quad (20)$$

The expression (13) for U^+ , U_x^+ and U_{xx}^+ are still valid, with the modification of d_1, d_2 and d_3

$$d_j = \left(\frac{c^-}{c^+} \right)^{j-1}, \quad (j = 1, 2, 3) \quad (21)$$

The linear equations for the γ 's remain the same, but with this new value of d_1, d_2 and d_3

Numerical result

Here we will show two examples. Figure 1 gives the numerical solution to the transport-dissipation model with jump $[U] = 0$ (left) and $[cU] = 0$ (right) at $\beta = \frac{5}{9}$, at $\alpha = -1$, at $t = 0, t = 0.15, t = 0.30$ and $t = 0.45$. The initial data is single hump given by

$$U(x, 0) = \begin{cases} \frac{1}{2}(1 + \cos((x - 0.28)\pi / 0.24)) & 0.04 \leq x \leq 0.52 \\ 0 & 0 \leq x < 0.04 \vee 0.52 < x \leq 1 \end{cases}$$

The wave speeds are $c^- = 2$ and $c^+ = 1$.

Figure 1. gives the solutions computed on grid size $h = 1/100$ and $k = 0.001$.

After reaching the interface, the wave speed changes accordingly on the other sides. For the case of the jump conditions $[U] = 0$, the solution values is continuous. The magnitude is not changed by the change of speed at the interface, but the wave length is changed by the change of speed at in the interface. The magnitude is changed by the term of dissipative.

For the case of the jump conditions $[cU] = 0$, the solution values is not continuous and there is sharp jump in the solution. The magnitude is changed by the change of speed at the interface, and the wave length is also changed by the change of speed at the interface. The magnitude is changed by the term of dissipative.

The exact solution of equation (4) is

$$U(x, t) = e^{\alpha t} f(x - c(x)t)$$

By this solution we only have one solution, but by the immersed interface method we have two solutions. It is still possible we can find more two solutions.

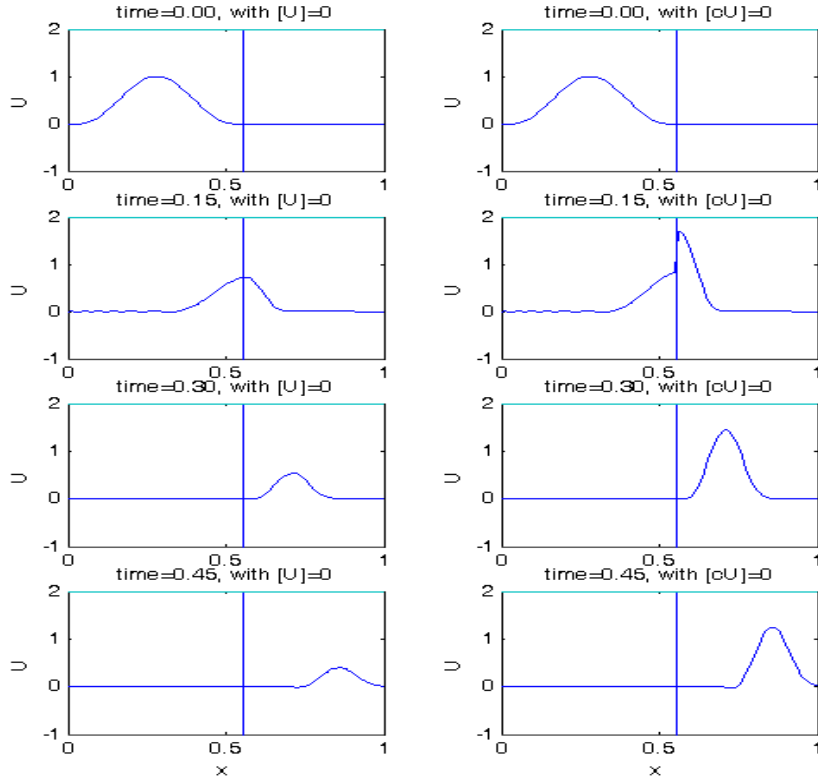


Figure 1. : Transport-dissipation model with jump $[U] = 0$ (left) and $[cU] = 0$ (right)

at $\beta = \frac{5}{9}$, at $\alpha = -1$, at $t = 0$, $t = 0.15$, $t = 0.30$ and $t = 0.45$.

3. TWO CHANNELS DISSIPATION MODEL

The main interest of this paper is to use the Immersed Interface Method to solve two channels dissipation model with discontinue constant speeds. The equation given by

$$\begin{aligned} \frac{\partial u(x,t)}{\partial t} + c_1(x) \frac{\partial u(x,t)}{\partial t} &= \alpha v(x,t) - \alpha u(x,t) \\ \frac{\partial v(x,t)}{\partial t} + c_2(x) \frac{\partial v(x,t)}{\partial t} &= \alpha u(x,t) - \alpha v(x,t) \end{aligned} \quad (22)$$

for the case the speeds have discontinuity at the interface $x = \beta$

$$c_1(x) = \begin{cases} c_1^- & x < \beta \\ c_1^+ & x > \beta \end{cases} \quad c_2(x) = \begin{cases} c_2^- & x < \beta \\ c_2^+ & x > \beta \end{cases} \quad (23)$$

The jump condition at the interface like in scalar transport dissipation model

$$[u] = 0, [v] = 0 \quad (24)$$

or

$$[c_1 u] = 0, [c_2 v] = 0. \quad (25)$$

To solve this problem let

$$U = \begin{pmatrix} u(x, t) \\ v(x, t) \end{pmatrix}, \quad C = \begin{pmatrix} c_1(x) & 0 \\ 0 & c_2(x) \end{pmatrix}, \quad A = \begin{pmatrix} -\alpha & \alpha \\ \alpha & -\alpha \end{pmatrix} \quad (26)$$

This yields a system of first order equations

$$U_t + CU_x = AU \quad (27)$$

To get second order accuracy we use the method that similar the equation (7), so we have scheme:

$$\begin{aligned} U_j^{n+1} = & \left(I + kA + \frac{k^2 A^2}{2h} - \frac{k^2 C^2}{h^2} \right) U_j^n + \left(\frac{kC}{2h} + \frac{k^2(AC + CA)}{4h} + \frac{k^2 C^2}{2h^2} \right) U_{j-1}^n \\ & + \left(-\frac{kC}{2h} - \frac{k^2(AC + CA)}{4h} + \frac{k^2 C^2}{2h^2} \right) U_{j+1}^n \end{aligned} \quad (28)$$

This can be used all points $j \neq J, J+1$, where J is such that

$$x_J < \beta < x_{J+1}$$

Analogous to the case scalar transport dissipation equation, we will use more general 3-point difference scheme of the form:

$$\begin{aligned} U_J^{n+1} &= U_J^n + \frac{k}{h} (\Gamma_{J,1} U_{J-1}^n + \Gamma_{J,2} U_J^n + \Gamma_{J,3} U_{J+1}^n) \\ U_{J+1}^{n+1} &= U_{J+1}^n + \frac{k}{h} (\Gamma_{J+1,1} U_{J+2}^n + \Gamma_{J+1,2} U_{J+1}^n + \Gamma_{J+1,3} U_J^n) \end{aligned} \quad (29)$$

But U is a vector variable and Γ 's are 2×2 matrices, in contrast to the problem in section before that U and γ 's are scalar. We want to choose the Γ 's to get second order accuracy. Again, consider the local truncation error at J :

$$\begin{aligned} L = & \frac{1}{h} (\Gamma_{J,1} U_{J-1}^n + \Gamma_{J,2} U_J^n + \Gamma_{J,3} U_{J+1}^n) \\ & + (-\alpha U + c(x) U_x) - \frac{1}{2} k (\alpha^2 U - 2\alpha c(x) U_x + c(x)^2 U_{xx}) + O(k^2) \end{aligned} \quad (30)$$

Expanding each U at $x = \beta$ and $h = O(k)$ gives

$$\begin{aligned}
 L = & \frac{1}{h} \Gamma_{J,1} \left(U^- + (x_{J-1} - \beta) U_x^- + \frac{1}{2} (x_{J-1} - \beta)^2 U_{xx}^- \right) \\
 & + \frac{1}{h} \Gamma_{J,2} \left(U^- + (x_J - \beta) U_x^- + \frac{1}{2} (x_J - \beta)^2 U_{xx}^- \right) \\
 & + \frac{1}{h} \Gamma_{J,3} \left(U^+ + (x_{J+1} - \beta) U_x^+ + \frac{1}{2} (x_{J+1} - \beta)^2 U_{xx}^+ \right) \\
 & + \left(-A(U^- + (x_J - \beta) U_x^-) + C^-(U_J^- + (x_J - \beta) U_{xx}^-) \right) \\
 & - \frac{1}{2} k \left(A^2 U^- - (AC^- + C^- A) U_x^- + (C^-)^2 U_{xx}^- \right) + O(k^2)
 \end{aligned} \tag{31}$$

where

$$C^- = \begin{pmatrix} c_1^- & 0 \\ 0 & c_2^- \end{pmatrix}, \quad C^+ = \begin{pmatrix} c_1^+ & 0 \\ 0 & c_2^+ \end{pmatrix}. \tag{32}$$

We can express U^+, U_x^+ and U_{xx}^+ in term U^-, U_x^- and U_{xx}^- by jump conditions, similar with section before as follows:

$$\begin{aligned}
 U^+ &= D_1 U^- \\
 U_x^+ &= D_2 U_x^- \\
 U_{xx}^+ &= D_3 U_{xx}^-
 \end{aligned} \tag{33}$$

where,

$$D_j = \begin{pmatrix} \frac{c_1^-}{c_1^+} & 0 \\ 0 & \frac{c_2^-}{c_2^+} \end{pmatrix}^j, \quad (j = 1, 2, 3) \tag{34}$$

for jump conditions $[CU] = 0$, and

$$D_j = \begin{pmatrix} \frac{c_1^-}{c_1^+} & 0 \\ 0 & \frac{c_2^-}{c_2^+} \end{pmatrix}^{j-1}, \quad (j = 1, 2, 3) \tag{35}$$

for jump condition $[U] = 0$.

Substituting interface conditions (33) into (31) gives

$$\begin{aligned}
 L = & \frac{1}{h} \Gamma_{J,1} \left(U^- + (x_{J-1} - \beta) U_x^- + \frac{1}{2} (x_{J-1} - \beta)^2 U_{xx}^- \right) \\
 & + \frac{1}{h} \Gamma_{J,2} \left(U^- + (x_J - \beta) U_x^- + \frac{1}{2} (x_J - \beta)^2 U_{xx}^- \right) \\
 & + \frac{1}{h} \Gamma_{J,3} \left(D_1 U^- + (x_{J+1} - \beta) D_2 U_x^- + \frac{1}{2} (x_{J+1} - \beta)^2 D_3 U_{xx}^- \right) \quad (36) \\
 & + \left(-A(U^- + (x_J - \beta) U_x^-) + C^-(U_J^- + (x_J - \beta) U_{xx}^-) \right) \\
 & - \frac{1}{2} k \left(A^2 U^- - (AC^- + C^- A) U_x^- + (C^-)^2 U_{xx}^- \right) + O(k^2)
 \end{aligned}$$

or

$$\begin{aligned}
 L = & \frac{1}{h} (\Gamma_{J,1} + \Gamma_{J,2} + D_1 \Gamma_{J,3} - hA - \frac{1}{2} h k A^2) U^- \\
 & + \frac{1}{h} \left((x_{J-1} - \beta) \Gamma_{J,1} + (x_J - \beta) \Gamma_{J,2} + D_2 (x_{J+1} - \beta) \Gamma_{J,3} - h(x_J - \beta) A + C^- h + \frac{kh}{2} (AC^- + C^- A) \right) U_x^- \\
 & + \frac{1}{2h} \left((x_{J-1} - \beta)^2 \Gamma_{J,1} + (x_J - \beta)^2 \Gamma_{J,2} + D_3 (x_{J+1} - \beta)^2 \Gamma_{J,3} + 2C^- h(x_J - \beta) + (C^-)^2 kh \right) U_{xx}^- + O(k^2) \quad (37)
 \end{aligned}$$

It will suffice for second order accuracy if we force coefficients of U^- , U_x^- and U_{xx}^- to be zero, so we obtain a system of matrix equations:

$$\begin{aligned}
 \beta_{11} \Gamma_{J,1} + \beta_{12} \Gamma_{J,2} + \beta_{13} D_1 \Gamma_{J,3} &= hA + \frac{1}{2} h k A^2 \\
 \beta_{21} \Gamma_{J,1} + \beta_{22} \Gamma_{J,2} + \beta_{23} D_2 \Gamma_{J,3} &= (x_J - \beta) A - C^- - \frac{k}{2} (AC^- + C^- A) \quad (38) \\
 \beta_{31} \Gamma_{J,1} + \beta_{32} \Gamma_{J,2} + \beta_{33} D_3 \Gamma_{J,3} &= -2C^- \frac{(x_J - \beta)}{h} + \frac{k}{h} (C^-)^2
 \end{aligned}$$

where

$$\beta_{ij} = \left(\frac{(x_{J+(j-2)} - \beta)}{h} \right)^{i-1}, \quad i, j = 1, 2, 3 \quad (39)$$

Similarly at the point x_{J+1} , we obtain the linear equations

$$\begin{aligned}
 \hat{\beta}_{11} \Gamma_{J+1,1} + \hat{\beta}_{12} \Gamma_{J+1,2} + \hat{\beta}_{13} D_1 \Gamma_{J+1,3} &= hA + \frac{1}{2} h k A^2 \\
 \hat{\beta}_{21} \Gamma_{J+1,1} + \hat{\beta}_{22} \Gamma_{J+1,2} + \hat{\beta}_{23} D_2 \Gamma_{J+1,3} &= (x_J - \beta) A - C^+ - \frac{k}{2} (AC^+ + C^+ A) \\
 \hat{\beta}_{31} \Gamma_{J+1,1} + \hat{\beta}_{32} \Gamma_{J+1,2} + \hat{\beta}_{33} D_3 \Gamma_{J+1,3} &= -2C^+ \frac{(x_J - \beta)}{h} + \frac{k}{h} (C^+)^2 \quad (40)
 \end{aligned}$$

where

$$\hat{\beta}_{ij} = \left(\frac{(x_{J-(j-2)} - \beta)}{h} \right)^{i-1}, \quad i, j = 1, 2, 3 \quad (41)$$

Numerical Result

Here we will show two examples. Figure 2 gives the numerical solution to the two channels dissipation model with jump $[U] = 0$ (left) and $[cU] = 0$ (right) at $\beta = \frac{5}{9}$, at $\alpha = 1$, at $t = 0$, $t = 0.15$, $t = 0.30$ and $t = 0.40$. The initial data are three humps given by

$$u(x,0) = \begin{cases} \frac{1}{2}(1 + \cos((x - 0.28)\pi / 0.08)) & 0.04 \leq x \leq 0.52 \\ 0 & 0 \leq x < 0.04 \vee 0.52 < x \leq 1 \end{cases}$$

$$v(x,0) = 0.5$$

The wave speeds are $C^- = \begin{pmatrix} 2.5 \\ 3.5 \end{pmatrix}$ and $C^+ = \begin{pmatrix} 1.5 \\ 2.0 \end{pmatrix}$.

Figure 2. gives the solutions computed on grid size $h = 1/200$ and $k = 0.001$.

After reaching the interface, the wave speed changes accordingly on the other sides. For the case of the jump conditions $[U] = 0$, the solutions of $u(x,t)$ is continuous. The magnitude of $u(x,t)$ is not changed by the change of speed at the interface, but the wave length of $u(x,t)$ is changed by the change of speed at in the interface. The magnitude is changed by the term of dissipative $u(x,t)$ and $v(x,t)$. The solution of $v(x,t)$ is also continuous and occurs three humps from the term of $u(x,t)$ in the second equation of two channels dissipation model.

For the case of the jump conditions $[cU] = 0$, the solution of $u(x,t)$ is not continuous at the interface. The magnitude of $u(x,t)$ is changed by the change of speed at the interface, and the wave length is also changed by the change of speed at the interface. The magnitude is also changed by the term of dissipative. The solution of $v(x,t)$ is not continuous at the interface and occurs three humps from the term of $u(x,t)$ in the second equation of two channels dissipation mod

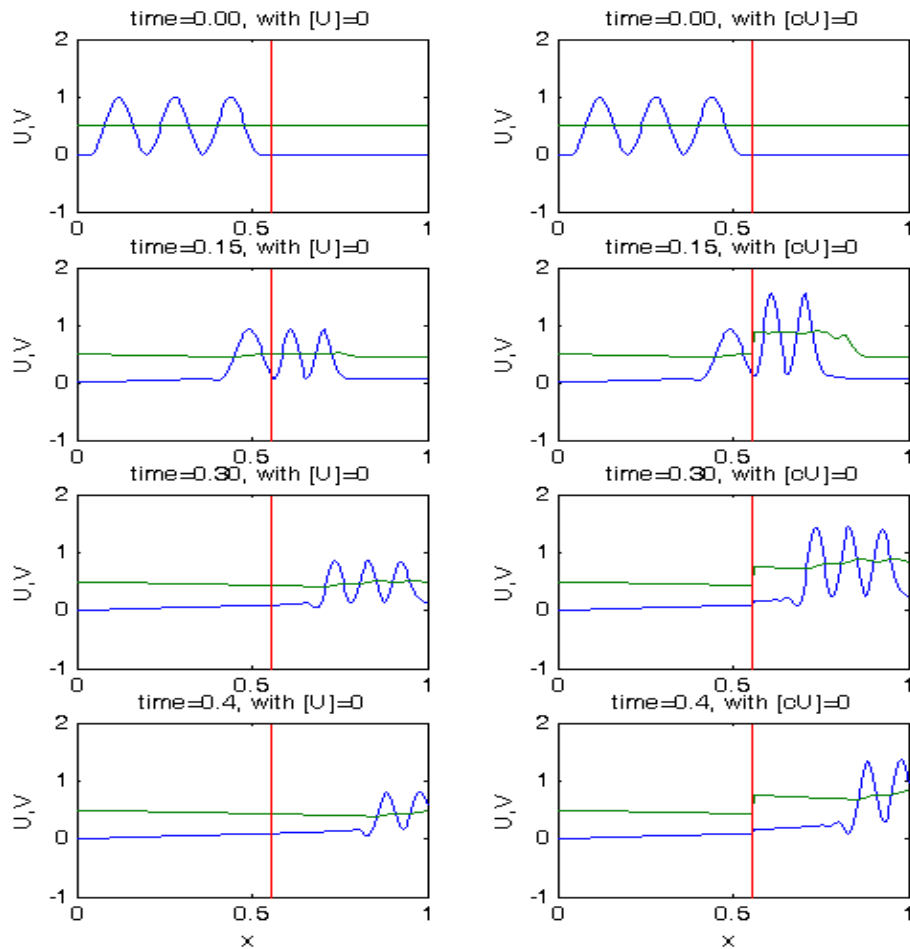


Figure 2 : Numerical solution the two channels dissipation model with jump $[U] = 0$ (left) and $[cU] = 0$ (right) at $\beta = \frac{5}{9}$, at $\alpha = 1$, at $t = 0$, $t = 0.15$, $t = 0.30$ and $t = 0.40$.

REFERENCES

1. Z. Li & R.J. Leveque , *The Immersed Interface Method for elliptic equations with discontinuous coefficients and singular source*. SIAM J. Num. Anal., 31, 1994
2. Sumardi, Suparna D, Lina Aryati, *Convergence of Finite Difference Approximation for Two Channel Dissipation Model*, Proceeding ICAM 2005
3. Sumardi, Suparna D, Lina Aryati, *Fourier Method for Two Channel Dissipation Model*, presented on Workshop in IMS, NUS, 2006
4. van Beckum F.P.H., Travelling wave solution of a coastal zone non-Fourier dissipation model, *Proceeding in Symposium on Coastal Zone Management*, ITB Bandung, 2003
5. C. Zhang , *Immersed Interface Method for hyperbolic system of Partial Differential Equations with Discontinuities coefficients*, PhD Thesis, University of Washington , WA, 1998.

SUMARDI , SOEPARNA DARMAWIJAYA , LINA ARYATI

Department of Mathematics, Gadjah Mada University

Email: mas_mardi@yahoo.com

ESTIMATING THE VALUE OF THE BASIC REPRODUCTION NUMBER IN AN AGE-STRUCTURED SIR HOST-VECTOR MODEL

ASEP K. SUPRIATNA & EDY SOEWONO

Abstract. Basic reproduction number is an important concept in mathematical epidemiology to study the long-term behavior of a disease transmission. This number characterizes the properties of the disease, such as the presence and absence of endemic state and its stability, and the minimum level of disease elimination. However, it is not easy to find the value of this basic reproduction number for practical purposes, since, it is a function of underlying biological and epidemiological parameters which are sometimes unavailable or unobservable. In this paper we estimate the value of the basic reproduction number as a function of available data such as life expectancy and mean age at infection.

Key words and Phrases: Basic reproduction number, host-vector model, age-structured model

1. INTRODUCTION

It is well known in literatures that in a direct transmission disease model the threshold number $\mathcal{R}_0$ can be estimated by the ratio of life expectancy and mean age at infection [2]. Examples are abundances, including [3] who considers a direct transmission in an age-structured model. However, for an indirect transmission model, such as for dengue disease model in which the disease in human is transmitted through mosquitoes, there is a little known if this estimation still holds. In this paper, we will show that, to some extent, if the survival functions of humans and mosquitoes are negative exponential functions and the life expectancy of the mosquitoes is far below the life expectancy of the humans (which is always the case), then this estimation is still valid for the indirect transmission model. The method in this paper gives a practical ways to obtain the magnitude of the basic

2000 Mathematics Subject Classification:

reproduction number. This finding is important since the minimal effort to eliminate the disease depends on the basic reproduction number through the relation $1 - \frac{1}{\mathcal{R}_0}$ [4,7].

2. MATHEMATICAL MODEL

The model discussed here is the generalization of the model in [1,6] to include recovery compartment and indirect transmission of the disease, such as in dengue disease transmission. To formulate the model, let us assume that the host population N_H is divided into three compartments, infectives (S_H), susceptibles (I_H) and recovered assumed immune individuals (R_H). We assume that vectors remain infective for life due to their short life period compared to the duration of the disease, and hence the vector population only comprises two compartments, namely infectives (S_V) and susceptibles (I_V). Recruitment rate for the host and the vector is B_H and B_V , with the death rate μ_H and μ_V , respectively. The disease transmission probability from vector to host and from host to vector is given by β_H and β_V . Suppose that there exist non-negative monotonically non-increasing functions of age, $Q_H(a)$ and $Q_V(a)$, describing the fraction of host and vector population, respectively, who survives to the age of a or more, such that $Q_H(0) = 1$ and $Q_V(0) = 1$. Furthermore, we assume that host life expectancy is finite, and hence $\int_0^\infty Q_H(a)da = L < \infty$, and $\int_0^\infty aQ_H(a)da < \infty$.

Assuming the initial conditions for the host population is given by $N_H(0)$ then we have

$$N_H(t) = \tilde{N}_H(t) + \int_0^t B_H Q_H(a) da, \quad (1)$$

with $\tilde{N}_H(t) = N_H(0)Q_H(t)$. Since the per capita rate of infection in host population at time t is $\beta_H I_V(t)$, the number of susceptibles at time t is given by

$$S_H(t) = \tilde{S}_H(t) + \int_0^t B_H Q_H(a) e^{-\int_{t-a}^t \beta_H I_V(s) ds} da, \quad (2)$$

with $\tilde{S}_H(t) = S_H(0)Q_H(t)e^{-\int_0^t \beta_H I_V(s) ds}$, and if we assume that the rate of recovery is γ then the number of infective host at time t is given by

$$I_H(t) = \tilde{I}_H(t) + \int_0^t B_H Q_H(a) [1 - e^{-\int_{t-a}^t \beta_H I_V(s) ds}] e^{-\int_{t-a}^t \gamma ds} da, \quad (3)$$

with $\tilde{I}_H(t) = I_H(0)Q_H(t)e^{-\int_0^t \gamma ds}$. Furthermore, since $R_H(t) = N_H(t) - S_H(t) - I_H(t)$ then we have

$$R_H(t) = \tilde{R}_H(0) + \int_0^t B_H Q_H(a) [1 - e^{-\int_{t-a}^t \beta_H I_V(s) ds}] [1 - e^{-\int_{t-a}^t \gamma ds}] da, \quad (4)$$

with $\tilde{R}_H(t) = \tilde{N}_H(t) - \tilde{S}_H(t) - \tilde{I}_H(t)$. Equations (2) to (4) constitute transmission dynamics in the host population. Analogously, the transmission dynamics in the vector population is governed by the following equations.

$$N_V(t) = \tilde{N}_V(t) + \int_0^t B_V Q_V(a) da, \quad (5)$$

with $\tilde{N}_V(t) = N_V(0)Q_V(t)$,

$$S_V(t) = \tilde{S}_V(t) + \int_0^t B_V Q_V(a) e^{-\int_{t-a}^t \beta_V I_H(s) ds} da, \quad (6)$$

with $\tilde{S}_V(t) = S_V(0)Q_V(t)e^{-\int_0^t \beta_V I_H(s) ds}$, and

$$I_V(t) = \tilde{I}_V(t) + \int_0^t B_V Q_V(a) [1 - e^{-\int_{t-a}^t \beta_V I_H(s) ds}] da, \quad (7)$$

with $\tilde{I}_H(t) = I_H(0)Q_H(t)$. It is also clear that $\lim_{t \rightarrow \infty} \tilde{N}_V(t) = \lim_{t \rightarrow \infty} \tilde{S}_V(t) = \lim_{t \rightarrow \infty} \tilde{I}_V(t) = 0$. The full model for the host-vector model is given by equations (2) to (4) and (6) to (7). In [7] we analyze these equations and show that there is a threshold epidemic $\mathcal{R}_0$ determining the occurrence of the non-trivial equilibrium point. We also show that this threshold is actually determining the stability of the non-trivial equilibrium point. In the following section we show that the value of the basic reproduction number can be estimated as a ratio of, usually available, demographic and epidemiological parameters: life expectancy L_H and mean age at infection $\bar{a}_H$.

3. ESTIMATION FOR THE BASIC REPRODUCTION NUMBER

We assume that survival functions are given by $Q_H(a) = e^{-\mu_H a}$ and $Q_V(a) = e^{-\mu_V a}$ for humans and mosquitoes, respectively. It is shown in [2] that mean age at infection is given by $\bar{a}_H = \frac{1}{\beta_H I_V^* + \mu_H}$ or $\beta_H I_V^* = \frac{1}{\bar{a}_H} - \mu_H = \frac{1 - \mu_H \bar{a}_H}{\bar{a}_H}$ and similarly by $\bar{a}_V = \frac{1}{\beta_V I_H^* + \mu_V}$ or $\beta_V I_H^* = \frac{1}{\bar{a}_V} - \mu_V = \frac{1 - \mu_V \bar{a}_V}{\bar{a}_V}$. Using a similar method as in [3] we find the basic reproduction number

$$\mathcal{R}_0 = B_H B_V \beta_H \beta_V \int_0^\infty a Q_H(a) \left(\int_0^\infty a Q_V(a) da \right) e^{-\gamma a} da, \quad (8)$$

or equivalently

$$\mathcal{R}_0 = \frac{B_H B_V \beta_H \beta_V \int_0^\infty a Q_H(a) \left(\int_0^\infty a Q_V(a) da \right) e^{-\gamma a} da}{\int_0^\infty B_V Q_V(a) \frac{[1 - e^{-\beta_V I_H^* a}]}{I_V^*} da}, \quad (9)$$

since (I_H^*, I_V^*) is an endemic equilibrium point (see [7] for details). Further calculation gives

$$\mathcal{R}_0 = \frac{[(\mu_H + \gamma)(\gamma \bar{a}_H + 1) + \beta_V B_H (\frac{1}{\mu_V} - \frac{\bar{a}_H \mu_H}{\mu_V})]}{\bar{a}_H (\mu_H + \gamma)^2}. \quad (10)$$

Finally, let us see a special case, that is, if there is no recovery or $\gamma = 0$. Since $\frac{1}{\mu_H} = L_H$ and $\frac{1}{\mu_V} = L_V$, and assuming the vector life expectancy is significantly lower than that for the host, so that $\frac{L_V}{L_H} \rightarrow 0$, then we have a well known relation

$$\mathcal{R}_0 = \frac{L_H}{\bar{a}_H}. \quad (11)$$

This shows that the value of the basic reproduction number can be estimated as a ratio of, usually available, demographic and epidemiological parameters: life expectancy L_H and mean age at infection $\bar{a}_H$.

Acknowledgement. Financial support for this work was provided by the Royal Netherlands Academy of Arts and Sciences (KNAW), via the SPIN Postdoc Program.

1. F. Brauer, A Model for an SI Disease in an Age-Structured Population. *Discrete and Continuous Dynamical Systems- Series B*, **no. 2**, pages 257-264, 2002.
2. O. Diekmann and J.A.P. Heesterbeek, *Mathematical Epidemiology of Infectious Diseases*, John-Wiley & Son, New York, 2000.
3. A.K. Supriatna, Estimating the Threshold Number for Dengue Disease Model in the Presence of Age Structure. A paper presented in *Seminar on Operation Research and Applied Mathematics*, Bandung 27 Aug 2005.
4. A.K. Supriatna & E. Soewono, Critical Vaccination Level for Dengue Fever Disease Transmission. *SEAMS GMU Proceedings of the International Conference on Mathematics and Its Applications*, pages 208-217, 2003.
5. A.K. Supriatna & E. Soewono, A Threshold Number for Dengue Disease Endemicity in an Age-Structured Model. A paper presented in the *International Conference on Applied Mathematics (ICAM05)*, ITB 22-26 Aug 2005.
6. A.K. Supriatna & E. Soewono, A Condition for Stability in an SIR Age Structured Disease Model with Decreasing Survival Rate. A Paper Presented in the *International Conference on Mathematics and Natural Sciences*, ITB 29-30 Nov 2006.
7. A.K. Supriatna & E. Soewono, Analysis of an Age-Structured Dengue SIR Model. *in prep.*

A. K. SUPRIATNA: Department of Mathematics, Padjadjaran University, Bandung, Indonesia.

E-mail: aksupriatna@bdg.centrin.net.id

E. SOEWONO: Industrial and Financial Mathematics Research Group, ITB, Bandung, Indonesia.

E-mail: esoewono@bdg.centrin.net.id

STABILITY ANALYSIS OF FTBS ON SHALLOW WATER EQUATIONS

L.H. WIRYANTO AND WARSOMA DJOHAN

Abstract. A set of shallow water equations is considered as a study case in determining a stability condition of a finite difference method, Forward Time Backward Space. The equations represent gravity wave on the water surface and the horizontal (mean) velocity, as the channel is declined. With that condition, we are able to perform smoothly the propagation of the wave, from the developed numerical procedure.

Key words and Phrases: shallow water equations, Forward Time Backward Space, gravity wave

1. INTRODUCTION

This paper is concerned with developing study in numerical partial different equations, using a finite difference method. Forward time backward space (FTBS) is a simple explicit method to solve a partial different equation, but it is restricted to a condition where the difference equation can become unstable for long run. We derive the stable condition for a set of shallow water equations and observe the numerical solution as the effect of that condition.

As references we can refer Hoffman [1] for the method, and Mei [2] for shallow water equations. Meanwhile relating to the result, Kranenburg [3] worked on similar equations, by including the effect of viscosity, to observe existing roll waves. Needham & Merkin [4] and Merkin & Needham [5] showed rigorously that steady, periodic solutions of the equations exist for $u_0 > 2\sqrt{gh_0}$, where u_0 and h_0 are the uniform velocity and water depth. In contrast, unsteady solutions are possible to obtain in a complement interval of that. We consider in this paper for u_0 near $\sqrt{gh_0}$

2000 Mathematics Subject Classification: 65m06

as restriction of the method FTBS. We perform the solution as wave propagation on the water surface.

In presenting the problem, we analyze the stability condition from linearized equations around uniform quantities. The system of equations is constructed from the difference equations for a monochromatic type solution. For larger time iterations, we expect no increasing the amplitude of the solution. We present this analysis in section 2, and the implement of this work is presented in section 3.

2. FINITE DIFFERENCE METHOD

2.1. Stability condition of FTBS

We consider shallow water equations adopted from the continuity and momentum equations

$$H_t + (UH)_x = 0 \quad (1)$$

$$U_t + UU_x + gH_x = \alpha g - \frac{g}{C^2} \frac{U|U|}{H} \quad (2)$$

describing water motion in a river with water surface height $H = H(x, t)$ and the cross sectional mean water velocity $U = U(x, t)$. x is the longitudinal coordinate along the bottom river having a constant slope α and t is time. The right hand side of the second equation represents gravity and friction forces respectively, with C as the Chezy coefficient.

We suppose that the solution is in form of

$$\begin{aligned} U(x, t) &= u_0 + u(x, t) \\ H(x, t) &= h_0 + h(x, t) \end{aligned}$$

where u_0 and h_0 satisfy

$$\alpha h_0 C^2 = u_0 |u_0| \quad (3)$$

as the constant solution. Mathematically, (3) is used to determine the physical quantity C , for given other quantities. The linear equations of (1) and (2) are

$$h_t + u_0 h_x + h_0 u_x = 0 \quad (4)$$

$$u_t + u_0 u_x + g h_x = \alpha g - \frac{g}{C^2} \frac{u |u_0|}{h_0} \quad (5)$$

When FTBS is applied to (4) and (5), we obtain difference equations

$$h_j^{n+1} = h_j^n - \frac{\Delta t}{\Delta x} [u_0(h_j^n - h_{j-1}^n) + h_0(u_j^n - u_{j-1}^n)] \quad (6)$$

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x} [u_0(u_j^n - u_{j-1}^n) + g(h_j^n - h_{j-1}^n)] \quad (7)$$

Here we neglect the right hand side of (5), and we define $h_j^n = h(j\Delta x, n\Delta t)$, $u_j^n = u(j\Delta x, n\Delta t)$ as the value of the variables at discrete space x and time t . We then suppose the solution as small wave

$$h_j^n = A^n e^{i\Delta x j} \quad (8)$$

$$u_j^n = B^n e^{i\Delta x j}, \quad (9)$$

where $i = \sqrt{-1}$. This type solution is to see the effect of time step in numerical iteration to the amplitude, i.e. the numerical procedure is stable for $|A| < 1$ and $|B| < 1$.

Furthermore, we substitute (8), (9) into (6), (7), and after some algebraic simplifications we write the result in a matrix equation

$$\begin{bmatrix} A^{n+1} \\ B^{n+1} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{11} \end{bmatrix} \begin{bmatrix} A^n \\ B^n \end{bmatrix} \quad (10)$$

where

$$a_{11} = 1 - \frac{\Delta t}{\Delta x} u_0 (1 - e^{-i\Delta x}),$$

$$a_{12} = -\frac{\Delta t}{\Delta x} h_0 (1 - e^{-i\Delta x}),$$

$$a_{21} = -\frac{\Delta t}{\Delta x} g (1 - e^{-i\Delta x})$$

The stable condition, relating to A and B , can be expressed in eigenvalues of the 2x2 matrix in (10), i.e.

$$\lambda = 1 - \frac{\Delta t}{\Delta x} (u_0 \pm \sqrt{gh_0}) - e^{-i\Delta x} \frac{\Delta t}{\Delta x} (u_0 \pm \sqrt{gh_0}).$$

In complex plane, λ is on a circle having center on the real axis $z = 1 - \frac{\Delta t}{\Delta x} (u_0 \pm \sqrt{gh_0})$ and radius $r = \frac{\Delta t}{\Delta x} (u_0 \pm \sqrt{gh_0})$. Therefore, the stable condition, $|\lambda| < 1$, is satisfied by restricting $0 < r < 1$, giving

$$\sqrt{gh_0} < u_0 < \sqrt{gh_0} + \frac{\Delta x}{\Delta t}. \quad (11)$$

Note that we choose negative sign, instead of $\pm$ in the solution of the characteristic eigenvalue equation, to give positive velocity u_0 . Moreover, the condition (11) shows that FTBS can be used to determine the disturbance of supercritical constant solution, i.e. $\sqrt{gh_0} < u_0$.

2.2. Initial and boundary conditions

In solving (4)-(5), initial condition $h(x, 0)$ is requested, and so is $u(x, 0)$ which is related to $h(x, 0)$; similarly the values at the left boundary, as indicated in (6)-(7). Practically, giving values of h is more reliable than u , and can be adopted from the physical condition.

Now we derive the relation between h and u by introducing $\zeta = x - ct$ representing moving coordinates with speed c . In term of this variable, (4) becomes

$$h_0 u_\zeta + (u_0 - c)h_\zeta = 0, \quad (12)$$

and (5) gives

$$gh_\zeta - (c - u_0)u_\zeta = 0.$$

Both equations are then combined by eliminating one of h_ζ or u_ζ resulting

$$c = u_0 + \sqrt{gh_0}.$$

This relation is used to replace c in (12), and it is integrated. The constant of integration is zero since $u, h \rightarrow 0$ for $\zeta \rightarrow \infty$. Therefore, the relation between u and h is

$$u(\zeta = x - ct) = \sqrt{\frac{g}{h_0}} h(\zeta)$$

that can be used to set the initial and boundary conditions of u when $h(x, 0)$ and $h(0, t)$ are given.

3. NUMERICAL RESULTS

The numerical analysis described above is applied in our calculations to obtain solutions. We first examine FTBS for the linearized equations neglecting the effect of gravity and friction forces, that can be in balancing. We calculate for $u_0 = 1$, $h_0 = 0.1$, $g = 9.8$, and we discretize the space and time using $\Delta x = 1$ and $\Delta t = 0.1$, so that they satisfy (11). As the initial condition we set $h(x, 0) = 0$, and the left boundary is

$$h(0, t) = \begin{cases} 0.02t/30, & \text{for } 0 < t < 30; \\ 0.02, & \text{for } t > 30. \end{cases}$$

These conditions can be seen when we open a water gate from $h_0 = 0.1$ to $h_0 + 0.02$ in time interval $[0, 30]$.

Figure 1 is typical water surface of the opening water-gate. We plot the surfaces for some different times in the same coordinates by shifting upward for larger times. We see the front of water surface traveling to the right, and the method is stable. This calculation is then repeated for different values of u_0 , and we found unstable solutions indicated by appearing ripples for $u_0 < 0.9$ on the water surface, as shown in Figure 2.

Now we solve the nonlinear shallow water equations and including the gravity and friction forces. The water gate described above is expanded by inclining the bottom of the channel with angle $\alpha = 0.003$, and we found that the method is stable for $u_0 > 0.97$ calculated up to $n = 1300$ using $\Delta t = 0.1$. For sinusoidal boundary, it generates waves propagating to the right followed by decreasing the

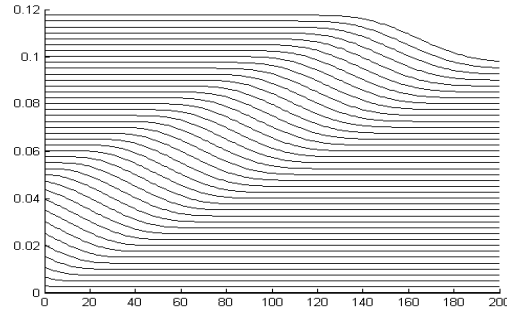


Figure 1: Plot of water surface corresponding to numerical calculation using $u_0 = 1.0$

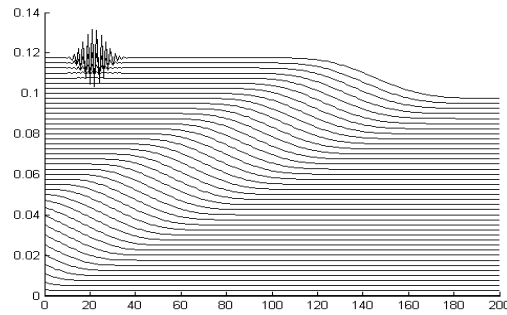


Figure 2: Plot of water surface corresponding to numerical calculation using $u_0 = 0.77$

amplitude further downstream. We show this in Figure 3 as the result from $h(0, t) = 0.02 \sin 0.2t$, $u_0 = 1.5$.

4. CONCLUSIONS

We have presented some numerical solutions of shallow water equations calculated by Forward Time Backward Space (FTBS). The method is restricted for u_0 close to $\sqrt{gh_0}$, otherwise the numerical iteration is unstable. We derive the stability condition based on linearize equations, and this analysis result is then conformed to some numerical calculations for opening of water gate. Other type of solution performed in this paper is wave propagation generated by sinusoidal boundary.

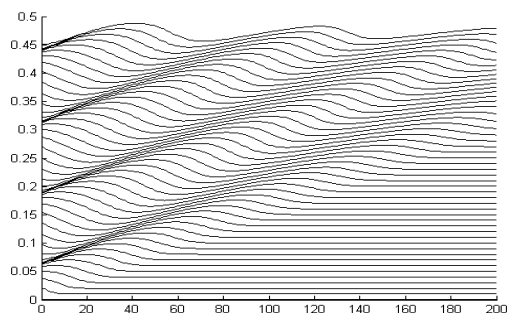


Figure 3: Plot of waves generated by boundary condition $h(0, t) = 0.02 \sin 0.2t$

Acknowledgement. The research and the publication of this paper were partly supported from a research grant of Research Group Bandung Institute of Technology, contract number 039/K01.7/PL/2008, and this is gratefully acknowledged.

REFERENCES

1. J. D. HOFFMAN, *Numerical methods for engineers and scientists*, Mc Graw-Hill, 1974.
2. C. C. MEI, *The applied dynamics of ocean surface waves*, John Willey & Sons, 1982.
3. C. KRANENBURG, On the evolution of roll waves, *J. Fluid Mech.*, **245**, 249–261, 1992.
4. D.J. NEEDHAM AND J. H. MERKIN, On roll wave down an open inclined channel, *Proc. R. Soc. Lond.*, **A 394**, 259–278, 1984.
5. J. H. MERKIN AND D. J. NEEDHAM, An infinite periodic bifurcation arising in roll waves down an open inclined channel, *Proc. R. Soc. Lond.*, **A 405**, 103–116, 1986.

L.H. WIRYANTO AND WARSOMA DJOHAN: Faculty Mathematics and Natural Sciences, Bandung Institute of Technology, Jalan Ganesha 10 Bandung, Indonesia.
E-mail: leo@math.itb.ac.id, warsoma@math.itb.ac.id

THE INITIAL FUNCTIONS IN WAVELET FRAME CONSTRUCTION

MAHMUD YUNUS, HENDRA GUNAWAN, JANNY LINDIARNI

Abstract. In this paper, we characterize the initial functions for cascade algorithm that also satisfy the extension principle in $L^2(\mathbb{R}^d)$. By applying these properties, we elaborate a procedural wavelet frame construction via multiresolution analysis, in multivariate setting, starting from solving refinement equation to get tight multiwavelet frame generators.

Key words and Phrases: multiresolution analysis, multiwavelets, wavelet frame, cascade algorithm, extension principle

1. INTRODUCTION

For d -copy of the set of integers $\mathbb{Z}^d$ and a finite subset Ψ of $L^2(\mathbb{R}^d)$, let

$$\Lambda(\Psi) = \{\psi_{j,k} : \psi \in \Psi, j \in \mathbb{Z}, k \in \mathbb{Z}^d\}.$$

where for all $j \in \mathbb{Z}$ and $k \in \mathbb{Z}^d$ we write $\psi_{j,k} = 2^{jd/2}\psi(2^j \cdot -k)$. We say that $\Lambda(\Psi)$ is a *frame* if there exist two positive numbers A and B such that

$$A \|f\|^2 \leq \sum_{g \in \Lambda(\Psi)} |\langle f, g \rangle|^2 \leq B \|f\|^2$$

for all $f \in L^2(\mathbb{R}^d)$; if $A = B$, then $\Lambda(\Psi)$ is a *tight frame* [2]. Such frames, which in Hilbert spaces are recognized as a generalization of orthonormal bases, play a fundamental role in signal processing, image processing, data compression and transmission, sampling theory and more, as well as being an interesting topic of research in modern and applied mathematics.

2000 *Mathematics Subject Classification:* 42C15, 42C30, 41A25.

Ron and Shen [7] derived the characterization of tight frames and gave sufficient conditions for the construction of tight wavelet frames from MRA (the multiresolution analysis). We know that in application people use the tight wavelet frames due to their efficiencies. Tight wavelet frames are efficient “bases” for separating several high-pass frequency parts from the low-pass frequency part. Also, the redundant property of tight wavelet frame is known to be useful for recovering information from the corrupted one.

In this paper, we use notation $\widehat{f}$ as a Fourier transform of f . We follow the notion of MRA for wavelet in [7] which is slightly a modified version of classical definition of MRA. Let $\phi \in L^2(\mathbb{R}^d)$ be a compactly supported function satisfying the following conditions:

- (i) There exists a trigonometric polynomial H in $e^{i\xi}$ such that

$$\widehat{\phi}(\xi) = H(\xi/2)\widehat{\phi}(\xi/2) \quad (1)$$

If we write $H(\xi) = 2^{-d} \sum h_k e^{i\xi \cdot k}$, for a sequence $\{h_k\}$ in $\ell^2(\mathbb{R}^d)$, then

$$\phi(x) = \sum_{k \in \mathbb{Z}^d} h_k \phi(2x - k), \quad \forall x \in \mathbb{R}^d. \quad (2)$$

- (ii) $\lim_{\xi \rightarrow 0} \widehat{\phi}(\xi) = 1$,

Next, we denote for all $j \in \mathbb{Z}$:

$$V_j = \overline{\text{span} \{2^{jd/2} \phi(2^j \cdot -k) : k \in \mathbb{Z}^d\}},$$

then V_j satisfy all conditions in classical definition of MRA except that ϕ and its shifts constitute a Riesz (or orthonormal) basis of V_0 .

Now, let $\phi \in L^2(\mathbb{R}^d)$ be a compactly supported which satisfies the refinement equation (1). We call H in (1) the *filter function* associated with ϕ . The well known method for constructing wavelet frame is the following condition called *Unitary Extension Principle* (UEP) developed by Ron and Shen in [7]:

$$H(\xi) \overline{H(\xi + \eta)} + \sum_{\ell=1}^N G_\ell(\xi) \overline{G_\ell(\xi + \eta)} = \begin{cases} 1, & \text{if } \eta = 0, \\ 0, & \text{if } \eta \in \{0, \pi\}^d \setminus \{0\} \end{cases} \quad (3)$$

In this case, we have to find trigonometric functions G_ℓ and define (multi)wavelet ψ_ℓ 's by

$$\psi_\ell(\xi) = G_\ell(\xi/2) \widehat{\phi}(\xi/2), \quad \ell = 1, \dots, r \quad (4)$$

An important reformulation of UEP was then given by Daubechies, Han, Ron, and Shen [4], and it is called *Oblique Extension Principle* (OEP). Let $S(\xi)$ be another trigonometric polynomial such that $S(\xi) \geq 0$ and $S(0) = 1$. We look for G_ℓ such that

$$S(2\xi) H(\xi) \overline{H(\xi + \eta)} + \sum_{\ell=1}^N G_\ell(\xi) \overline{G_\ell(\xi + \eta)} = \begin{cases} S(\xi), & \text{if } \eta = 0, \\ 0, & \text{if } \eta \in \{0, \pi\}^d \setminus \{0\}. \end{cases} \quad (5)$$

With these G_ℓ 's we can define wavelet frame generators ψ_ℓ in terms of Fourier transform as in (4).

Although we have UEP and OEP as the methods for constructing tight wavelet frames, we have not known yet any procedural wavelet frame construction using these two methods (see e.g. [2, 8]). Then, using the method of UEP and OEP, we will provide a procedural wavelet frame construction via MRA in multivariate setting. In this paper, we will discuss the first step of procedure, that is, finding the initial functions in *cascade algorithm* that are also satisfying the conditions in extension principles.

2. MAIN RESULT

In order to construct tight wavelet frame via MRA, we start from finding solution of refinement equation (1) for a scaling function ϕ of the MRA, by using the cascade algorithm (see e.g. [1, 3, 5]). In this paper, our focus is in characterization of the initial functions for the cascade algorithm following [6] and adapted to $L^2(\mathbb{R}^d)$.

Starting with a compactly supported function ϕ_0 , we define for $n = 0, 1, \dots$

$$\phi_n := \sum_{k \in \mathbb{Z}^d} h(k) \phi_{n-1}(2 \cdot -k). \quad (6)$$

This relation (6) is called *cascade algorithm* for the filter sequence h . It is shown in [1] that the cascade algorithm always converges to ϕ in distribution if $\hat{h}(0) = 1$; if $\{\phi_n\}$ is bounded in $L^2(\mathbb{R}^d)$, then it converges weakly to $\phi \in L^2(\mathbb{R}^d)$. The L^2 -convergence of the cascade algorithm can be observed from its transition operator as follows.

For any $\psi \in L^2(\mathbb{R}^d)$, we define the sequence

$$[\phi](j) := \int_{\mathbb{R}^d} \phi(x) \overline{\phi(x-j)}, \quad j \in \mathbb{Z}^d \quad (7)$$

The sequence $[\phi_n]$, for each $n = 0, 1, \dots$, is finitely supported, and from (6) and (7) we have

$$[\phi_n](j) := \sum_{k \in \mathbb{Z}^d} h'(2j-k) [\phi_{n-1}](k), \quad j \in \mathbb{Z}^d \quad (8)$$

where $h'(j) = \sum_{k \in \mathbb{Z}^d} h(k) \overline{h(k-j)}$ is the autocorrelation of h . Further, if ϕ is a refinable function, i.e. the solution of (1) or (2), then its autocorrelation $f(x) = \int_{\mathbb{R}^d} \phi(t) \overline{\phi(t-x)} dt$ is the solution of refinement equation

$$f(x) := \sum_{k \in \mathbb{Z}^d} h'(k) f(2x-k). \quad (9)$$

Equation (8) can be rewritten as

$$[\phi_n] = T_{h'} [\phi_{n-1}] \quad (10)$$

where $T_{h'} : \ell^2(\mathbb{Z}^d) \rightarrow \ell^2(\mathbb{Z}^d)$ is a linear transformation such that for every $b \in \ell^2(\mathbb{Z}^d)$,

$$(T_{h'}b)(j) := \sum_{k \in \mathbb{Z}^d} h'(2j - k) b(k), \quad j \in \mathbb{Z}^d. \quad (11)$$

Let Ω be an *invariant support set* for the associated transition operator $T_{h'}$ (see e.g. [5]). The class of all complex sequences which vanish outside Ω is denoted by $\ell(\Omega)$. The operator $T_{h'}$ restricted to $\ell(\Omega)$ will be called the *restricted transition operator*, and can be represented by the following finite order matrix

$$T_{h'} := (2^d h'(2j - k))_{j, k \in \Omega}$$

Following Strang [9], we say that the restricted transition operator $T_{h'}$ satisfies *Condition E* if it has a simple eigenvalue 1 and all its other eigenvalues lie inside the unit circle. A sequence h is called to be *fundamental* if it satisfies

$$\sum_{j \in \mathbb{Z}^d} h(2j) = \sum_{j \in \mathbb{Z}^d} h(2j + 1) = 2^{-d}.$$

By this definition of fundamental sequence, it is easy to see that if h' is an auto-correlation of h and h' is fundamental, then h is also fundamental (see e.g. [9]).

Theorem 2.1 ([9]) *Suppose that h is fundamental. Then the following are equivalent:*

(a) *The cascade algorithm for h converges in $L^2(\mathbb{R}^d)$ for any compactly supported initial function $\phi_0 \in L^2(\mathbb{R}^d)$ satisfying*

$$\sum_{k \in \mathbb{Z}^d} \phi_0(x - k) = 1, \quad x \in \mathbb{R}^d. \quad (12)$$

(b) *The restricted transition operator $T_{h'}$ satisfies Condition E.*

Lemma 2.2 ([9]) *Suppose that h is fundamental and $\{\phi_n\}$ is the corresponding cascade sequence with starting function ϕ_0 . If ϕ_0 satisfies (12) then every ϕ_n satisfies*

$$\sum_{k \in \mathbb{Z}^d} \phi_n(x - k) = 1, \quad \text{for almost all } x \in \mathbb{R}^d. \quad (13)$$

Lemma 2.3 ([5]) *If h is fundamental and $\phi \in L^2(\mathbb{R}^d)$ is compactly supported refinable function normalized such that $\hat{\phi}(0) = 1$, then the shift of ϕ form a partition of unity, i.e.*

$$\sum_{k \in \mathbb{Z}} \phi(x - k) = 1, \quad \text{for almost all } x \in \mathbb{R}^d. \quad (14)$$

Using the facts above, we will prove our main result that is stronger than Theorem 2.1.

Proposition 2.4 *Suppose that h satisfies $\sum_{k \in \mathbb{Z}^d} h(k) = 1$. The following are equivalent:*

- (a) *The cascade algorithm for h converges strongly in $L^2(\mathbb{R}^d)$ for any initial function ϕ_0 satisfying (12).*
- (b) *The operator $T_{h'}$ satisfies Condition E and h is fundamental.*

Proof. Suppose that (b) is satisfied. The cascade sequence $\{\phi_n\}$ converges to ϕ weakly in $L^2(\mathbb{R}^d)$. We need only to show that $\|\phi_n - \phi\| \rightarrow 0$ as $n \rightarrow \infty$.

Since $\int \phi_0 dx = 1$, from the relation (6) we have

$$\int_{\mathbb{R}^d} \phi_n(x) dx = 1, \quad \text{for all } n = 0, 1, \dots \quad (15)$$

Multiplying both sides of (13) by $\overline{\phi_n(x)}$ and integrating gives

$$\sum_{k \in \mathbb{Z}^d} [\phi_n](k) = 1. \quad (16)$$

Since $\int_{\mathbb{R}^d} \phi(x) dx = 1$, using (14) we have

$$\sum_{k \in \mathbb{Z}^d} [\phi](k) = 1. \quad (17)$$

Recall that $[\phi]$ is an eigenvector of the transition operator $T_{h'}$ with eigenvalue 1. Since $T_{h'}$ satisfies Condition E, $[\phi]$ is the unique eigenvector of the eigenvalue 1. Now we have

$$\lim_{n \rightarrow \infty} [\phi_n] = \gamma[\phi]. \quad (18)$$

The relations (16), (17), and (18) give $\gamma = 1$, and (18) gives

$$\|\phi\|^2 = [\phi](0) = \lim_{n \rightarrow \infty} [\phi_n](0) = \lim_{n \rightarrow \infty} \|\phi_n\|^2.$$

Therefore ϕ_n converges strongly to ϕ in $L^2(\mathbb{R}^d)$.

Conversely, suppose that $\|\phi_n - \phi\| \rightarrow 0$ as $n \rightarrow \infty$ for any compactly supported initial function $\phi_0 \in L^2(\mathbb{R}^d)$ satisfying (12). Consider the function $\phi_0^\alpha := \chi_{[-1/2, 1/2]^d}(\cdot - \alpha)$, $\alpha \in \Omega$. With ϕ_0^α as an initial function, the cascade algorithm for h generates a sequence $\{\phi_n^\alpha\}_{n=0}^\infty$ which converges in $L^2(\mathbb{R}^d)$ to ϕ for each $\alpha \in \Omega$. For each $\alpha \in \Omega$ and $n = 0, 1, \dots$ define a sequence b_n^α by

$$b_n^\alpha(k) := \int_{\mathbb{R}^d} \phi_n^0 \overline{\phi_n^\alpha(x - k)} dx, \quad k \in \Omega. \quad (19)$$

Then $b_n^\alpha = T_{h'} b_{n-1}^\alpha$ and $b_0 = \delta(\cdot - \alpha) =: \delta_\alpha$, where $\delta(0) = 1$ and $\delta(\gamma) = 0$ for $\gamma \in \Omega \setminus \{0\}$. It follows that

$$b_n^\alpha = T_{h'}^n b_0^\alpha = T_{h'}^n \delta_\alpha, \quad \text{for all } n = 0, 1, \dots \quad (20)$$

for any $\alpha \in \Omega$. Since $\|\phi_n^\alpha - \phi\| \rightarrow 0$ as $n \rightarrow \infty$ for every $\alpha \in \Omega$, we have

$$b_n^\alpha(j) := \int_{\mathbb{R}^d} \phi_n^0(x) \overline{\phi_n^\alpha(x-j)} dx \rightarrow \int_{\mathbb{R}^d} \phi(x) \overline{\phi(x-j)} dx = [\phi](j), \quad (21)$$

as $n \rightarrow \infty$, for all $j \in \Omega$. Relations (20) and (21) give

$$T_{h'}^n \delta_\alpha \rightarrow [\phi], \quad \text{as } n \rightarrow \infty \quad (22)$$

for every $\alpha \in \Omega$. Since δ_α , $\alpha \in \Omega$, form a basis of $\ell(\Omega)$, it follows that for any sequence $b \in \ell(\Omega)$

$$T_{h'}^n b \rightarrow \gamma[\phi], \quad \text{where } \gamma = \sum_{k \in \Omega} b_k. \quad (23)$$

This implies that the matrix sequence $\{T_{h'}^n\}$ converges. Hence the spectral radius $r(T_{h'}) \leq 1$ and 1 is the only eigenvalue on the unit circle. To show that 1 is a simple eigenvalue of $T_{h'}$, we need only to show that $[\phi]$ is the unique eigenvector (up to multiple of constant) of the eigenvalue 1 of $T_{h'}$.

Let $b \in \ell(\Omega)$ be an eigenvector of $T_{h'}$ corresponding to the eigenvalue 1. Then $\lim_{n \rightarrow \infty} T_{h'}^n b = b = \gamma[\phi]$. Hence $[\phi]$ is the only eigenvector as desired. To show that h is fundamental, it suffices to show that its autocorrelation h' is fundamental.

Recall that for any $\alpha \in \Omega$ and $\delta_\alpha := \delta(\cdot - \alpha) \in \ell(\Omega)$

$$T_{h'} \delta_\alpha \rightarrow [\phi] \quad \text{as } n \rightarrow \infty. \quad (24)$$

Let $v \in \ell(\Omega)$ be a left eigenvector of $T_{h'}$ with eigenvalue 1, then

$$v^T T_{h'}^n = v^T, \quad \text{for all } n = 0, 1, \dots \quad (25)$$

Hence, for each $\alpha \in \Omega$,

$$v(\alpha) = (v^T T_{h'}^n) \delta_\alpha = v^T (T_{h'}^n \delta_\alpha), \quad \text{for all } n = 0, 1, \dots$$

It follows from (24) that

$$v(\alpha) = v^T [\phi] = \sum_j v(j) [\phi](j) =: c, \quad \text{for some constant } c.$$

This shows that $v = ce$ where $e = \sum_{\alpha \in \Omega} \delta_\alpha$. Hence $e^T T_{h'} = e^T$ which is equivalent to

$$2^d \sum_{p \in \mathbb{Z}^d} h'(2p - q) = 1, \quad \text{for all } q \in \Omega.$$

Therefore, h' is fundamental and hence h is fundamental. $\square$

3. CONCLUDING REMARKS

From Proposition 2.4, it is known that if the cascade algorithm converges for a compactly supported initial function $\phi_0 \in L^2(\mathbb{R}^d)$ with $\hat{\phi}_0(0) = 1$ and $[\phi_0] = \delta$, then $T_{h'}$ satisfies Condition E and h is fundamental. Also we can conclude that if $\hat{h}(0) = 1$ and that the cascade algorithm converges for a compactly supported initial function $\psi_0 \in L^2(\mathbb{R}^d)$ with $\hat{\psi}_0(0) = 1$ and $[\psi_0] = \delta$, then the cascade algorithm converges for any initial function $\phi_0 \in L^2(\mathbb{R}^d)$ satisfying (12).

To do wavelet frame construction, then, we can use the above initial function ϕ_0 in order to have scaling function ϕ and its corresponding filter function H . Next, we can use the relation (3) and (5) to find the filter functions G_ℓ 's. Finally, the desired wavelet frame

$$\left\{ 2^{jd/2} \psi_\ell(2^j \cdot -k) : j \in \mathbb{Z}, k \in \mathbb{Z}^d \right\}$$

can be obtained by using (4).

REFERENCES

- [1] A. S. Cavaretta, W. Dahmen, and C. A. Micchelli. Stationary subdivision scheme. *Memoir Amer. Math. Soc.*, 1:1–186, 1991.
- [2] O. Christensen. Recent developments in frame theory. In A.H. Siddiqi, I.S. Duff, and O. Christensen, editors, *Modern Mathematical Models, Methods and Algorithms for Real World Systems*, pages 159–195. Anamaya Publishers, New Delhi, 2006.
- [3] I. Daubechies. *Ten Lectures on Wavelets*. CBMS-NSF Regional Conference Series in Appl. Math. SIAM, Philadelphia, 1992.
- [4] I. Daubechies, B. Han, A. Ron, and Z. Shen. Framelets: MRA-based constructions of wavelet frames. *Appl. Comp. Harmonic Analysis*, 14:1–46, 2003.
- [5] T. N. T. Goodman and S. L. Lee. Convergence of cascade algorithms. In *Mathematical Methods for curves and surfaces II*, pages 191–212. Vanderbilt Univ. Press, Nashville, 1998.
- [6] B. Han and R-Q. Jia. Multivariate refinement equations and subdivision scheme. *SIAM J. Math. Anal.*, 29:1177–1199, 1998.
- [7] A. Ron and Z. Shen. Affine systems in $L^2(\mathbb{R}^d)$: the analysis of the analysis operator. *J. Funct. Anal.*, 148:408–447, 1997.
- [8] M. Skopina. On construction of multivariate wavelets with vanishing moments. *Appl. Comput. Harmon. Anal.*, 20(3):375–390, 2007.

- [9] G. Strang. Eigenvalues of $(\downarrow 2)H$ and convergence of cascade algorithm. *IEEE Trans. Signal. Proc.*, 44, 1996.

MAHMUD YUNUS: Mathematics PhD Student at FMIPA-ITB, Bandung, Indonesia
(permanent address: Department of Mathematics, FMIPA-ITS, Surabaya, Indonesia)
E-mail: yunusm@matematika.its.ac.id

HENDRA GUNAWAN: Analysis and Geometry Group, FMIPA-ITB, Bandung, Indonesia.
E-mail: hgunawan@math.itb.ac.id

JANNY LINDIARNI: Analysis and Geometry Group, FMIPA-ITB, Bandung, Indonesia.
E-mail: janny@math.itb.ac.id

PART IV

Computer Science

TEXTURE FABRIC DEFECT DETECTION SYSTEM UTILIZING IMAGE PROCESSING AND ARTIFICIAL NEURAL NETWORK

SRI HARTATI

Abstract. A prototype of a texture fabric defect detection system has been developed to automate process of recognizing texture fabric defect by utilizing both image processing and feed forward artificial neural networks.

The system consists of image acquisition, pre-processor, feature extractor, response selector and output. The input transducer captures texture fabric image. After a digital image of texture fabric has been acquired, the pre-processor process that image to improve the image in ways that increase the chances for success of the feature extraction, and response selection processes.

The feature extractor performs a function to extract features to be processed in the unit response selector. The image enhancement and image segmentation are performed to prepare an accurate input for response selector.

The unit response selector is in the form of a feed forward artificial neural network that processes the extracted feature. The extracted feature of texture pattern then is processed by processor layer of the artificial neural network to produce information that texture fabric defective or not defective.

The experimental result shows that the system is capable of recognizing different types of texture fabric defects. The texture fabric defects having strong contrast with the background image can easily extracted. Having 70 sample texture images being analyzed, only a false detection is found. The system successfully recognizes different types of texture fabric defects with an accuracy of 98.6%.

Key words and Phrases : image processing, artificial neural network, detection, defect

1. INTRODUCTION

Artificial neural network are analytical systems that address problems whose solutions have not been explicitly formulated. The topology of neural networks is a logical structure in which multiple nodes communicate with each other through synapses that interconnect them. This topology imitates of the structure of biological nervous systems.

Many applications of artificial neural networks focus at 3 major areas such as data analysis, pattern recognition and controlling systems [2]. Artificial neural network applications on data analysis includes evaluation of loan application, analysis of trading commodity, time series forecasting, marketing analysis, pattern of consumer activity, searching and analysis of criminal data [4], and applications on control quality process [7], on physics [11], and on product classification [8, 9]. In the area of pattern recognition, artificial neural network widely used for quality control, industrial inspection where complex pattern recognition problems exist Artificial neural network is widely applied in speech recognition, remote sensing analysis [2] and radar or sonar echoes [8,4], artificial neural network (ANN) is also used in medical domain [12] and in

electrical power [10], it was used to detect a red blood spot dirty spot that occurs in an egg yolk or the surrounding area [9]. Defect detection applications have been developed for many areas [7, 8].

The work here is aimed to develop a prototype of fabric texture defect detection system utilizing both image processing and artificial neural network.

2. RESEARCH METHOD

The fabric defect detection system developed here has components such as shown in Figure 1. These components were developed using the visual programming.

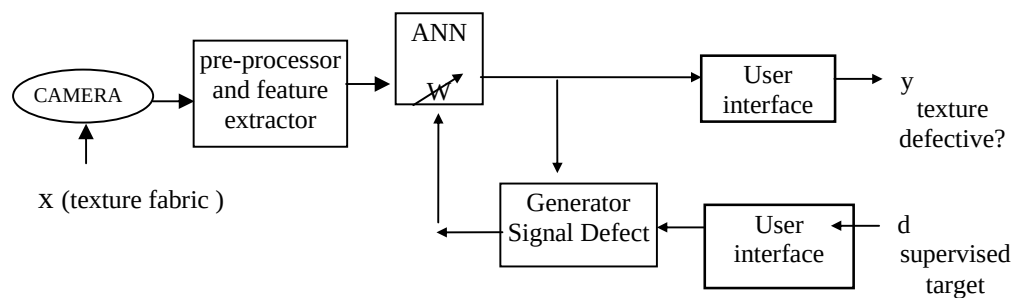


Figure 1: Texture defect detection system

Modules for pre-processing and feature extraction has been developed to process the result of an image acquisition process and to prepare an input to artificial neural network (ANN) system, such as shown in Fig.1. The pre-processing carries out the result of an image acquisition, in order to remove noises occurred during the acquisition. The feature extractor developed here are process of histogram equalization, edge detection, edge linking. The supervised ANN system is utilized to detect fabric detection.

The unit “pre-processor and feature extractor” consists of edge detection and edge linking processes [6]. An edge of an image shows different gray levels of two regions. If a gray level of a fabric defect is known well, for example by extracting the border region between two gray levels then the appropriate defect will be known correctly.

For this purpose, an edge detection process is developed using Sobel operators [6]. A shareholding method is applied to the computed image to determine edges of fabric texture defect. An examples of fabric texture defect is shown in Fig. 2 and the result of edge detection is shown in Fig. 3

The figure 3 shows that there are some noises occurred. These noises seem to be a pattern that similar to the result of edge detection. These noises may contribute errors during defect detection process, due the fact that these noises will be a part of input to an artificial neural network system, and will be processed further. In order to avoid this problem, the unit pre-processor and extractor is provided an edge linking process [6]. This process forms a complete border surrounding the result of the edge detection process. This link covers completely all fabric defect pattern. Figure 4 shows the result of the edge linking process.

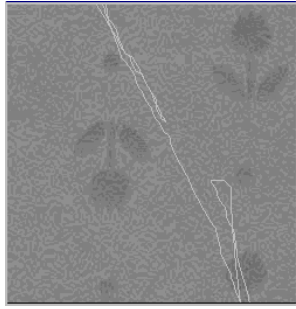


Figure 2: An example of fabric defect

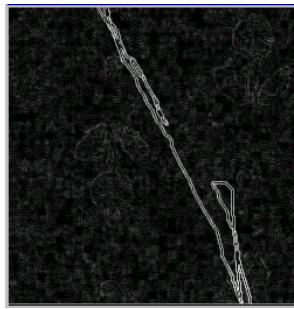


Figure 3: A result of edge detection"

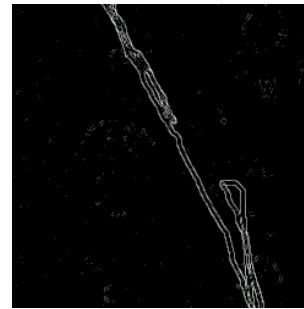


Figure 4: A result of edge linking preprocess

The results of edge detection and edge linking processes are features of fabric texture defect. These features then will be extracted to become a set of input of an artificial neural network (ANN).

3. ARTIFICIAL NEURAL NETWORK FOR DEFECT DETECTION

An appropriate artificial neural network (ANN) model needs to be selected in order to make an efficient computation. One of methods being considered here is to make the number of nodes in layers of ANN as small as possible. For that purposes, the process of feature extraction is carried out by applying of threshold value of 128 to pixel values of the image, and a binary image is produced. The values of the histogram of the pre-processed binary image are normalized to 1, and then they are used as an input to ANN unit.

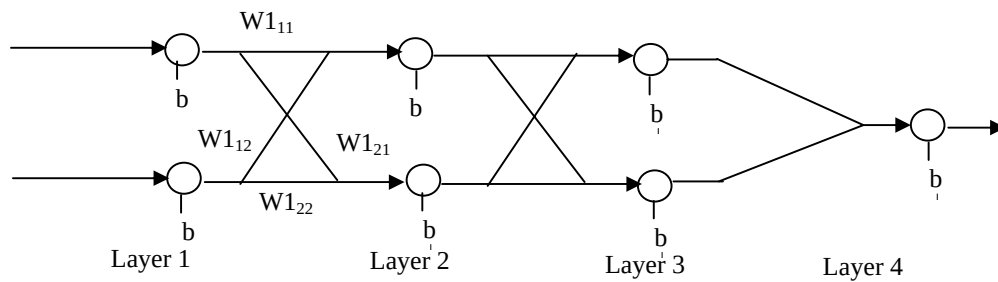


Figure 5: Topology ANN for detecting fabric texture defect

This type of input makes the choice of ANN architecture is quite simple. The selected ANN architecture for defect detection system is feed forward ANN [3, 5], which consist of one layer input with two input nodes, two layers processing with two nodes of each layer, and one layer output with two input nodes such as shown in Figure 5. This network has a topology of fully connected, and consists of (2x2) synaptic weights. These values are adjusted during training process until the values are convergent.

For detecting fabric defect, the ANN unit carries out training and detecting processes. This system unit does training process to learn a set of different defective texture fabric images and a set of non defective texture fabric images. These images are captured during image acquisition. The training

process carried out and followed the algorithm discussed in [2, 4, and 9]. The activation of the neuron or node is formulated as follows

$$O_j = F(\sum W_{ji} O_i - \theta_j) \quad (1)$$

Where W_{ji} is synaptic weight from input node i to output node. O_i and O_j is the input node i and output node j , respectively. The θ_j is "threshold" node, which equals to *bias node * bias weight factor*. F is the activation function, and in this research the sigmoid function is adopted such that

$$F(a) = 1 / (1 + e^{-(a)}) \quad (2)$$

At the beginning of the experiment, the value of W_j and θ_j are set to the small random values. Then the training of the network begins, with the objective of is to differentiate two classes of inputs of defective texture and not defective texture. Specially, the goal is to have the node output $Y(t)$ equal to 1 for defective texture class and 0 for not defective texture class. If the computation results in an output 1 or 0 exactly, the weights are not changed. On the contrary, if the node has not achieved 1 or 0, the synapses weights are updated according to

$$W_{ji}(t+1) = W_{ji}(t) + \Delta W_{ji}(t) \quad (3)$$

ΔW_{ji} is the change in synaptic weight. In this experiment the delta rule is applied

$$\Delta W_{ji} = \eta(T_j(t) - Y_j(t))O_i(t) \quad (4)$$

where represent learning rate constant, generally between [0-1], and δ_j is error gradient at node j , which is

$$\delta_j = T_j - O_j \quad (5)$$

where T_j is a desired activation output, and O_j is the computed output value of node j

The gradient error value for the hidden node is given by equation:

$$\delta_j = O_j(1 - O_j) \sum_k \delta_k W_k \quad (7)$$

where δ_k is the gradient error at the node k and it has a connection of the hidden node j . The iteration processes are repeated until convergent values of synaptic weights are found, and they are achieved whenever the *Error* signal is less than the given tolerance value. This error

$$Error = \frac{1}{2} \sum_j (T_j - O_j)^2 \quad (8)$$

is a result of error accumulation collected from different types of images that are trained by the system.

The repeated iterations include inputting different types of fabric texture images, computing activation functions and updating synaptic weights.

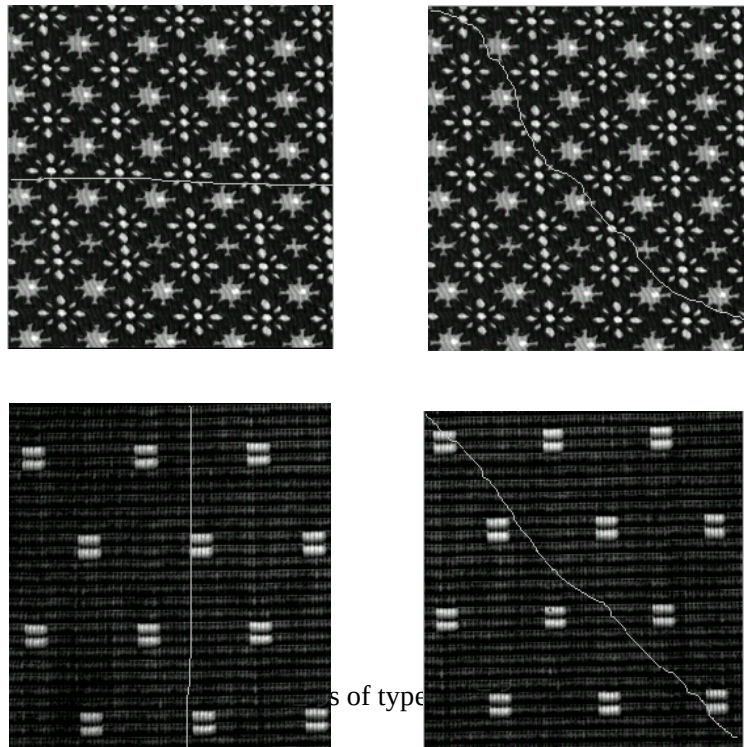
After carrying out the training processes of different types of texture images, the system now is known to be smart and is ready for recognizing whether the fabric texture is defective or not. As a smart system, it generates convergent synaptic weights that are appropriate to detect texture defect.

4. EXPERIMENTAL RESULTS

The system has been implemented using visual programming and tested for detecting a number of different fabric textures. The result of training process for recognizing 10 types of fabrics textures (5 types of defective textures and 5 types of not defective textures) is shown in Table 1. The type of fabric textures are not shown here due to number page limitation, however some examples are shown in Fig. 5, but they are labelled as shown in the table.

Table 1: Results of training process with $\eta = 0.03$, the tolerance = 0.2

	Type of textures	Fabric condition	Input to ANN		Target	Training output
			Node 1	Node 2		
1	A1	Defective	0.9849700	0.0150300	1	0.861
2	B1	Not defective	0.9999200	0.0000800	0	0.195
3	A2	Defective	0.9860800	0.1391600	1	0.828
4	B2	Not Defective	0.9999800	0.0000200	0	0.193
5	A3	Defective	0.9875950	0.0124050	1	0.775
6	B3	Not Defective	1.0000000	0.0000000	0	0.194
7	A4	Defective	0.9851840	0.0148160	1	0.853
8	B4	Not Defective	0.9999400	0.0000600	0	0.193
9	A5	Defective	0.9998993	0.0001070	1	0.713
10	B5	Not Defective	0.9998900	0.0001100	0	0.194



The training process used $\eta = 0.03$, the tolerance value = 0.2. The number of iterations for adjusting the synaptic weight is 901759. The final values of synaptic weights are shown in Table 2.

Table 2: The final values of synaptic weights

	Output node 1	Output Node 2	Synaptic weight W11	Synaptic weight W12	Synaptic weight W21	Synaptic weight W22
Lay. 1 to 2	0.160	0.105	-1.616	-28.981	-2.134	-10.822
Lay. 2 to 3	0.810	0.198	8.240	1.784	-9.993	1.901
Lay. 3 to 4	0.202		-9.367	31.377		

These synaptic weights values are then used for detecting fabric defects. The results of detecting fabric defect using learning rate = $\eta = 0.03$, tolerance value = 0.2 are shown in Table 3.

Another test shows that the system has been trained by 10 types of fabric textures mentioned above, and 30 types of fabric textures which are 10 types of textures used in training process and 20 types of other textures. The result of training process using a learning rate $\eta = 0.04$, tolerance value of 0.2 is tabulated in Table 4.

Table 3: Defect detection results with $\eta = 0.03$, tolerance = 0.2

	Type of textures	Output of ANN	Results of detection
1	A1	0.861472	Defective
2	B1	0.191380	Not Defective
3	A2	0.831393	Defective
4	B2	0.192045	Not Defective
5	A3	0.781937	Defective
6	B3	0.191522	Not Defective
7	A4	0.856100	Defective
8	B4	0.193620	Not Defective
9	A5	0.712694	Defective
10	B5	0.194205	Not Defective

Table 4: Result training process with $\eta = 0.04$, tolerance value = 0.2

	Type of textures	Texture condition	Input to ANN		Target	Output
			Node 1	Node 2		
1	A1	Defective	0.9849700	0.0150300	1	0.890936
2	B1	Not Defective	0.9999200	0.0000800	0	0.196608
3	A2	Defective	0.9860800	0.1391600	1	0.861290
4	B2	Not Defective	0.9999800	0.0000200	0	0.197158
5	A3	Defective	0.9875950	0.0124050	1	0.810851
6	B3	Not Defective	1.0000000	0.0000000	0	0.196649
7	A4	Defective	0.9851840	0.0148160	1	0.985184
8	B4	Not Defective	0.9999400	0.0000600	0	0.198693

9	A5	Defective	0.9998993	0.0001070	1	0.738053
10	B5	Not Defective	0.9998900	0.0001100	0	0.200239

The number of iterations for updating synaptic weights is 648016. The results of defect detection on 10 types of textures mentioned above are tabulated in Table 5.

Table 5: The results of defect detection having $\eta = 0.04$, tolerance = 0.2

	Type of textures	Input to ANN		Output of ANN	Detection results
		Node 1	Node 2		
1	A1	0.9849700	0.0150300	0.890936	Defective
2	B1	0.9999200	0.0000800	0.196608	Not defective
3	A2	0.9860800	0.1391600	0.861293	Defective
4	B2	0.9999800	0.0000200	0.197158	Not defective
5	A3	0.9875950	0.0124050	0.810851	Defective
6	B3	1.0000000	0.0000000	0.196649	Not defective
7	A4	0.9851840	0.0148160	0.985184	Defective
8	B4	0.9999400	0.0000600	0.198693	No defective
9	A5	0.9998993	0.0001070	0.738053	Defective
10	B5	0.9998900	0.0001100	0.200239	No defective

The defect detection results on 20 types of textures having pattern images are stored with different names and they are shown in texture image column in Table 6.

Table 6: Results of defect detection on 20 types of textures

No.	Type of Textures	Input to ANN		Output	Detection result
		Node 1	Node 2		
1	N6M1	0.9902190	0.009781	0.616694	Defective
2	N7M1	0.9861300	0.01387	0.693933	Defective
3	N8M1	0.9848630	0.015137	0.859940	Defective
4	N9M2	0.9939120	0.006088	0.893468	Defective
5	N10M1	0.9998930	0.000107	0.486265	Not Defective*
6	N6	0.9999690	3.051758 10e-5	0.200239	Not Defective
7	N7	0.9999690	1.525879 10e-5	0.197668	Not Defective
8	N8	0.9999985	1.525879 10e-5	0.197158	Not Defective
9	N9	0.9999850	1.525879 10e-5	0.197158	Not Defective
10	N10	0.9999850	1.525879 10e-5	0.197158	Not Defective
11	N1M2	0.9923860	0.0076140	0.574957	Defective
12	N2M2	0.9930110	0.0069890	0.538668	Defective
13	N3M2	0.9925540	0.0074460	0.565259	Defective
14	N4M2	0.9923860	0.0076140	0.574957	Defective
15	N5M2	0.9925840	0.0074160	0.563492	Defective
16	N6M2	0.9925840	0.0075840	0.563492	Defective
17	N7M2	0.9924160	0.0076140	0.573196	Defective
18	N8M2	0.9927370	0.0072623	0.554643	Defective
19	N9M2	0.9922737	0.0072630	0.554643	Defective
20	N10M2	0.9924320	0.0075680	0.572316	Defective

* must be deformed

The system has been tested using learning rate of $\eta = 0.04$, tolerance value of $= 0.1$, and yielding training output such as shown in Table 7, the number required iterations yielding final synaptic weights is 649338.

Table 7: Training results with $\eta = 0.04$, error tolerance $= 0.1$

	Type of textures	Input to ANN		Output	Detection results
		Node 1	Node 2		
1	A1	0.9849700	0.0150300	0.941640	Defect
2	B1	0.9999200	0.0000800	0.142705	Not Defective
3	A2	0.9860800	0.1391600	0.918674	Defective
4	B2	0.9999800	0.0000200	0.143234	Not Defective
5	A3	0.9875950	0.0124050	0.874494	Defective
6	B3	1.0000000	0.0000000	0.142726	Not Defective
7	A4	0.9851840	0.0148160	0.937769	Defective
8	B4	0.9999400	0.0000600	0.144767	Not Defective
9	A5	0.9998993	0.0001070	0.800097	Defective
10	B5	0.9998900	0.0001100	0.146314	Not Defective

The result of defect detection of 20 types of fabric textures using learning rate $\eta = 0.04$, tolerance $= 0.1$ is shown in Table 8.

Table 8: Results of defect detection using $\eta = 0.04$, tolerance $= 0.1$

	Type of textures	Input to ANN		Output	Detection results
		Node 1	Node 2		
1	N6M1	0.9902190	0.009781	0.658885	Defective
2	N7M1	0.9861300	0.013870	0.751721	Defective
3	N8M1	0.9848630	0.015137	0.917572	Defective
4	N9M2	0.9939120	0.006088	0.943495	Defective
5	N10M1	0.9998930	0.000107	0.500920	Defective
6	N6	0.9999690	3.051758 10e-5	0.146314	Not defective
7	N7	0.9999690	1.525879 10e-5	0.143743	Not defective
8	N8	0.9999985	1.525879 10e-5	0.143234	Not defective
9	N9	0.9999850	1.525879 10e-5	0.143234	Not defective
10	N10	0.9999850	1.525879 10e-5	0.143234	Not defective
11	N1M2	0.9923860	0.0076140	0.606135	Not defective
12	N2M2	0.9930110	0.0069890	0.559322	Not defective
13	N3M2	0.9925540	0.0074460	0.593696	Not defective
14	N4M2	0.9923860	0.0076140	0.606135	Not defective
15	N5M2	0.9925840	0.0074160	0.591423	Not defective
16	N6M2	0.9925840	0.0075840	0.591423	Not defective
17	N7M2	0.9924160	0.0076140	0.602972	Defective
18	N8M2	0.9927370	0.0072623	0.580015	Defective
19	N9M2	0.9922737	0.0072630	0.580015	Defective
20	N10M2	0.9924320	0.0075680	0.622753	Defective

The value of learning rate affects the number of training that carried out to obtain the final synaptic weights, such as shown in Table 9, while the value of error tolerance determines the accuracy of defect detection process. The lower error tolerance the more accuracy of defect detection can get. By decreasing error tolerance by 0.1 the false detection shown in Table 6 can be accomplished such as shown in Table 10.

Table 9: Learning rate and the number of training

Learning	The number of training
0.03	901759
0.04	648015

Table 10: Result of false detection

No.	Type of fabric	Input to ANN		Output	Detection results
5	N10M1	0.99998930	0.000107	0.486265	Not defective*

5. CONCLUDING REMARKS

The fabric texture defect detection system is developed here utilizing image processing and artificial neural network method. The performance of the system has been examined for detecting 70 types of fabric textures, using learning rate values of 0.03 and 0.04 and error tolerances 0.1 and 0.2. The system yields to 98.6 % accuracy using the worst parameters in the system (learning rate 0.04 and error tolerance 0.2)

REFERENCES

1. Callan, R. 1999. *The essence of neural networks*. England: Harlow.
2. Chang, E., and Tong, D. *Detecting Symmetry with a Hopfield Net*, Proceedings of the International Joint Conference on Neural Networks, Vol. 2, January, 327-33, 1990.
3. Fausett, L. 1994. *Fundamental of neural networks architectures, algorithms and applications*. New Jersey: Prentice Hall.
4. Fu, L.M. *Neural Networks in Computer Intelligence*, McGraw-Hill, Inc., New York, 1994.
5. Haykin, S. 1999. *Neural Network: A Comprehensive Foundation*. New Jersey: Prentice-Hall.
6. Gonzales, P. *Digital Image Processing*, Addison-Wesley, New York, 1990.
7. Harjoko, A, and Hartati, S, *A Defect Detection Method for Quality Control in Ceramic Tile Industry*, Proceedings The First Jogja Regional Physics Conference, Section E Geophysics and Applied Physics, Sept. 2004.
8. Harjoko, A and Hartati, S, *Pengolahan Citra Digital Berbasis Fitur Untuk Pengklasifikasian Buah-Buahan* (Feature Based Digital Image Processing for

- Classifying Fruits), International Conference on Statistics and Mathematics and Its Application in the Development of Science and Technology, , Bandung. Vol.4 Nomor 1, October 2004.
9. Harjoko, A., and Prayogo *Pemanfaatan Jaringan Tiruan Dan Thresholding Linier Dalam Klasifikasi Telur* (Utilizing Artificial Neural Networks and Linear Threshold in Egg Classification), *Proceeding of the STNI Tarumanegara Conference*, Vol.2 No. 1 . B-14, p.71-78, October 2005.
 10. Matsuda, S. and Akimoto, Y. The representation of Large Numbers in Neural Networks and Its Application to Economical Load Dispatching of Electrical Power", *Proceeding of the International Joint Conference on Neural Networks*, vol. 1, p. 587-592, 1989
 11. Perret, D, 1992, *New Computing Techniques in Physics Research II*, World Scientific, New Jersey.
 12. Wallace, V. P., Bamber, J. C., and Crawford, D. C. 2000. Classification of reflectance spectra from pigmented skin lesions, a comparison of multivariate discriminate analysis and artificial neural network. *Journal Physical Medical Biology* 45: 2859-2871

Sri Hartati: Electronics and Instrumentation Lab., Faculty of Mathematics and Natural Science Gadjah Mada University, Indonesia.
E-mail: shartati@ugm.ac.id

SRI HARTATI

KNOWLEDGE MANAGEMENT SYSTEM BASED ON ONTOLOGY MODEL A CASE STUDY ON PLANT DATA SEARCHING

.HERLINA JAYADIANTI, F.SOESIANTO, ABDUL KADIR AND AZHARI

Abstract The relational model approach is still used broadly in many database and information systems. The main restriction of using the model is its static nature that can not support storing information in more semantically knowledge. On the other hand, if information intend to be considered more formally concrete knowledge as in reality, it's difficult to retrieve the persistent data or information on the various ways of the user need. This paper has researched, analysed, and discussed data modelling by using an ontology model on constructing the data bank of plant information as a knowledge base model. A protege ontology software is used to arrange a model plant ontology in form of classes hierarchies, slots, and instances. The plant ontology model that has built can indicate much more semantics of plant information. In addition, the model can be used easily in form of many way of *user perception that is needed in order to made queries and result*.

Key words and Phrases : *Semantics Data Model, Ontology, Protege, Plant information*

1. INTRODUCTION

Today, as increasingly advance in Internet field, especially web browser, it enables users to get much scientific information, as well as to just having fun. Internet has been a primary requirement for users till now, especially in exchanging information and getting new, inexpensive information quickly. Nonetheless, the constraints begin to be felt as increasingly the number of users and information searchers throughout the world. One of the constraints is, for instance, that desired data are not displayed as required by users. The causes are that saved data were not be able to accommodate the search on various perceptions and meanings used by the users. Programmers and information technology (IT) experts have began to find the way out of these problems, primarily how to save the meanings and semantics contained in information based on intended perceptions. The programmers have been continuing to develop knowledge-oriented documents-saving processes with no changing on document formats and remain to look after semantic and contextual values of saved words (Smith, 1997).

Most of browsers have been using data relational-based saving processes. Relational database have many advantages including implementation amenity, but the saving process has not yet been able to save meanings and contents of a knowledge domain. It is caused by programmers have to keep it away from redundancy and anomaly. In other side, it is an advantage to avoid duplications on database, but it has high possibilities to lose semantic information, so that can not be used for knowledge-searching process from many way of user perception. The other weakness of relational database model is a disability to follow the high speed of knowledge growth (Azhari and Wardoyo, 2005; Bertino and Catania, 2001).

A way to develop *web browser* that able to search data based on many way of user perception is the use of *ontology model*. Currently, there are several *web browsers* already use

ontology model to provide amenity for users in searching information, one of these webs is *Amazon*. Hereinafter, a *web browser* using ontology model is called a *semantic web*, as it is not only able to save solely data, but also to save meanings of the objects.

Ontology model has some advantages such as support to *knowledge management system* and likely to migrate from document-oriented view to relevant knowledge views, and then they could be combined, and reuse more dynamic and flexible. Ontology is a way to represent knowledge regarding object meanings, object properties, and object relations, which is conducive on a certain knowledge domain (Chandrasekaran and Josehson, 1999; Eberhart, 2004; Tamma, 2002).

However, some prior researches found that ontology model is little applied by *web browsers*. For a developer, ontology model is useful in data processing to be knowledge since data were saved in separately *classes* among others, so that adding or eliminating data only has little effect on the data structures. However, since the model is relatively new and need to be further developed, it is not yet to be used broadly.

By present paper, the author tries to develop and manage knowledge based on ontology model. The experimental objective is plant information as it contains much knowledge information can be saved, including species, leaf form, flower configuration, spreading region, and color of flower. By ontology model, user can save data much more semantic of plant information, so that he or she can made query or reuse saved data according to his or her perception of plant information.

Authenticity of the Research

Several researches referred by the author have not studied development of *knowledge management system* using ontology model approach. It can be said that present research is a new study and need to be developed continuously in the future.

As for some references, Noy and McGuinness (2001) had studied on ontology of wine information by which the researchers made answers various questions about wine, while Azhari and Sholichah (2006) have made ontology of information database for scheduling air transport. In Azhari and Sholichah (2006), air transport schedule information are represented as *knowledge management system* model that more semantic in the form of ontology, so that can responds various queries in the same intention of air transport schedule, and also displays routes of the flights.

In present research, it emphasized on implementation of ontology model on plant information, and relationship between ontology model and *First Order Logic*. Ontology model is expected to be used for saving data of plant information and various queries in the same intention of users on plant.

The research is aimed to develop plant information saving model using ontology model, and examines cases of searching semantic model of plant knowledge developed by varying query forms, using protégé. This research is useful for saving plant documents in the form of semantic knowledge in order to make query or reutilize saved data according to perception and knowledge level of the user.

Theoretical Bases Knowledge Management

According to Soesianto (2007), knowledge is a part of information technology besides of data and information. Knowledge as a component of technology is one's awareness to his/her environment. Computer-based information technology is explained by refer to Figure 1 Soesianto (2007).

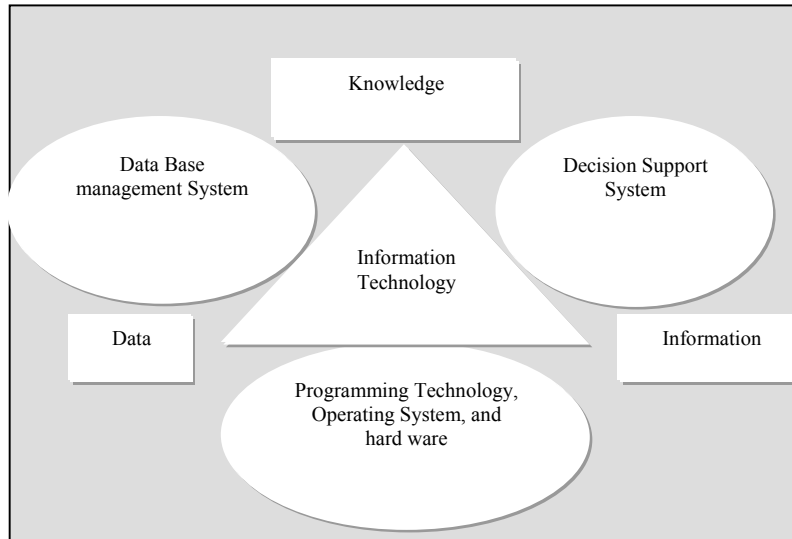


Figure 1. Information Technology

Referring to Hendrik (2003), knowledge is data and information combined with ability, intuitive, experience, idea, and motivation derived from competent sources. Hendrik (2003), then, classified knowledge into three classes, these are:

1. Tacit Knowledge

Information would be tacit knowledge when processed by one's mind.

2. Explicit Knowledge

Knowledge that can be formulated in the form of words or formula, and it can be directly transferred to the other person completely.

3. Shared Knowledge

Explicit knowledge used together in a community.

Knowledge management process can be presented as shown in **Figure 2** (Wiryana and Hasibuan, 2003).

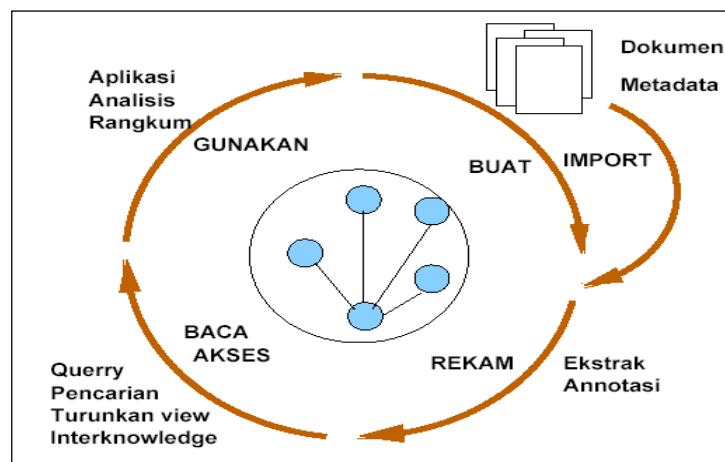


Figure 2. from Document to Knowledge

Sequence of steps in knowledge management as follows:

1. **Creation or Knowledge import**
Contents of knowledge management are made and converted from other inputs, so that corresponds with existing system.
2. **Recording (Capture)**
Knowledge should be determined whether it is important or not to the organization.
3. **Access (Retrieval) and Reading**
This step provides answers for knowledge searching by user.
4. **Reusing**
User will process knowledge received for further reusing.

1.2.2 Ontology Concept in Computer

Ontology growth begins from *Artificial Intelligence*, and then it expands to *Knowledge Engineering*, *Natural Knowledge*, *Knowledge Management*, and *e-commerce*, primarily in interoperability and integration of data (Bertino and Catania, 2001).

In Artificial Intelligence (AI) field, ontology has two senses related one another. First ontology is representation vocabulary that is commonly majored for a certain domain or discussion subject. Second, ontology is a *body of knowledge* for explaining a certain discussion, commonly by *common sense knowledge* domain and *representation vocabulary* (Noy and McGuinness, 2001).

According to Benjamins (2002), ontology constitutes a definition of basic sense and relation vocabulary from an area as the order of combination between term and relation to define vocabulary. Following with Gruber (1993), ontology is an explicit specification of conceptualism. Borst (1997) has modified Gruber's definition by suggesting that ontology is formal specification of received conceptual (*share*).

In accordance with Chandrasekaran (1999), ontology is a theory of object meaning, object property, and object relation, which is conducive on a certain knowledge domain. Noy and McGuinness (2001) argued that ontology is important as it can be used to explain structure of a discipline. Ontology enables to likely migrate from document-oriented view to relevant knowledge, then these are combined for usage in the future. Maedche and Staab (2002) stated that ontology is a formal conceptual of a certain domain using together in a community.

Eberhart (2004) found that currently many people have awareness in importance of *knowledge management system* for getting more benefits from knowledge and to conserve knowledge assets. A primary question in knowledge management system is how to integrate much knowledge, which spread across various spaces in an organization. It might be solved by constructing expert system to acquire knowledge and represent it into simple format enable for all components of a organization.

According with Bertino and Catania (2001), ontology model is a theory regarding existence of all kind of entities defined in a domain and stated in a semantic network graph consisting of class hierarchy, slot, and instance, which is expected to support in knowledge engineering for constructing knowledge management system.

1.2.3 Ontology and First Order Logic

Logic is an order for anyone in determining the truth/authenticity of a statement or argument. The order is also referred as inferential made in common form and independent on specific things or specific languages used.

One of the orders is *First Order logic*, which is representation language enabling a programmer for expressing required knowledge to translate into programming language and to be saved. Knowledge representation can be carried out by *Quantifier*, that is, symbol by which a set or horizon of variable is expressed in a logic expression. Quantifier is divided into two; first, *universal quantifier* ($(\forall x)p(x)$) means that nature of p owned by any x in its pronounce context.

Second, *existential quantifier* $(Ex)p(x)$ means that nature of p at least owned by an x in its pronounce context.

1.2.4 Components of Ontology Model

In accordance with Guarino (1998), supporting components to be focused in constructing ontology model, these are; 1) *concept* and *class* are used in the broadly sense. A concept could be a thing said, so that could also be explanation of task, function, action, strategy, and others, 2) *relation* is representation of a type from interaction among concepts of a domain, 3) *functions* or *Mother-of* is a particular relation in which element of n of unique relation to element of $n-1$, 4) *axioms* are used in modeling a correct *sentence*, 5) *instances* are used to represent elements.

In sum, ontology components consist of *class*, *instance*, and *slot* (Noy and McGuinness, 2001).

1.2.5 Basic Principles to be followed in Constructing Ontology Model

Basic criterions should be concerned in constructing ontology model are: 1) *clarity* and *objectivity*, ontology should be prepared with meanings which are term definitions prepared as objective definition in a documentation of natural language, 2) *completeness*, a definition is stated in important, adequate term, 3) *coherence*, consistency in defining, 4) *consistency*, it is possible to cross-check with contradictive of a valid definition, 5) minimizing semantic distance to *sibling* concept, 7) minimizing commitment ontology, 8) standardizing the name as it is possible.

1.2.6 Architecture of Ontology Model

Maedche and Staab (2002) proposed that there are four main components in architecture of ontology, these are:

1. **Management Ontology.** As a fundament to learn ontology comprehensively with ability of ontology arrangement and graphic display, it supports to user in processing manually ontology design.
2. **Resource Process.** This component contains techniques to discover, import, analyze, and implement relevant data inputs.
3. **Algorithm Library.** It acts as *algorithmic* of framework. A number of algorithms are provided to summarize and maintain entities in ontology model.
4. **Coordination.** Ontology user utilizes this component to connect management ontology, resource, and algorithm library. Architecrture of Ontology is shown in following **Figure 3**.

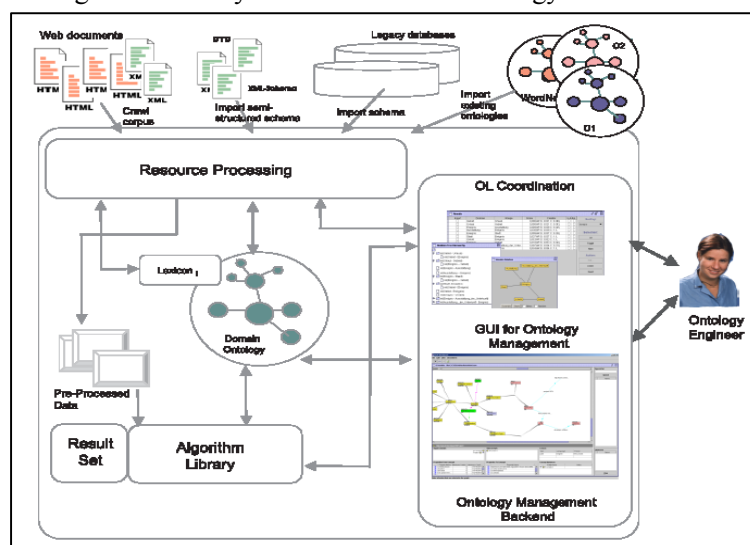


Figure 3. Conceptual Architecture of Ontology Model

1.2.7 Process of Ontology Model Development

Following with Noy and McGuinness (2001) and Yao and Etzkorn (2005), ontology development process could be explained as follows:

1. **Determination of Domain.** It is a first step of digitalized process of knowledge conducted by answering some competent queries.
2. **Reutilization of Ontology.** It is a step of reutilization and justification of constructed ontology.
3. **Reeling of Terms.** It is a process of determining terms used to make statement or explain the same thing.
4. **Defining Class and Class hierarchy .** It creates some definitions from concept of hierarchy and then elaborates property of the concept (Noy and McGuinness, 2001)
5. **Defining Property.** In general, there are some kinds of object property (nature) that these could be slot in ontology.
6. **Defining Constraint.** Slot is possible to have different type of value, appropriate value, cardinality, and other slot values.
7. **Instancing.** Instantiation is filling of slot and defining of individual class.

1.2.8 Software for Constructing Ontology Model

Software might be used to represent knowledge are Ontoedit, Protégé2000, OILEd, Ontolingua, Appollo, WebOnto and WebODE. Software for constructing ontology model in this paper used protégé, with protégé, user can model display of acquisition knowledge, enter knowledge domain, make *form*, and enter data. Protégé helps user in constructing knowledge-based ontology domain easily. Protégé is able to visualize relation of *subclasses of tree*, and support in constructing *multiple inheritances*.

1.2.9 Hypothesis

By several book references discussed early, hypothesis of present research is that ontology model is able to save knowledge more semantically, so that it's able to provide answers for queries proposed by users with different perceptions or viewpoints to intended object. This hypothesis will be examined and proved by experimental process of this paper.

1.3 Methodology Of Research

The research carried out by following steps:

a. Experimental Tools and Materials

Tools and materials prepared for examining are plants, ontology protégé software, computer Intel(R) Pentium 4 (R) 3.00 GHz 512 MHz of RAM with platform of Windows XP.

b. Steps and Way of Research

This experiment is carried out in several steps as shown in **Figure 4** below.

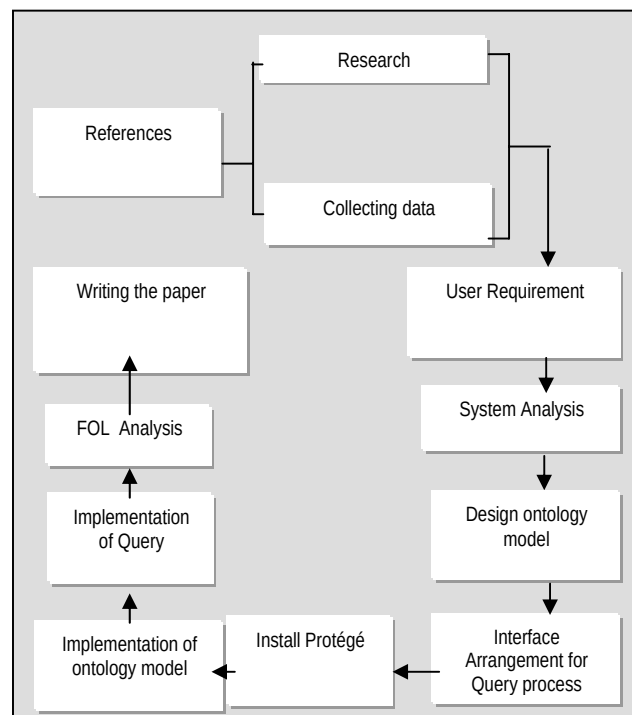


Figure 4 Experimental Process

Description of steps and way of the research are explained as follows:

- 1) **Preparation (stocktaking of coverage and information need)**
Initial data and information are taken by book references. On this step, tools preparation is also accomplished, including preparing computer and required software.
- 2) **Analysis**
Analysis and arrangement of plant knowledge are corresponded with ontology phases.
- 3) **Arrangement of Ontology Model**
Arrangement of ontology model is constructed by using principles of ontology model development as cleared in Sub-section 2.7.
- 4) **Interface Arrangement for Query process**
Arrangement of query result examination using protégé might be made by selecting desired class in accordance with user's knowledge.

2. MAIN RESULTS

2.1 Implementation of Class and Class Hierarchy

To model classes and compiling them hierarchically, first is making of new class and named with class of area, class of leaf, class of flower, class of root, class of soil, class of plant media, class of environment, and class of plant species classification. All class Hierarchies are shown in **Figure 5** below.



Figure 5. Class Hierarchy

2.2 Implementation of Property-Slot

To represent property (slot) of class, it is needed to fill slot value, including name, cardinality, type, and constraint (facet). An example of **Figure 6**, it shows slot of area class and slot is filled with the country that has value of symbol, cardinality 1 (single), it means that each country has at least one name.

Template Slots			
Name	Cardinality	Type	
NEGARA	single	Symbol	

Figure 6. Slot in Class of Area

2.3 Implementation of Instance

An example of making individual instance is making of *area class instance* as shown in **Figure 7**. To fill instance in class of Asia Tenggara, it is carried out by filling text in *instance editor*.

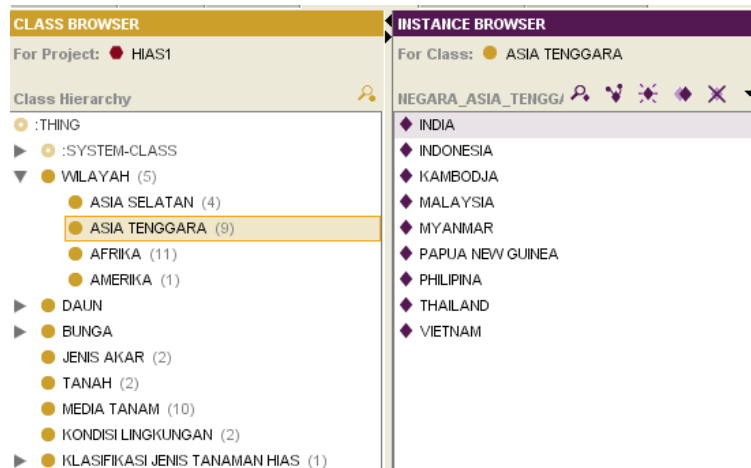


Figure 7. Instance in Class of Asia Tenggara

Visual Display of Ontology Model

Relation and Synchronization of class and instance that has been implemented in protégé could be seen on panel of Jambalaya. Panel of Jambalaya might represent class and node, and slot (property) is visualized in a clearly display, so that user can see relation occurred in constructed ontology model, whether or not that relation as according with desired, as shown in **Figure 8** below.

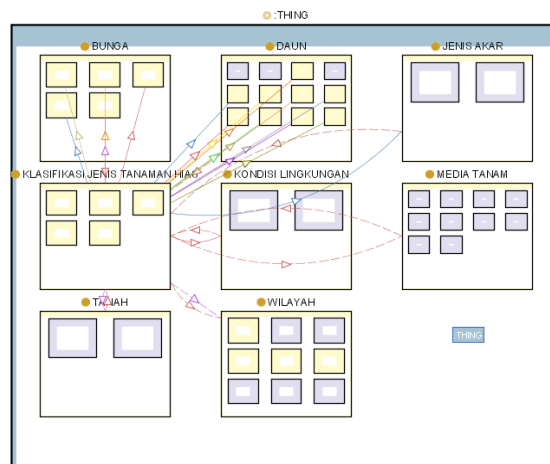


Figure 8. Graph Visualization Display of Ontology Model of Plant Species Classification

2.5 Implementation of Query

User 1: Adenium Somalense is a species of Adenium, not a hybrid of Adenium.

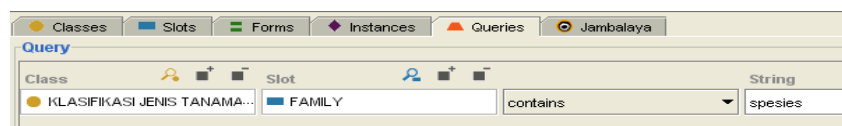


Figure 9. Searching for Adenium Somalense based on Family

Answers: Obesum, Socotranum, Swazikum, Oleifolium, Somalense, Mutiflorum, Arabicum, Boehmianum

The answers are seen as shown in **Figure 10**

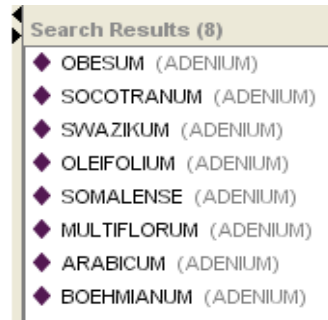


Figure 10. Answers of Adenium Somalense based on *Family*

User 2: Adenium Somalense is adenium with leaf likes lancet.

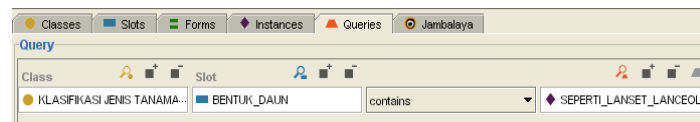


Figure 11. Searching for Adenium Somalense based on *Leaf Form*

Answers: Somalense, Black Ruby, Calypso Stripe, Big Mama, Arabicum, Boehmianum, Andromeda.

The answers are seen as shown in **Figure 12**.



Figure 12. Answers of Adenium Somalense based on *Leaf Form*

2.6 Analysis of Query by First Order Logic

Query 1: Adenium Somalense is a species of Adenium, not a hybrid of Adenium

Formation:

- Classification of plant species (name of flower, family)
- Family (Species, Hybrid)
- Name of Flower (Obesum, socotranum, somalense, mutiflorum, boehmianum, arabicum, swazikum, oleifolium, Two Tones, Black Ruby, Calypso Stripe, ... Red surath)
- Species (Obesum, Socotranum, Swazikum, Oleifolium, Somalense, Mutiflorum, Arabicum, Boehmianum)

R (name of flower):=

Species (σ species=hybrid: Classification of Plant Species)

Answers: Obesum, Socotranum, Swazikum, Oleifolium, Somalense, Mutiflorum, Arabicum, Boehmianum

3. CONCLUDING REMARKS

3.1 Conclusions

From discussion above, it can be taken by following conclusions:

1. Knowledge management system of plant classification might be built using plant data modeled ontologically, in the form of classified hierarchy as *semantics class, subclass, slot, and instance*.
2. This Ontology Model of Plant classification can be accessed and used to answer the query semantically in the same intention in different way by protégé software.
3. The use of ontology model can be made in the form of semantic database to modify the query.
4. Information and knowledge can be flexibly, detailed displayed. It is enabled as ontology model can be tracked from various class name, subclass, slot, attribute, or relation, and its result represents relation link of class, subclass, slot, attribute, instance, or their combinations.

3.2 Suggestion

The author suggests that future researches should develop and expand ontology model of plant classification in the form of *semantic web*, so that it is more easily to be used

REFERENCES

1. Azhari dan Sholichah, M., 2006, Model Ontology untuk Informasi Jadwal Penerbangan Menggunakan Protégé, Jurnal Informatika, vol7(1). Hal 67-76, Mei 2006, Jurusan Teknik Informatika, Fakultas Teknologi Industri, Universitas Kristen Petra, Surabaya.
2. Azhari dan Wardoyo, R., 2005, Pendekatan Model Data Semantis pada Sistem BasisData Cerdas, Prosiding Seminar Nasional Revitalisasi Penelitian Bidang Matematika dan Ilmu Pengetahuan Alam, FMIPA UGM, Yogyakarta.
3. Benjamins R., RB, 2000, *Knowledge System Technology: Ontologies and Problem Solving Methods*.
4. Bertino, E. dan Catania, B., 2001, *Intelligent Database Systems*, ACM Press, Addison-Wesley, Pearson Education, London.
5. W.N. Borst, , 1997, *Construction of Engineering Ontologies*, Phd Thesis, University of Twente, Enschede.
6. Chandrasekaran, B. dan Josehson, J., 1999, *What are Ontologies, and Why Do We Need Them?*.
7. Eberhart, A., 2004, *Ontology-based Infrastructure for intelligent Applications*, Disertasi Doktor, der Naturwissenschaftlich-Technischen Fakultät I der Universität des Saarlandes, Saarbrücken.
8. T.R.Grubner, 1993, *A translation approach to portable ontologies*, *Knowledge Acquisition*.
9. N.Guarino, 1998, *Formal Ontology in Information Systems, Proceedings of FOIS'98*. Hal. 3-15, 6-8

- June, IO Press, Italy.
Hendrik., 2003, *Sekilas Tentang Knowledge Management*, <http://www.ilmukomputer.com/populer/hendrik/hendrik-km-zip> (accessed 21/Mei/2006).
10. A. Maedche, , dan S. Staab, 2002, *Ontology Learning*, <http://www.aifb.uni-karlsruhe.de/Forschungsgruppen/WBS> (accessed 8/Mei/2007).
 11. N. F. Noy,, dan McGuinness, D. L., 2001, *Ontology Development 101: A Guide to Creating Your first Ontology*, <http://www.ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness.pdf> (accessed 8/Mei/2007).
 12. P. Smith, , 1997, *An Introduction to Knowledge Engineering*, International Thompson Computer Press.
 13. F. Soesianto, 2007, Matematika dan Komputasi dalam Teknologi Informasi, Pidato Pengukuhan Jabatan Guru Besar Fakultas Teknologi Industri, Universitas Atma Jaya, Yogyakarta.
 14. V.A.M.Tamma, , 2002, *An Ontology Model Supporting Multiple Ontologyes for Knowledge Sharing*, PhD Thesis, University of Liverpool.
 15. Y. Tijerino, , 2003, *Ontology Generation From Tables*, In *proc. Of the 4th Internasional Conference on Web Information System Engineering (WISE 2003)*, Rome, Italiy.
 16. I. M. Wiryana, dan E. Hasibuan, 2002, Menuju Ontology Pendukung Pengembangan Kelautan Indonesia, *Workshop Kelautan*, Hamburg.
 17. H.Yao, dan I. Etzkorn, 2005, *Cohesion Metrics for Ontology Design and Application*.

.Herlina Jayadianti

Student on Department of Electrical Engineering Faculty of Engineering , Gadjah Mada University
e-mail:Jayadianti@lycos.com

F.Soesianto: Department of Electrical Engineering Faculty of Engineering
Gadjah Mada University

Abdul Kadir: Department of Electrical Engineering Faculty of Engineering
Gadjah Mada University

Azhari: Semantic Web Research Group, Computer Science
Mathematics and Natural Science Faculty, Gadjah Mada University
e-mail: arisn@ugm.ac.id

A CASE STUDY OF XML TO RELATIONAL DATABASE TRANSLATION USING XML-ENABLED DATABASE

ANNY KARTIKA SARI

Abstract. This paper illustrates an example of the unexpected results in translating XML document into relational database. Some guidance concerning a creation of Data Type Definition (DTD) whose XML documents are to be translated into relational database is then discussed. The guidance is much related to the presented case study to give a concrete example of transferring an XML document into one of XML-enabled database using the export facility provided by one of XML editor tools.

Key words: DTD, XML Document, XML-enabled database, XML-to-Relational database Translation

1. INTRODUCTION

Sometimes it is said that XML and its surrounding technologies constitutes a “database” in the sense that it is a sort of database management system. Nevertheless, the possibility of using it as a database will fail in most environments that have many users, strict data integrity requirements, and the need for good performance [1]. The requirements are important in the e-business applications. Thus, for e-business applications that use XML as the basis of structure, it is better to find the more proper database rather than use the XML document itself as a database.

However, to make the XML document which can be translated properly into database is not easy. If we are not careful, the database produced will not conform to the database we want. An unexpected result in translating XML document into relational database is presented in this paper. Furthermore, some guidance of transferring the XML document and its Data Type Definition (DTD) into relational database in order to get the database that fits in the requirements is discussed.

An article by Bourret [2] discusses the mapping of DTDs to databases. In his paper, Bourret mentioned two types of mappings, i.e. table-based mapping and an object-relational (object-based) mapping. He gave description and examples of each type of mapping. In this paper, the method of mapping used is the table-based mapping.

The XML to database translation is commonly used in e-business applications to enable other non-XML applications to use the data. Besides, for some people, accessing and processing the data from the database is easier than from an XML document. Some technologies are available for transferring data from XML document to database. The technology used in this paper is the XML-enabled database, specifically Microsoft Access. XMLSPY tool is also

used to transfer data from an XML document into an Access database. One case study will be presented in this paper to give a concrete example of translating an XML document into relational database. The simple example is depicted from an e-business application, i.e. a database of orders in a book shop.

2. XML TO RELATIONAL DATABASE TRANSLATION

Translating an XML document into database seems to be easy. However, if we only simply transfer the XML document without considering whether the structure is suitable to be transferred into relational database, sometimes the database that is produced does not fit in our requirements. In this section, an example of the unexpected result in the translation is shown. After that, some guidance in making the DTD for XML documents to be translated into relational database will be presented.

2.1 An Example of the Unexpected Results in the Translation

Consider the example of an XML document depicted in Figure 1.

```
<shop>
  <book>
    <bookID>S0001</bookID>
    <title>XML How to Program</title>
    <year>1</year>
    <author>Deitel </author>
    <type>science</type>
    <publisher>Prentice Hall</publisher>
    <price>50</price>
    <stock>10</stock>
  </book>
  <customer>
    <custID>0001</custID>
    <name>
      <lastName>Sari</lastName>
      <firstName>Anny</firstName>
    </name>
    <address>
      <city>Enschede</city>
      <street>Beatrixstraat</street>
      <number>109</number>
      <postCode>7511 KR</postCode>
    </address>
    <phone>0645382588</phone>
    <email>sari@cs.utwente.nl</email>
    <account>034714577</account>
  </customer>
</shop>
```

Figure 1. An example of an XML document that produces some unexpected tables

The document in Figure 1 is made to record the data in a shop. The data to be recorded is the data about the book and customer of a shop, which are represented by the *book* and the *customer* elements. Since the name of the customer consists of last name and first name, the *name* element has 2 child elements, i.e. *lastName* and *firstName*. With the same reason, the *address* element has 4 child elements, i.e. *city*, *number*, *postCode*, and *street*. It seems just usual, nothing will cause problem. The result expected is two tables, i.e. *customer* and *book*. The *customer* table is expected to contain the complete *name* and *address*.

However, when we see the result of the translation using Microsoft Access import facility, something strange happens, especially in the *customer* table. There are three tables produced, i.e. *customer* table, *name* table, and *address* table, as shown by Figure 2. This means that an element, regardless whether it is parent element or child element, will be automatically mapped into a table. Thus, in creating a child element, we have to consider the possible result as the effect of the rule.

customer					
PrimaryKey	ForeignKey	account	custID	email	phone
1	0	34714577	1	sari@cs.utwente.nl	645382588

name			
PrimaryKey	ForeignKey	firstName	lastName
1	1	Anny	Sari

address					
PrimaryKey	ForeignKey	city	number	postCode	street
1	1	Enschede	109	7511 KR	Beatrixstraat

Figure 2. The Access tables produced in the translation process

The above example shows that to make an XML document that can be translated into relational database, we have to consider many things in order that the document fits in our requirement. Some guidance in making a DTD for XML document to be translated into relational database will be explored in the next section.

2.2 The Guidance in Creating a DTDs for XML Documents to be translated into Relational Database

A DTD (Document Type Definition) defines an XML document's structure [3]. It determines what elements, attributes, etc. are permitted in the document. In order that an XML document can be translated into relational database, its DTD should define the structure of the document that conforms to the requirement of the translation. In this section, some guidance that can be used to construct a DTD as a validator for its document so that the XML document can be translated into database correctly, or at least it will be suitable with our need, is presented. However, most of the guidance is intended to the use of XMLSPY for translating an XML document into MS Access.

Below is the guidance in creating a DTD for XML document in order that the documents can be translated into relational database:

- Elements with the same name will be contained in one table.
Elements with the same name will be mapped into one table, even though they are child elements. Thus, the creation of an element must be done carefully in order to avoid unexpected tables. It is better to keep only 1 child element for each element, i.e. the child element to represent the column values.
- Use attributes to represent the column values of the tables, except the translation is done by XML-enabled database such as Microsoft Access.
Actually, the use of attributes to represent the values has an advantage. If

we choose to store table column values as the content of elements, rather than attributes, we will need to be on the look out for possible name collisions between columns in different tables with the same name. This is not an issue when using attributes. Moreover, attributes provide more useful types than element. For instance, attribute provides the tokenized type that enables the DTD to check the uniqueness of the values of the ID type attribute. Nevertheless, the representation of the column values using the content of elements is still possible as long as there is no name collision, and it is suitable when we use XML-enabled database import facility, like what Microsoft Access provides, to transfer the document into database. In the translation of XML document into Access database using the import facility of MS Access, attributes will be ignored. This has to be taken into account when one wants to translate an XML document into Access using the import facility provided by Access. The data has to be child elements. For this approach, it is important to define the occurrence indicators for the content of element. If the column can be empty, the optional suffix (?) is used; if the column must have value, the optional suffix cannot be added. For the representation using attributes, the attribute default needs to be defined. The *#REQUIRED* default is suitable for a column that does not allow nulls, whereas the *#IMPLIED* default is suitable for a column that allows nulls.

- c. For attributes that are expected to be keys, use attributes with *ID* types.
To uniquely identify each element, we can create an attribute for each element that can distinguish the element from others. In the translation into a database, this attribute will become primary key of the table. Moreover, the attribute must be of type *#REQUIRED*. The establishment of the *ID* attribute is also important for the element that needs to be referred to by other element, which is known as foreign key in relational database. A unique *ID* will need to be created for each instance of an element that is generated. We need to ensure that these *IDs* are unique not only across all elements of a specific type, but across all elements in our document. One way to do this is to use a number prefixed by the name of the element or the initial name of the element in which it appears.
- d. Use *IDREF* type attribute for elements that refer to other elements.
Sometimes, an element has an attribute whose value is obtained by another element. In the term of relational database, this is known as foreign key of a table that refers to a primary key of another table.
- e. Use *enumerated* attributes for choices.
Sometimes a column of tables only allows particular values. To provide such requirement, in the DTD that defines the document, we can use *enumerated* attributes. For instance, shipping methods allowed for an order are postal service, federal service, and UPS. Thus, the element or attribute of shipping must contain one of the three methods.

3. CASE STUDY: THE DATABASE OF ORDERS IN A BOOK SHOP

For case study, we will discuss the database of orders in a book shop. We imagine that the book shop firstly get the data using an XML document. For further processing, such as calculating the total profits of the shop, the data should be transferred into a database. The database used here is Microsoft Access. We use XMLSPY to translate the data to Access database, and not use the import facility provided by Access. We will show each step to be taken in creating the DTD as well as the XML document such that they provide the suitable structure to be translated into a database. The creation of DTD follows the guidance in section 2.2.

3.1 Creating the Elements Included in the Document

The database is intended to save the data of the customers who order books from the shop. We use attributes to represent the contents of the elements, as guided by point b in the section 2.2. Furthermore, the name of each element should be different from other elements, since elements with the same name will be contained in one table, as point a of section 2.2 says. The elements that should be defined in the document are as follow.

- a. *Customer* element, which contains the information about the customer.
The information included here is the customer ID, the name of the customer, the address, the account number, the email address, and the phone number. Furthermore, the name consists of last name and first name, while address consists of street, street number, postcode, and city. However, each of the content should be defined as other attributes of *Customer* element, since we have discussed that the creation of, say, name element with the child elements last name and first name will produce another table that is unexpected.
- b. *Book* element, which contains the information about books.
The information that should be included in this element is book ID, title, type, author, publisher, year, price, and stock. The book ID attribute is used to identify uniquely book with particular title and year. Thus, the real primary key is the title and year. The type attribute describes the type of the book, such as novel, science, biography, etc. The stock attribute represents the number of the books that are still available in the shop.
- c. *Shipping* element, which contains the information about shipping methods that can be used.
The information included in this element is the shipping ID, method, the city, and the cost. The method describes the method to be used. The city represents the destination city where a book has to be sent.
- d. *Order* element, which contains the information about orders.
The information included in this element is order ID, customer ID, book ID, number of items, shipping ID, status of payment, and status of order. Again, the order ID is used to distinguish each order. The customer ID refers to the customer ID in the Customer element, while the book ID refers to the book ID in the Book element. The number of items represents the number of books the customer ordered. The shipping ID refers to the shipping ID attribute in the Shipping element. The status of payment attribute shows whether the order has been paid by the customer. And the status order attribute shows whether the order has been sent. Further, in the database, we can add one more important field, i.e. total cost, that is calculated from the price of the book, the number of the book, and the method to be used.

However, this needs some query which is not discussed here.

3.2 Determining the default and the type of the attributes

Since we use XMLSPY for translating data from XML document to MS Access, the use of attributes to represent the data is possible, as point b of section 2.2 says. Since all the information contained in the document is represented by the attributes, all elements, except the root element, can be defined as *EMPTY*. The default and the type of each attributes are defined using the guideline in section 2.2, especially point c, d, and e. Below are the definition of the default and the type of each attribute.

1. customer ID in the *Customer* element, named as *customerID*
Since this is an attribute to uniquely identify the element, the default must be *#REQUIRED*, and *ID* type is the most suitable type to be used.
2. last name in the *Customer* element, named as *lastname*
This attribute must appear in the element. Thus, the default should be *#REQUIRED*. The type can be *CDATA*.
3. first name in the *Customer* element, named as *firstname*
This attribute can be left out for a customer who has only last name. The default can be *#IMPLIED*, with *CDATA* type.
4. street in the *Customer* element, named as *street*
This attribute is very important for sending the order, and thus cannot be left out. The default should be *#REQUIRED* with the type of *CDATA*.
5. street number in the *Customer* element, named as *number*
Again, this attribute is very important for sending the order. The default should be *#REQUIRED* with the type of *CDATA*.
6. post code in the *Customer* element, named as *postcode*
With the same reason, it should be of *#REQUIRED* default and *CDATA* type.
7. city in the *Customer* element, named as *city*
Other than for sending process, this attribute is also used to determine the cost of the sending, later in the database. Thus, the default should be *#REQUIRED* and the type is *CDATA*.
8. account number in the *Customer* element, named as *account*
This attribute cannot be left out because it is used to bill the customer. Thus, the default is *#REQUIRED* with the type of *CDATA*.
9. email address in the *Customer* element, named as *email*
This attribute must be filled. The default should be *#REQUIRED* with the type of *CDATA*.
10. phone number in the *Customer* element, named as *phone*
This attribute is needed, since it is possible that the shop owner needs to inform something important to the customer. Thus, the default is *#REQUIRED* with the type of *CDATA*.
11. book ID in the *Book* element, named as *bookID*
Since this is an attribute to uniquely identify the Book element, the default must be *#REQUIRED*, and *ID* type is the most suitable type to be used.
12. title in the *Book* element, named as *title*
This attribute must appear in the element. Thus, the default should be *#REQUIRED*. The type can be *CDATA*.
13. type in the *Book* element, named as *type*
This attribute can be filled by some choices. The choices that are provided are science, fiction, biography, politics, art, and sport. Thus, this must be enumerated attribute with the type values of science, fiction, biography, politics, history, art, and sport. The default value is science.

14. author in the *Book* element, named as *author*
Since this attribute must not be left out, the default must be *#REQUIRED*, and *CDATA* is the suitable type to be used.
15. publisher in the *Book* element, named as *publisher*
This attribute must appear in the element. Thus, the default should be *#REQUIRED*. The type can be *CDATA*.
16. year in the *Book* element, named as *year*
This attribute cannot be left out. Thus, this must be of *#REQUIRED* default. The type can be *CDATA*.
17. price in the *Book* element, named as *price*
Again, this attribute must not be left out. The default should be *#REQUIRED* with the type of *CDATA*.
18. Stock in the *Book* element, named as *stock*
This is a required attribute. The default should be *#REQUIRED* with the type of *CDATA*.
19. shipping ID in the *Shipping* element, named as *shippingID*
Since this is an attribute to uniquely identify the element, the default must be *#REQUIRED*, and *ID* type is the most suitable type to be used.
20. method in the *Shipping* element, named as *method*
This attribute can be filled by some choices. The choices that are provided are standard, priority, and high priority. Thus, this must be *enumerated* attribute with the type values of *standard*, *priority*, and *high*. The default value is set to *standard*.
21. city in the *Shipping* element, named as *city*
This attribute must appear in the element. Thus, the default should be *#REQUIRED*. The type can be *CDATA*.
22. cost in the *Shipping* element, named as *cost*
This attribute cannot be left out. The default must be *#REQUIRED*, with the *CDATA* type.
23. order ID in the *Order* element, named as *orderID*
This attribute is to distinguish the element. The default should be *#REQUIRED* with the type of *ID*.
24. customer ID in the *Order* element, named as *customerIDREF*
This attribute refers to the *customerID* attribute from the Customer element. It must not be left out. Thus, the default should be *#REQUIRED* with the type of *IDREF*.
25. book ID in the *Order* element, named as *bookIDREF*
The attribute refers to the *bookID* attribute from the Book element. It should appear in the element. The suitable default is *#REQUIRED*, and the type is *IDREF*.
26. number of items in the *Order* element, named as *items*
The attribute should not be left out. The default should be *#REQUIRED* with the type of *CDATA*.
27. shipping ID in the *Order* element, named as *shippingIDREF*
This attribute refers to the *shippingID* attribute from the Shipping element. It must not be left out. Thus, the default should be *#REQUIRED* with the type of *IDREF*.
28. status of payment in the *Order* element, named as *paid*
This attribute can be filled in by some choices, i.e. yes, means that the order has been paid, and no. Thus, this should be an *enumerated* attribute with the choices of *yes* and *no*. The default is set to *no*.

29. status of order in the *Order* element, named as *orderstatus*

Again, this attribute should be *enumerated* because it can be filled in by some choices, i.e. *yes*, means that the order has been sent, and *no*. The default is set to *no*.

3.3 Creating the DTD that defines the document

Figure 3 shows the DTD that has been created which is based on the elements and attributes that have been determined in the previous section. There are 4 elements in the DTD, i.e. *Customer*, *Book*, *Shipping*, and *Order*.

```
<!ELEMENT ShopOrder ( Customer+, Book+, Shipping+, Order)>
<!ELEMENT Customer EMPTY>
<!ATTLIST Customer
    customerID ID #REQUIRED
    lastname CDATA #REQUIRED
    firstname CDATA #IMPLIED
    street CDATA #REQUIRED
    number CDATA #REQUIRED
    postcode CDATA #REQUIRED
    city CDATA #REQUIRED
    account CDATA #REQUIRED
    email CDATA #REQUIRED
    phone CDATA #REQUIRED
>
<!ELEMENT Book EMPTY>
<!ATTLIST Book
    bookID ID #REQUIRED
    title CDATA #REQUIRED
    type ( science | fiction | biography | politics | history | art | sport ) "science"
    author CDATA #REQUIRED
    publisher CDATA #REQUIRED
    year CDATA #REQUIRED
    price CDATA #REQUIRED
    stock CDATA #REQUIRED
>
<!ELEMENT Shipping EMPTY>
<!ATTLIST Shipping
    shippingID ID #REQUIRED
    method ( standard | priority | high ) "standard"
    city CDATA #REQUIRED
    cost CDATA #REQUIRED
>
<!ELEMENT Order EMPTY>
<!ATTLIST Order
    orderID ID #REQUIRED
    customerIDREF IDREF #REQUIRED
    bookIDREF IDREF #REQUIRED
    items CDATA #REQUIRED
    shippingIDREF IDREF #REQUIRED
    paid ( yes | no ) "no"
    orderstatus ( yes | no ) "no"
>
```

Figure 3. The DTD used to define the document for the shop orders.

3.4 Creating the XML Document that conforms to the DTD

The next step is creating XML document that conforms to the DTD for the shop orders. Due to the large space required for the XML document produced, the XML document is not presented in this paper.

3.5 Translating the XML Document into Relational Database

There are four tables produced from the translation of the document into Microsoft Access database. The tables are depicted by Figure 4.

Customer									
account	city	customerID	email	firstname	lastname	number	phone	postcode	street
770482357	Enschede	C0001	sari@cs.utwente.nl	Anny	Sari	109	645382588	7511 KR	Beatrixstraat
476233890	Enschede	C0002	erfianto@cs.utwente.nl	Bayu	Erfianto	109	645382567	7509 DE	Matenweg
825377322	Hengelo	C0003	fachry@cs.utwente.nl	Khairun	Fachry	80	645381344	7411 GT	Julianstraat
566794566	Enschede	C0004	pulungan@cs.utwente.nl	Reza	Pulungan	210	645382568	7510 BH	Emmastraat
76588754	Hengelo	C0005	jliu@cs.utwente.nl	Jin	Liu	6	645382587	7412 GA	Javaststraat

Book							
author	bookID	price	publisher	stock	title	type	year
Deitel	S1001	80	Prentice Hall	10	XML How to Program	science	2001
Kurose	S1002	75	Addison Wesley	15	Computer Networking	science	1999
Lewis	S1003	50	Addison Wesley	7	Java Software Solutions	science	2000
Peterson	S1004	65	Morgan Kaufmann	4	Computer Networks	science	2000
Fraknoi	S1005	30	Saunders College	8	Voyages through the Universe	science	1997
Ernest Hemingway	F2001	10	Arrow	3	For Whom the Bell Tolls	fiction	1994
John Grisham	F2002	5	Arrow	5	The Partner	fiction	1997
Gabriel Garcia Marquez	F2003	7	Penguin Books	4	One Hundred Years of Solitude	fiction	1970

Shipping			
city	cost	method	shippingID
Enschede	5	standard	S01
Enschede	8	priority	S02
Hengelo	7	standard	S03
Hengelo	10	priority	S04
Hengelo	12	high	S05
Almelo	12	standard	S06
Almelo	15	priority	S07
Almelo	17	high	S08

Order						
bookIDREF	customerIDREF	items	orderID	orderstatus	paid	shippingIDREF
S1001	C0001	1	O0001	no	yes	S01
S1001	C0002	1	O0002	no	no	S02
F2001	C0004	1	O0003	yes	yes	S01
S1001	C0003	1	O0004	no	yes	S04
S1005	C0001	1	O0005	yes	yes	S02

Figure 4. The relational tables produced from the translation of XML document

The name of each table in the relational database is the same as the name of each element in the XML document as well as in DTD. The records in the *Customer* table represent the *Customer* elements of the document, the records in the *Book* table represent the *Book* elements of the document, the records in the *Shipping* table represent the *Shipping* elements in the document, and the records in the *Order* table represent the *Order* elements of the *ShopOrder* document. The number of the fields of each table corresponds to the number of attributes of each element to be represented. However, the tables are still native, means that there has been no sufficient equipment like the usual database. For instance, the primary key has not been assigned, although in the design of the document the *ID* of each element has been established. Also, the navigation, i.e. which table refers to which table or which table function as look-up table, has not been assigned yet. This will be the business of the database. Further, it is possible that some queries are needed in order that the database can fulfil the requirement of the data for the book shop.

4. DISCUSSION

The XML-to-database translation has some advantages as well as some drawbacks. The first advantage is that the translation into database will make the data easier and quicker to be accessed because there is no parsing and text conversion, which has to be done whenever the data is not transferred into database. Furthermore, the presentation of the data by the 'real' database will provide some facilities, such as efficient storage, indexes, security, transactions and data integrity, multi-user access, triggers, queries across multiple document, etc. Those cannot be found in the XML documents, at least not as good as the database. Thus, while it may be possible to use an XML document or documents as a database in environments with small amounts of data, few users, and modest performance requirements, this will fail in most production environments, which have many users, strict data integrity requirements, and the need for good performance [1].

The other advantage is that the data transferred to database, in this case the Access database, can be translated back into the XML document easily using the export facility provided by Access. However, the translation must be further processed in order to get the XML document that is exactly the same with the original one. Using the Access export feature, each field is by default translated into a child element of a parent element which represents the record. Thus, the XML document produced by the default translation will be different to the original XML document. This translation is not done well using the import facility provided by XMLSPY, but is performed well using the export facility provided by Access.

One of the drawbacks that can be found from the translation is that the translation requires a particular structure of the document that is translated into the database, as discussed in the second chapter. The error structure will be brought into the database that will make the database not meet our requirement.

The guidance to create a DTD for XML documents to be translated into database presented above can also be used to make guidance for creating XML Schema for XML documents with the same intention. The concepts used in the DTD can be used in XML Schema because most concepts defined for both document are similar, except that every XML Schemas is also an XML document. The creation of guidance for creating XML Schema for XML documents to be translated into database will benefits more to the problem of XML into relational database translation.

The XML-to-database translation discussed so far uses traditional database to store the data for further processing. The database enables the translation from an XML document into the database. This type of database is called the XML-enabled database. Some examples of XML-enabled databases are Microsoft Access (developed by Microsoft), as we have seen in the case study, DB2 (developed by IBM), Cache (developed by Inter System), FileMaker (developed by FileMaker), and SQL Server (developed by Microsoft) [4]. Another alternative of storing data obtained from XML document is using the native-XML database. A native-XML database is a database designed especially for storing XML documents. The database is specialized for storing XML data and stores all components of the XML model intact [5]. Documents go in and documents come out in the database. Examples of native-XML databases are eXtensible Information Server (XIS, developed by eXcelon), dbXML (developed by dbXML Group), 4Suite (developed by FourThought), and Birdstep RDM XML (developed by Birdstep) [4].

One of the differences between XML-enabled and native-XML database is that XML-enabled database uses schema-specific structures that must be mapped to the XML document at design time. Native-XML database uses generic structures that can contain any XML document. Actually there are still some other types of database related to the translation, such as Middleware, XML Servers, and Content Management System, but we do not discuss them in this paper.

Among the two types of database, i.e. the XML-enabled database and the native-XML database, the XML-enabled database is more suitable for data-centric documents. Data centric document is characterized by fairly regular structure and a little or no mixed content [4], such as the ShopOrder document. The data itself is more interesting than the documents. Thus, the important part to be stored is only the data, and not the document. The use of the XML-enabled database is much more appropriate than native-XML database. Moreover, since most of native-XML databases can only return the data as XML, this will not be practical for data-centric documents in which the data might be processed using another format. If the application needs the data in another format, it must parse the XML before it can use the data.

The other type of XML documents is document-centric document, which is usually designed for human consumption and has several characteristics from its nature [6]. Examples are books, email, advertisements, and almost any hand-written XHTML document. They are characterized by less regular or irregular structure, larger grained data (that is, the smallest independent unit of data might be at the level of an element with mixed content or the entire document itself), and lots of mixed content [1]. Document-centric documents are usually written by hand in XML or some other format, such as RTF, PDF, or SGML, which is then converted to XML. Unlike data-centric documents, they usually do not originate in the database. Thus, the XML native database is suitable for storing this type of documents.

5. CONCLUSION

XML-enabled database is one of the technologies for translating an XML document into database. This technique is easy to be performed, except that we need to make a document such that it fulfils the requirements of the documents that can be translated into database. For this necessity, a DTD needs to be defined. The guidance of creating the DTDs has been presented in this paper and has been proven to work properly in the case study.

Using Microsoft Access import facility to translate XML document into Access database, individual data values must be in child elements and table/column names must match element names. In this way, attributes are ignored. This will not be the case when we use XMLSPY to transfer the data from the document into Access.

Compared to the native-XML database, the XML-enabled database is more suitable to store the data obtained from data-centric documents XML, since the data is the most important part in the document. Thus, the XML-enabled database is suitable for XML documents applied in e-business application, in which the documents tend to be data-centric.

For future works, a semi-automatic tool can be developed for translating XML documents into relational database. The tool should give choices to the users in the translation so that the users can determine the rules of the translation, e.g. which attributes should be translated into fields, by themselves. This tool will be an alternative of the tools for such translations available today, such as XML editor and XML-enabled database, in which the translation is done automatically by the tools without any consideration to the users' requirements.

REFERENCES

1. R. Bourret, "XML and Databases", 2005.
Available at: <<http://www.rpbouret.com/xml/XMLAndDatabases.htm>>, last accessed: May 2007.
2. R. Bourret, "Mapping DTDs to Databases", 2001.
Available at: <<http://www.xml.com/pub/a/2001/05/09/dtdtodbs.html>>, last accessed: April 2007.
3. H.M. Deitel, P.J. Deitel, *XML How to Program*. Upper Saddle River, NJ: Prentice Hall, 2001.
4. R. Bourret, "XML Database Products", XML and Databases, 2005.
Available at: <<http://www.rpbouret.com/xml/xml/XMLDatabaseProds.htm>>, last accessed: May 2007.
5. K. Staken, "Introduction to Native XML Databases", O'Reilly xml.com, 2001.
Available at: <<http://www.xml.com/pub/a/2001/10/31/nativexmlldb.html>>, last accessed: April 2007.
6. T. Tashiro, "An Approach to Realizing User-Friendly Authoring Tool for Document-centric XML", Tech Conference, 2007.
Available at: <[2007/xtech.org/public/asset/attachment/169](http://2007.xtech.org/public/asset/attachment/169)>, last accessed: April 2007

Anny Kartika Sari: Computer Science Study Program, Gadjah Mada University, Indonesia.

E-mail: anny_sari@gadjahmada.edu

TAR-QL: An SQL-like Query Language for Querying Discovered Temporal Association Rules

EDI WINARKO

Abstract. Many rule discovery algorithms in data mining generate a large number of rules, often greatly exceeds the size of the underlying database, and only a small fraction of that large volume of rules is of any interest to the user. It is generally understood that interpreting the discovered rules to gain a good understanding of the domain is an important phase of the knowledge discovery process. However, when there are a large number of generated rules, identifying and analyzing interesting rules become difficult. Providing the user with a list of rules ranked by their confidence and support might not be a good way of organizing the set of rules as this method would overwhelm the user. Therefore, to be useful, a data mining system must manage the large amount of generated rules by offering flexible tools for further selecting rules. In order to address this post-processing problem we propose a retrieval system to facilitate the selection of interesting rules. A query language called TAR-QL is proposed for specifying the criteria of rules to be retrieved.

Temporal rules, query language, retrieval:

1. INTRODUCTION

Finding interesting rules is difficult task when the number of generated rules is large. Several approaches for post-processing discovered association rules have been reported. One approach is to group ‘similar’ rules among the discovered rules [15, 9, 10, 11]. This approach works well for a moderately large number of rules. However, for a larger number of rules it will produce too many clusters.

A more flexible approach is by identifying the rules that are of special importance to the user through templates or *data mining queries*. This approach can

complement the rule grouping approach. Templates have been used to specify interesting and uninteresting classes of rules (association and episode rules) [7]. The importance of data mining queries has been highlighted by Imielinski and Mannila (1996) by introducing the concept of Inductive Databases, in which Knowledge and Data Discovery Management Systems (KDDMS) manage KDD applications just as DBMSs successfully manage business applications. In addition to allowing the user to query the data, KDDMS also give the users the ability to query patterns, rules, and models extracted from these data.

Several query languages have been proposed in the literature that provide the user the ability to query the data, for example, Mine-Rule [13], DMQL (Han et al. 1996), and OLE DB [14]. However, these languages are not designed for querying the discovered rules, but for generating rules from the data. A more powerful data mining query language called MSQL is proposed, which can be used not only for rule generation, but also for querying the discovered rules (using SelectRules operator) [6]. One of the main limits of the four languages is the weak support of rule post-processing. Their post-processing capabilities are limited to a few predefined built-in primitives. Tuzhilin & Liu (2002) have proposed Rule-QL for querying multiple sets of association rules and developed efficient algorithms for processing the queries.

This paper describes a query language TAR-QL for facilitating the selection of interesting rules during the post-processing of a large set of discovered richer temporal association rules. The query language specification is based on the subset of Rule-QL, extended by adding several additional operators and functions to deal with richer temporal association rules. The paper is organized as follows. In Section 2 the general framework of the post-processing model is described. Section 3 defines the rulebase on which the query will be applied. Section 4 discusses the specification of TAR-QL. Section 5 concludes the paper.

2. FRAMEWORK OF THE POST-PROCESSING

The KDD process and methodology discussed in Fayyad *et al.* (1996) underline the importance of interaction and iteration. Typically, the iterative process occurs between the discovery and presentation phases, as shown in Figure 1(a). This paper adopts a different approach introduced by Klementtinen (1996) and illustrated in Figure 1(b). This approach puts the emphasis of the iteration process on the presentation phase. Two reasons to justify this approach are mentioned in Klementtinen (1996). First, in exploratory data analysis, it can be hard to specify beforehand what is considered interesting. If the goal is to discover previously unknown information, this can only be achieved by iterative exploration. Second, pattern generation is often time consuming. Although many efficient algorithms exist, the pattern discovery phase is still a barrier for smooth interaction.

The framework has been used to support interactive exploration of large collection of association and episode rules generated from telecommunication alarm databases [8]. The approach uses templates to make a selection from a rulebase, as well as to eliminate a selection from consideration. In Tuzhilin and Adomavicius

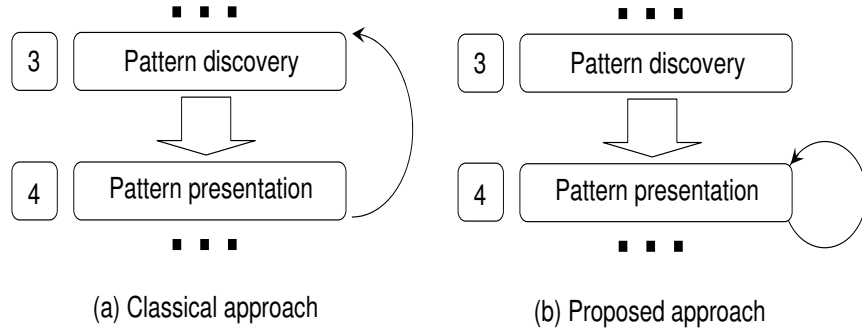


Figure 1: Classical vs proposed approaches

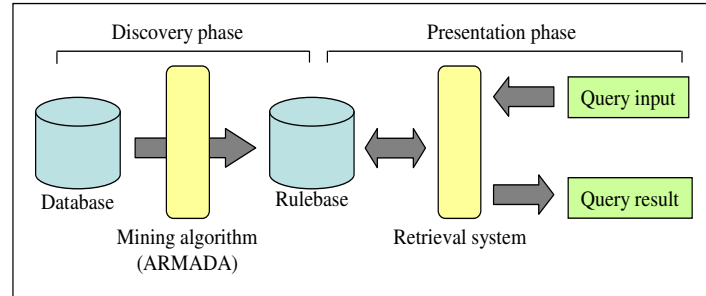


Figure 2: The post-processing framework

(2002), the use of templates is extended to the biological domain to address the problem of analysing very large numbers of discovered rules in the generated from microarray data. When there are multiple sets of discovered rules that have been generated over several time periods, this framework can be employed for analysing the behavior of rules [1, 12, 17].

In the context of richer temporal association rules, the framework is schematically shown in Figure 2. As shown in the figure, the framework puts the emphasis on the rule presentation phase where interesting rules iteratively can be found without repeating rule discovery phase that is more time consuming. The rulebase is generated at once with ARMADA [18], and the iteration is performed mostly in the presentation phase through the retrieval system that provides facilities for selecting interesting rules.

In order to be useful for a wide range of rule retrieval scenarios, the retrieval system should provide flexible accesses to the rulebase. Although the templates are useful for specifying the criteria of association rules to be retrieved, they are not fully suitable for the richer temporal association rules, due to the complex format of the rules. As an alternative, this paper considers the use of query languages

for selecting the rules. It was mentioned above that although several data mining query languages have been proposed, most of them are designed for rule mining and only a few provide enough support of rule retrieval. In contrast to this trend, the query language required in this framework is only to facilitate the retrieval of rules from the rulebase and is never concerned with the mining data itself. Rule-QL [17] is one of the languages satisfying this requirement.

However, since Rule-QL is specifically designed for association rules, it must be extended to deal with richer temporal association rules. Therefore, in this paper, a query language TAR-QL is proposed as a query specification language for selecting richer temporal association rules. TAR-QL is based on the subset of Rule-QL, extended by adding several additional operators and functions to deal with richer temporal association rules. The specification of TAR-QL together with examples are presented in Section 4.

3. RULEBASE OF TEMPORAL ASSOCIATION RULES

The rule set, called *rulebase*, contains richer temporal association rules generated from the database of interval sequences using ARMADA described in [18]. Let R be a rulebase, then R contains the rules of the form $\beta \rightarrow \alpha$, where α and β are normalized temporal patterns such that $\beta \sqsubset \alpha$. Each rule is associated with at least the support and confidence values. In this paper, a temporal pattern β is called LHS (left-hand side) of the rule, while α is RHS (right-hand side) of the rule.

Temporal patterns in the LHS and RHS of the rules are normalized temporal patterns that do not contain temporal extensions because it has been abstracted from the time intervals in a specific interval sequence. In this paper, normalized temporal patterns are defined over a set of state S and a set of interval relations $\{equals (=), contains (c), is-finished-by (fi), starts (s), overlaps (o), meets (m), before (b)\}$. Let $Rel = \{=, c, fi, s, o, m, b\}$ be the set of these seven relations. As an example, Figure 3 presents four normalized temporal patterns and two richer temporal association rules generated from the patterns. It is assumed that the temporal patterns are defined over a set of states $S = \{A, B, C, D\}$.

4. TAR-QL QUERY LANGUAGE SPECIFICATION

This section describes the syntax of TAR-QL for querying a rulebase of richer temporal association rules. Each SELECT statement in TAR-QL must have at least the SELECT clause and the FROM clause. The other four clauses (WHERE, GROUP BY, HAVING and ORDER BY) are optional as shown in the following:

```
SELECT [FUNCT] r
FROM Rulebase r
[WHERE {pruning_condition|rule_state_condition|
      rule_format_condition}
  [{AND|OR} {pruning_condition|rule_state_condition|
      rule_format_condition}] ...
```

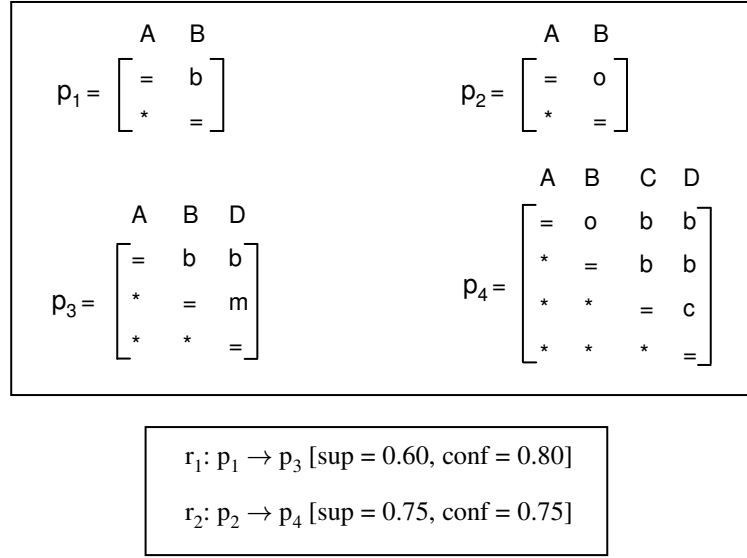


Figure 3: Example of temporal patterns and temporal rules

```

]
[GROUP BY {LHS|RHS|SUPPORT|CONFIDENCE}]
[HAVING condition]
[ORDER BY {LSIZE|RSIZE|SUPPORT|CONFIDENCE} [DESC|ASC]]

```

The FROM clause specifies the rulebase over which the queries are performed, r is the rule variable defined over the rulebase. The optional FUNCT after the SELECT clause specifies aggregation operations over the resulting set of rules.

4.1 Aggregate Functions

A number of aggregate functions exist: COUNT, MAX, MIN, and AVG. These functions can be used in the SELECT clause or in a HAVING clause which will be discussed later. The description of the functions is as follows.

- COUNT: returns the number of rules in the resulting set of rules,
- AVG_CONF, AVG_SUP, AVG_LSIZE, AVG_RSIZE: return the average levels of confidence, support, and the size of rules' LHS/RHS in the resulting set,
- MAX_CONF, MAX_SUP, MAX_LSIZE, MAX_RSIZE: return the maximum values of confidence, support, and the size of rules' LHS/RHS in the resulting set,

- MIN_CONF, MIN_SUP, MIN_LSIZE, MIN_RSIZE: return the minimum values of confidence, support, and the size of rules' LHS/RHS in the answer set.

4.2 Conditional Expressions

Three types of conditions after the WHERE clause are: *pruning conditions*, *rule state conditions*, and *rule format conditions*.

Pruning conditions

The pruning conditions are used to specify queries involving rule attributes, such as support, confidence, the size of rules, and the name of states. They have the following format:

`Rule_func(r) <rel_op> const`

The *Rule_func* is one of the following defined functions:

- CONF(r): returns the confidence value of the rule *r*,
- SUP(r): returns the support value of the rule *r*,
- SIZE_LHS(r): returns the number of states of the rule *r* left-hand side,
- SIZE_RHS(r): returns the size of the rule *r* right-hand side,
- STATE_LHS(r, n): returns the *n*-th state of the rule *r* left-hand side,
- STATE_RHS(r, n): returns the *n*-th state of the rule *r* right-hand side.

The *rel_op* is one of the relational operators $\leq$, $\geq$, $<$, $>$, and $=$. The *const* is a constant.

Query 1 Find rules whose support is greater than 50% and confidence is greater than 70%.

```
SELECT r
FROM Rulebase r
WHERE SUP(r) > 50% AND CONF(r) > 70%
```

Query 2 Find rules whose number of states on the LHS of the rules is fewer than 4 and second state on the RHS of the rules is 'B'.

```
SELECT r
FROM Rulebase r
WHERE SIZE_LHS(r) < 4
      AND STATE_RHS(r, 2) > 'B'
```

Rule state conditions

The rule state conditions allow the user to specify that the rule's LHS/RHS (left-hand side/right-hand side) must contain a certain set of states, be a subset of a given set of states, or be equal to a set of given states. These conditions have the format:

`states_func <set_op> const,`

The *states_func* is one of the following defined functions:

- STATES_LHS(*r*): returns the set of states of the rule *r* left-hand side,
- STATES_RHS(*r*): returns the set of states of the rule *r* right-hand side.

The *set_op* is one of the set operators SUPERSET_OF, SUBSET_OF and EQUAL. The *const* is a set of states.

Query 3 Find rules whose the RHS of the rule contains a set {A, B, D}.

```
SELECT  r
FROM Rulebase r
WHERE STATES_RHS(r) SUPERSET_OF {A, B, D}
```

Rule format conditions

The rule format condition allows the user to specify that the rule's LHS/RHS must contain a temporal pattern, be contained in a temporal pattern, or be equal to a temporal pattern. It has the format:

`pattern_func <pat_op> const`

The *pattern_func* is one of the following two defined functions:

- LHS(*r*): returns the left-hand side of the rule *r*,
- RHS(*r*): returns the right-hand side of the rule *r*.

The *pat_op* is one of the temporal pattern containment operators CONTAINS, CONTAINED_IN and EQUAL. The *const* is a temporal pattern constant.

In order to simplify the writing of temporal patterns in the query, a function TO_PATTERN is introduced to convert a pair containing a list of states and a list of relationships to a normalized temporal pattern. As an example, TO_PATTERN((A,

C, B, D), (o, b, m, b, b, o)) represents a pattern

$$\begin{bmatrix} A & C & B & D \\ = & o & b & b \\ * & = & m & b \\ * & * & = & o \\ * & * & * & = \end{bmatrix}.$$

Query 4 Find rules whose the RHS of the rule contains a pattern $\begin{bmatrix} A & B & D \\ = & o & b \\ * & = & m \\ * & * & = \end{bmatrix}$.

```
SELECT r
FROM Rulebase r
WHERE RHS(r) CONTAINS TO_PATTERN(<A, B, D>, <o, b, m>)
```

4.3 Grouping and Ordering Rules

The GROUP BY clause can be used to group the resulting rules based on various grouping criteria. After the rules have been grouped, aggregate functions can be applied to the individual group. Four keywords that can be used with the GROUP BY clause are LHS, RHS, SUPPORT, and CONFIDENCE, for grouping rules based on the structure of LHS/RHS of the rules, support, and confidence.

Query 5 Find average confidence of rules that have similar structure on their left-hand side.

```
SELECT AVG_CONF(r)
FROM Rulebase r
GROUP BY LHS
```

Sometimes we want to retrieve the values of aggregate functions only for groups that satisfy certain conditions. HAVING clause is used to eliminate all the groups for which the conditional expression of the HAVING clause does not evaluate to TRUE.

Query 6 For each group of rules on which there are more than four rules, find average confidence of rules in the group.

```
SELECT AVG_CONF(r)
FROM Rulebase r
GROUP BY LHS
HAVING COUNT > 4
```

The ORDER clause specifies an order for displaying the resulting set. The ORDER BY LSIZE, ORDER BY RSIZE, ORDER BY SUPPORT, and ORDER BY CONFIDENCE are used to order rules based on the sizes of LHS and RHS of the rules, support, and confidence, respectively. Furthermore, the resulting rules may be sorted in ascending (ASC) or descending (DESC) order.

5. CONCLUSION

In this paper we have described the framework of a retrieval system for selecting interesting rules from a collection of generated temporal association rules. Furthermore, we have also discussed the specification of our proposed TAR-QL query language for the retrieval of richer temporal association rules. TAR-QL is based on the subset of Rule-QL [17], extended by adding several additional operators and functions to deal with richer temporal association rules.

REFERENCES

1. S. Baron and M. Spiliopoulou. Monitoring change in mining results. In Yahiko Kambayashi, Werner Winiwarter, and Masatoshi Arikawa, editors, *Proceedings of the 3rd International Conference on Data Warehousing and Knowledge Discovery (DaWak'01)*, volume 2114 of *LNCS*, pages 51–60, Munich, Germany, 2001. Springer.
2. M. Botta, J.F. Boulicaut, C. Masson, and R. Meo. Query languages supporting descriptive rule mining: A comparative study. In Rosa Meo, Pier Luca Lanzi, and Mika Klemettinen, editors, *Database Support for Data Mining Applications*, volume 2682 of *LNCS*, pages 24–51. Springer, 2004.
3. U. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. The KDD process for extracting useful knowledge from volumes of data. *Communication of the ACM*, 39(11):27–34, 1996.
4. J. Han, J.Y. Chiang, S. Chee, J. Chen, Q. Chen, S. Cheng, Wan Gong, Micheline Kamber, Krzysztof Koperski, Gang Liu, Yijun Lu, Nebojsa Stefanovic, Lara Winstone, Betty B. Xia, Osmar R. Zaiane, Shuhua Zhang, and Hua Zhu. DBMiner: a system for data mining in relational databases and data warehouses. In *Proceedings of the International Conference on Knowledge Discovery in Databases and Data Mining (KDD'96)*, 1996.
5. T. Imielinski and H. Mannila. A database perspective on knowledge discovery. *Communication of the ACM*, 39(11):58–64, 1996.
6. T. Imielinski and A. Virmani. MSQL: A query language for database mining. *Data Mining and Knowledge Discovery*, 3(4):373–408, 1999.
7. M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A.I. Verkamo. Finding interesting rules from large sets of discovered association rules. In N.R. Adam, B.K. Bhargava, and Y. Yesha, editors, *Proceedings of the 3rd International Conference on Information and Knowledge Management*, pages 401–407, Gaithersburg, Maryland, 1994. ACM Press.
8. M. Klemettinen, H. Mannila, and H. Toivonen. Interactive exploration of discovered knowledge: A methodology for interaction, and usability studies. Technical Report C-1996-3, Department of Computer Science, University of Helsinki, Finland, 1996.

9. B. Lent, A.N. Swami, and J. Widom. Clustering association rules. In W. A. Gray and Per-Åke Larson, editors, *Proceedings of the 13th International Conference on Data Engineering (ICDE'97)*, pages 220–231. IEEE Computer Society, 1997.
10. B. Liu, W. Hsu, and Y. Ma. Pruning and summarizing the discovered associations. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'99)*, pages 125–134, San Diego, CA, USA, 1999.
11. B. Liu, M. Hu, and W. Hsu. Multi-level organization and summarization of the discovered rules. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'00)*, New York, NY, USA, 2000. ACM Press.
12. B. Liu, Y. Ma, and R. Lee. Analyzing the interestingness of association rules from the temporal dimension. In *Proceedings of IEEE International Conference on Data Mining (ICDM'01)*, pages 377–384, Silicon Valley, USA, November 2001.
13. R. Meo, G. Psaila, and S. Ceri. A new SQL-like operator for mining association rules. In T. M. Vijayaraman, Alejandro P. Buchmann, C. Mohan, and Nandlal L. Sarda, editors, *Proceedings of the 22th International Conference on Very Large Data Bases (VLDB'96)*, pages 122–133, 1996.
14. A. Netz, S. Chaudhuri, U.M. Fayyad, and J. Bernhardt. Integrating data mining with SQL databases: OLE DB for data mining. In *Proceedings of the 17th International Conference on Data Engineering (ICDE'01)*, pages 379–387, 2001.
15. H. Toivonen, M. Klemettinen, P. Ronkainen, K. Hatonen, and H. Mannila. Pruning and grouping of discovered association rules. In *ECML-95 Workshop on Statistics, Machine Learning, and Knowledge Discovery in Databases*, pages 47–52, Heraklion, Greece, 1995.
16. A. Tuzhilin and G. Adomavicius. Handling very large numbers of association rules in the analysis of microarray data. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'02)*, pages 396–404, New York, NY, USA, 2002. ACM Press.
17. A. Tuzhilin and B. Liu. Querying multiple sets of discovered rules. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'02)*, pages 52–60, New York, NY, USA, 2002. ACM Press.
18. E. Winarko and J.F. Roddick. ARMADA - an algorithm for discovering richer relative temporal association rules from interval-based data. *Data and Knowledge Engineering*, 63:76–90, 2007.

EDI WINARKO: Department of Mathematics, Gadjah Mada University, Yogyakarta, Indonesia.

E-mail: ewinarko@ugm.ac.id

PART V

Statistics

INTRODUCTION OF SPATIAL STRUCTURAL EQUATIONS MODEL

ISAH AISAH, BUDI NURANI AND
SUTAWANIR DARWIS

Abstract. Structural Equation Models (SEM), also called simultaneous equation models, are multivariate regression models. Unlike the more traditional multivariate linear model, however, the response variable in one regression equation in an SEM may appear as a predictor in another equation. These structural equation means to represent causal relationship among the variables in the model.

In this paper we introduce Spatial Structural Equations Model has three variant: spatial lag model (SPM-*l*), spatial error model (SPM-*e*), and the combined spatial lag-error (SPM-*le*).The three spatial models will be presented in the SEM using the simple (latent or observe) regression model.

Key words : Structural Equation Model (SEM)

1. INTRODUCTION

Many application of SEM at field such as: social econometric, Phsycology, physics, etc. SEM is combined between analysis factor (path analysis) and Regression analysis. The class of Structural Equation Models (SEM) makes it possible to simultaneously estimate theoretical and correspondence statements. A SEM is made up of two related sub models :

- A latent variables measurement model which represents the relationship between the latent variables and their observable indicators.
- A structural model, representing the relationship between the latent variables.

SEM that involve spatial data is called spatial SEM.

2. MAIN RESULTS

2.1 The Structural Equation Model

Oud (2000) gave the most general form of SEM as follow

$$\mathcal{G} = \Lambda \mathcal{G} + \omega \quad \text{with} \quad \text{cov}(\omega) = \Omega \quad (1)$$

$$y = \Lambda \mathcal{G} + \varepsilon \quad \text{with} \quad \text{cov}(\varepsilon) = \Theta \quad (2)$$

The model consist of two equations, equation (1) is called latent structural equation and equation (2) is called measurement equation, where :

Vector $\mathcal{G}$ contains the k latent variables,

Δ specifies the structural relationships between the latent variables, and

Ω is the covariance matrix of the latent structural errors in vector ω .

vector y contains the observed variables,

Λ specifies the loading or regression coefficients of the observed variables on the latent variables, and

Θ is the measurement error covariance matrix.

The measurement errors in ε are assumed to be uncorrelated with the latent variables in $\mathcal{G}$ as well as with the structural errors in ω .

It should be noted that (1), which is in structural form with the same vector v at its left-hand and right-hand side, can be written irreduced form with v with left-hand side only as follows $v = (I - \Delta)^{-1} \omega$ and $vv' = [((I - \Delta)^{-1} \omega)((I - \Delta)^{-1} \omega)']$

$$\begin{aligned} \text{such as } E(vv') &= E [((I - \Delta)^{-1} \omega)((I - \Delta)^{-1} \omega)'] \\ &= E [((I - \Delta)^{-1} \omega)(\omega'(I - \Delta')^{-1})] \\ &= (I - \Delta)^{-1} E(\omega\omega')(I - \Delta')^{-1} \\ &= (I - \Delta)^{-1} \Omega (I - \Delta')^{-1} \end{aligned} \quad (3)$$

Following Bollen (1989), we can instead of model (1)-(2) one often encounters a more elaborate version of the SEM model in which endogenous and exogenous variables are put in different vectors:

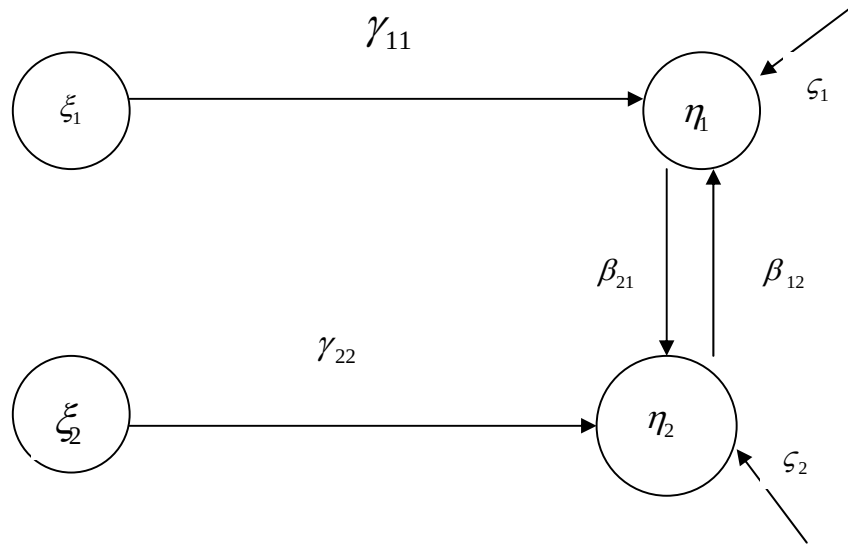


Figure 2.1 Path diagram of SEM

The latent variable model for figure 2.1 is

$$\begin{aligned} \eta_1 &= \beta_{12}\eta_2 + \gamma_{11}\xi_1 + \zeta_1 \\ \eta_2 &= \beta_{21}\eta_1 + \gamma_{22}\xi_2 + \zeta_2 \end{aligned} \quad (4)$$

Equation (4) may be rewritten in matrix notation:

$$\begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} 0 & \beta_{12} \\ \beta_{21} & 0 \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} + \begin{bmatrix} \gamma_{11} & 0 \\ 0 & \gamma_{22} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix} \quad (5)$$

which is more compactly written as

$$\eta = B\eta + \Gamma\xi + \zeta \quad \text{with} \quad \text{cov}(\zeta) = \Psi \quad (6)$$

Equation (6) is a structural equation with assumptions $E(\eta) = 0$, $E(\xi) = 0$,

$E(\zeta) = 0$, ζ uncorrelated with ξ , and

η $m \times 1$ vector, m latent endogenous variables

ξ $n \times 1$ vector, n latent exogenous variables

ζ $m \times 1$ vector, latent errors in equations

B the $m \times m$ -matrix, with $(I - B)$ non singular, as coefficient matrix for latent endogenous variables.

Γ the $m \times n$ - matrix, as coefficient matrix for latent exogenous variables.

Consider figure 2.2 below:

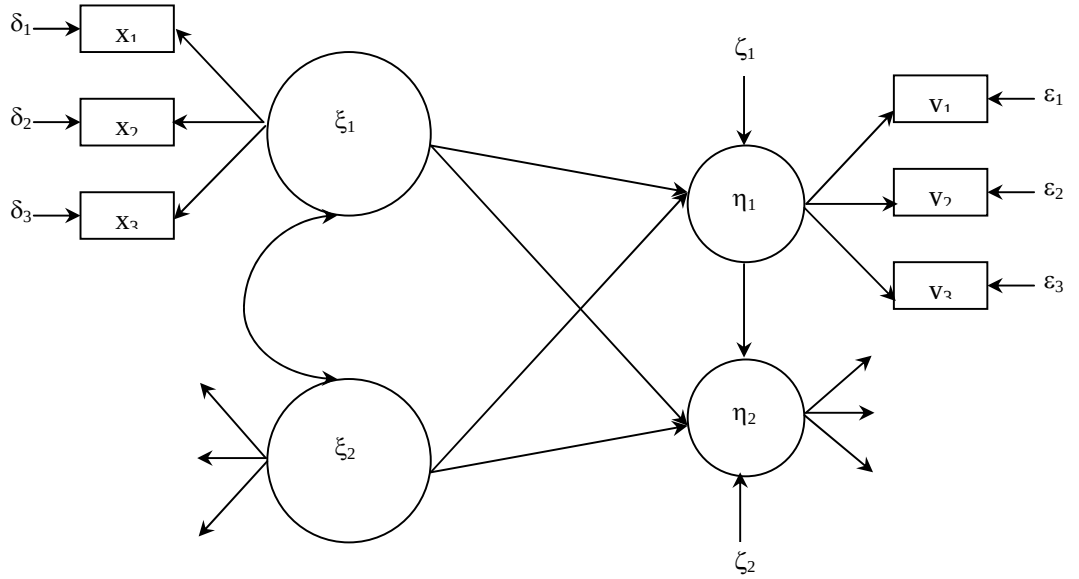


figure 2.2 path diagram of measurement model
figure 2.2 provide a measurement model can be written as:

$$y = \Lambda_y \eta + \varepsilon \quad \text{with} \quad \text{cov}(\varepsilon) = \Theta_\varepsilon \quad (7)$$

$$x = \Lambda_x \xi + \delta \quad \text{with} \quad \text{cov}(\delta) = \Theta_\delta \quad (8)$$

Assumptions: $E(\eta) = 0$, $E(\xi) = 0$, $E(\varepsilon) = 0$, $E(\delta) = 0$, ε uncorrelated with η, ξ and δ , δ uncorrelated with η, ξ and ε .

x as observed indicators of ξ (the $q \times 1$ - matrix)

y as observed indicators of η (the $p \times 1$ - matrix)

Λ_x as coefficients relating x to ξ (the $q \times n$ - matrix)

Λ_y as coefficients relating y to η (the $p \times m$ - matrix)

Θ is the measurement error covariance matrix.

Several parameter estimation methods can be used in most SEM programs, such as the ML method, which maximizes the log likelihood function of the free elements in the parameter matrices Δ, Ω, Λ , and Θ , for given data in Y :

$$\ell(\theta|Y) = -\frac{N}{2} \log|\Sigma| - \frac{N}{2} \text{tr}(\Sigma^{-1} S) - \frac{pN}{2} \log 2\pi \quad (9)$$

θ in (7) contains the free parameters, $Y_{p \times n}$ is the data matrix (N columns of independent replications of the p - variate vector y , typically originating from a sample of randomly drawn subjects), $\Sigma_{p \times p}$ is the model-implied covariance or moment matrix:

$$\Sigma = \Lambda(I - \Delta)^{-1} \Omega (I - \Delta)^{-1} \Lambda' + \Theta \quad (10)$$

Which is a function $\Sigma(\theta)$ of θ , and $S_{p \times p} = \frac{1}{N} Y Y'$ is the sample covariance or moment matrix. The ML-estimator $\hat{\theta} = \arg \max \ell(\theta|Y)$ chooses that value of θ which maximizes $\ell(\theta|Y)$. However, instead of maximizing the log likelihood function in (9) standard soft ware for SEM analysis minimizes the fit function

$$F_{ML} = \log|\Sigma| + \text{tr}(\Sigma^{-1} S) - \log|S| - p \quad (11)$$

with the same result.

2.2 The Spatial SEM

Oud (2004) studied about spatial econometric's with the assumption of independent replications is dropped and spatial dependence between sampling units is accounted for ($N \times N$) weight matrix W describing the spatial arrangement of the N units. Although in practice W is often found to be symmetric, the formulas are valid for non symmetric W .

The spatial model (SPM) has three variants: spatial lag model (SPM- l), spatial error model (SPM- e), and the combined spatial lag-error model (SPM- le). The three spatial models will be presented in the SEM framework using the simple (latent or observed) regression model

$$\eta = \Gamma \xi + \zeta \quad (12)$$

Which is obtained from (6) by specifying $B = 0$.

We rewrite (6) in sampling units formulation by replacing the 1- column vectors η, ξ , and ζ by N - column matrix H, X , and Z :

$$H = \Gamma X + Z \quad (13)$$

The SPM- l consists in adding $H \delta W_\delta$, such as (13) become:

$$\begin{aligned} H &= H \delta W_\delta + \Gamma X + Z \text{ or} \\ H &= (TX + Z)(I - \delta W_\delta)^{-1} \end{aligned} \quad (14)$$

The spatial lag model in SEM as follows:

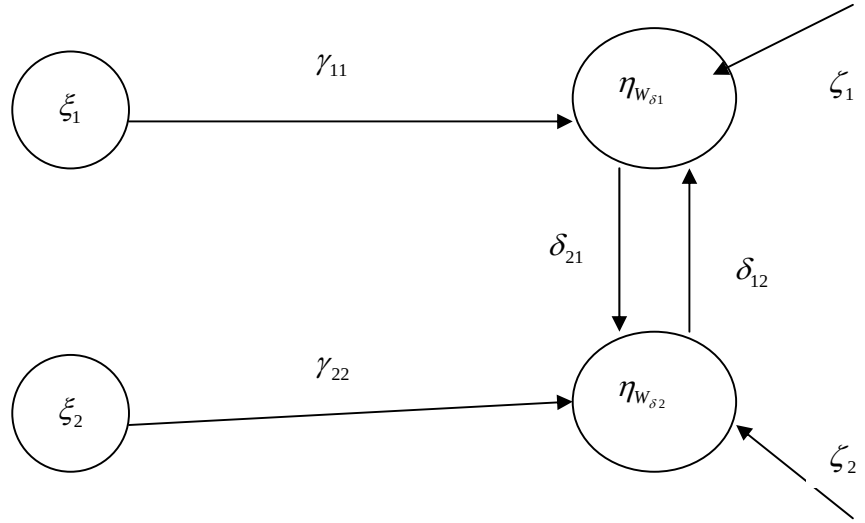


figure 2.3 path diagram of spatial lag model

Figure 2.3 provide a spatial lag model is:

$$\eta = \delta \eta_{w_{\delta}} + \Gamma \xi + \zeta \quad (15)$$

Where

η is taken to represent the dependent variable

δ is the spatial dependence parameter measuring the average influence of contiguous

Observations on η

$\eta_{w_{\delta}}$ is the spatially lagged dependent variable

Γ is regression coefficients of the explanatory variables

ξ is the (N x N) matrix of observations of the explanatory variables

ζ is stochastic disturbances

2.3 Time Series SEM

Stef Van Buuren (1997) explained about Time Series SEM with structural model by Bentler-Weeks in Time Series SEM is defined

$$\eta = \beta_0 \eta + \gamma \xi \quad (16)$$

where β_0 ($m \times m$) and ($m \times n$) are parameter matrices. Let Φ denote the covariance matrix of , which contains elements that are either free, constrained or fixed. Let G be a known matrix that selects observed variables from η and ξ . The covariance structure of the observed variables under model (13) can be represented as

$$\Sigma = G(I - B_0)^{-1} \Gamma \Phi \Gamma' (I - B_0)^{-1} G' \quad (17)$$

Where B_0 and Γ are specified as

$$B_0 = \begin{pmatrix} \beta_0 & 0 \\ 0 & 0 \end{pmatrix}, \quad \Gamma = \begin{pmatrix} \gamma \\ 0 \end{pmatrix}$$

Let the $T \times 1$ vector v_0 contain the time series to be analyzed, and that v_p is the p -th lagged variable of v_0 . The ARMA $(p,0)$ can be expressed as a single equation in the Bentler-Weeks model as

$$v_0 = \gamma \xi = (\Phi_1, \dots, \Phi_p, 1) \begin{pmatrix} v_1 \\ \cdot \\ \cdot \\ \cdot \\ v_p \\ f_0 \end{pmatrix} \quad (18)$$

The variance-covariance of ξ in the model is

$$\Phi = \begin{pmatrix} \tau_0 & \tau_1 & \cdot & \cdot & \cdot & \tau_{p-1} & 0 \\ \tau_1 & \tau_0 & \cdot & \cdot & \cdot & \tau_{p-2} & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \tau_{p-1} & \tau_{p-2} & \cdot & \cdot & \cdot & \tau_0 & 0 \\ 0 & 0 & \cdot & \cdot & \cdot & 0 & \sigma^2 \end{pmatrix} \quad (19)$$

Where τ_i denotes the i -th order auto covariance in the ARMA $(p,0)$ model and where σ^2 is the variance of the noise process f_0 . For $p = 1$, by substituting (18) and (19) into (117) yields $\tau_0 = \sigma^2 / (1 - \Phi_1^2)$, which is identical to the known theoretical variance of the AR(1)- model. Because Box-Jenkins express variances of higher order AR-model in term of infinite MA models, The variance structure of (15) for $p > 1$ cannot be directly related to their results.

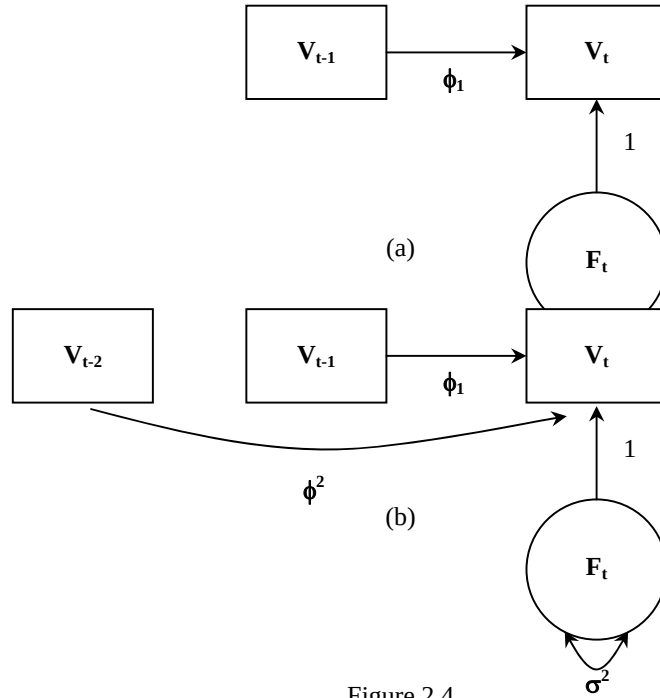


Figure 2.4
Path diagrams of The ARMA(1,0) model (a) and the ARMA(2,0) model (b)

2.4 Generalized Spatial Structural Equation Models (GSSEM)

Melanie (2003) studied about GSSEM provide a method for researchers to perform ecological regressions incorporating many correlated variables in a meaningful way while taking account of spatial structure. Melanie supported that the model could be extended to incorporate both space and time. In this paper we follow the space time model from Pfeiffer (1979) has studied the Space Time Autoregressive model (STAR-model) based on Box-Jenkins model. The STAR model as a combination of a time series Autoregressive (AR) model and a spatial model can be chosen to describe time dependent of an oil production in several wells.

The STAR (1;1) model can be represented by:

$$Z_{(Nx1)}(t) = \phi_{10} Z_{(Nx1)}(t-1) + \phi_{11} W_{(NxN)}^{(1)} Z_{(NxN)}(t) + e(t) \quad (20)$$

Lack of the STAR model from Pfeiffer is only applicable for homogenous location because all of the parameters models are the same for all locations.

Ruchjana, B.N (2002) extended the STAR model from Pfeiffer into be GSTAR (Generalized Space Time Autoregressive) which assume that the autoregressive and the space time parameter are different for all location. The GSTAR order one model on the space and time is called GSTAR (1;1), notify as:

$$Z_{(Nx1)}(t) = \Phi_{10(NxN)} Z_{(Nx1)}(t-1) + \Phi_{11(NxN)} W_{(NxN)}^{(1)} Z_{(NxN)}(t-1) + e_{(Nx1)}(t) \quad (21)$$

2.5 Research Plan

Oud (2004) proposed SEM approach to spatial dependence model and SEM can be applied in a straight forward way to estimate spatial dependence model. Melanie (2003) extended SEM become GSSEM while Stefan van Buuren developed SEM with Time Series SEM. According the suggestion from Melanie (2003), we want to extend Spatial Time SEM to be Space Time SEM, i.e STAR SEM or GSTAR SEM..

3. CONCLUDING REMARKS

- Most generally, SEM combines the ideas of factor analysis with regression. SEM traditionally used in sociology and psychology while SEM development most applied geographically.
- Spatial Structural Equations Model has three variants : spatial lag model (SPM-l), spatial error model (SPM-e), and the combined spatial lag-error (SPM-le). The three spatial models will be presented in the SEM using the simple (latent or observed) regression model.
- ARMA model of Box and Jenkins (1976) can be formulated and fitted as a structural equations model. Writing an ARMA model as a structural model is useful because this opens up the possibility to extend the univariate ARMA model.
- SEM can be extended as Space Time SEM

REFERENCES

1. Bollen, K.A.. *Structural Equations with Latent Variables*. New York : Wiley. 1989
2. Bennet, R.J. *Spatial Time Series : Analysis-Forecasting-Control*. London : Pion
3. Cressie, N. *Statistics for Spatial Data. Revised Edition*. New York : John Willey & Sons, Inc.1993
4. Oud Han and Folmer Henk. *Spatial Structural Equations Models*
5. van Buuren, S . *Fitting ARMA Time Series by Structural Equation Models* : The Psychometric Society.1997.
6. Ruchjana, B.N. An Generalized Space Time Autoregressive Model , Dissertation Unpublished, ITB, 2002

Isah Aisah: Ph D student at The Department of Mathematics and Natural Science, Padjadjaran University.

E-mail : isah_aish@yahoo.co.id

Budi Nurani: A lecturer at the Department of Mathematics, Padjadjaran University

E-mail : bnurani@yahoo.com

Sutawanir Darwis: A lecturer at the Department of Mathematics, Institut Teknologi Bandung

E-mail : sdarwis@dns.math.itb.ac.id

ESTIMATING THE ABILITY OF *Scylla serata*, *Pinctada margaritifera* AND *Perna viridis* AS ADSORBENT FOR LEAD FROM AQUEOUS SOLUTION: USING LINEAR REGRESSION MODELS

KAMSIA BUDIN, AMRAN AHMED, SITI RAHAYU MOHD.
HASHIM, NORAINI ABDULLAH, ALICE NGO CHIN-CHIN

Abstract. The ability of crab, oyster and mussel shells as adsorbents to remove lead (Pb) from aqueous solution were estimated using linear regression models by taking into account factors, such as time of exposure, types of adsorbent, particle size and amount of adsorbent. Results show that, the adsorbents reached its maximum absorption in 6 hours in most cases. The percentage of removal efficiency was influenced by time of exposure, amount and particle size of adsorbent used in the experiment. Negative correlations were observed between the particle size and percentage of adsorption for crab and oyster. However, it seemed that the Pb adsorptions using mussel were significantly influenced only by time.

Key words and Phrases : crab, oyster, mussel, adsorbent, percentage of removal efficiency, mass, size particle, lead, linear regression models

1. INTRODUCTION

The presence of heavy metals in water bodies continues to be of great concern worldwide. One of the metals released into the environment from a number of sources is lead (Pb) which often used in solder for electronics, as a major constituent of lead-acid batteries, as a coloring element in paint and ceramic glazes and also in building construction. It is a poisonous metal and even at a very low concentration, it could cause anemia and brain disorders especially in children [1,2]. Various technologies have been developed to remove lead and other heavy metals contamination from water bodies and wastewater, such as ion-exchange, evaporation, electrolysis, filtration, reverse-osmosis and activated carbon sorption [3,4]. However, these technologies sometimes are inadequate, expensive and create metal bearing sludge.

Therefore, adsorbents from crab, mussel and oyster are industrially attractive alternatives because they are capable of lowering the transition metal concentration, are widely available (can be obtained cheaply from by products or wastes in the food processing industries) and also being environmental friendly [5]. These adsorbents are also highly tolerant for organics and regeneration [6,7]. All these properties encourage the development of crab, mussel and oyster as the new low-cost operators, simple and yet still highly effective metal removal. In this present work, the ability of crab shell (*Scylla serata*), oyster shell (*Pinctada margaritifera*), and green mussel shell (*Perna viridis*) to remove lead from aqueous solution was examined to determine their removal efficiency.

However, in order to transfer the technologies from the laboratory studies to full-scale applications, it is necessary to develop mathematical models for the ion adsorption by the shells. The models should be helpful in interpreting the experimental data and predicting the treatment results when the conditions are changed. In this study a simple regression model was developed to estimate the ability of the three types of shell as adsorbents for Pb based on three factors which were amount of adsorbents (g) and particle size (mm) and time of exposure (minute).

2. MATERIALS AND METHODS

2.1 Shell collection

Crab shell (*Scylla serata*), oyster shell (*Pinctada margaritifera*), and green mussel shell (*Perna viridis*) were obtained as wastes from around the shores along the coast of Kota Kinabalu in Sabah, Malaysia. The shells were separated from their meat by steaming or boiling, washed and then dried without any special treatment. They were then pulverized to a geometric mean particle size of 0.25mm, 0.5mm and 1.0mm.

2.2 Laboratory analysis and data collection

The sorption experiment were conducted under pH 5.0 and room temperature with 150 rpm shaking condition to study the effects of particle sizes (0.25mm, 0.5mm and 1.0mm), mass of adsorbents (0.25g, 0.5g and 1.0g), and the effects of time of exposure at 5, 30, 60, 180, 360, 540 and 720 minutes. The Pb concentrations in the supernatant were analyzed by flame absorption spectrophotometer (AAS Polarized Zeeman Z- 5000) and the percentage removal efficiency were determined as follows:- $\{ (Co - Ct)/Co \} \times 100$, where Co is the initial concentration of Pb (mg/l), Ct is the concentration (mg/l) of Pb in the aqueous in t minutes [5].

2.2 Statistical Methodology

The data were run using Statistical Package for Social Sciences (SPSS) version 12 for the statistical analysis. Skewness and kurtosis coefficients were used to check on the normality of the data. The strength of association between the variables measured by the Pearson correlation coefficient (1), r is given by:-

$$r = \frac{\sum_{i=1}^n X_i Y_i - \frac{\sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{n}}{\sqrt{\left(\sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i \right)^2}{n} \right) \left(\sum_{i=1}^n Y_i^2 - \frac{\left(\sum_{i=1}^n Y_i \right)^2}{n} \right)}} \quad (1)$$

where, X, Y = the variables being compared

n = sample size

Linear regression model (2) was used to regress the amount of the criterion variable with respect to the manipulation of the amount or values of the regressor variables;

$$Y_i = \alpha + \sum_{i=1}^k \beta_i . X_i \quad (2)$$

Where, Y_i = the percentage of Plumbum

α = the intercept

β_i = the parameter coefficients for the regressors

k = the number of regressors

X_1, X_2, X_3 = the regressor variables

The proportion of the total variability in the percentage of Plumbum, Y , attributable to the dependence of Y on all the regressors was measured by the adjusted coefficient of multiple determination (3), R_a^2 .

$$R_a^2 = 1 - \frac{MSE}{MS(Total)} \quad (3)$$

Where, MSE = mean sum of squares of error

$MS(Total)$ = mean sum of squares of total

[8]

3. MAIN RESULTS

The results thus obtained were presented and discussed in this following section.

3.1 Time of exposure

The experimental results for the adsorbents showed rapid initial adsorption rate followed by a slower rate. Initially, the adsorption sites were opened and the metal ions interacted easily with the site within the first 180 minutes and hence higher rate of adsorption were observed and started to deliberate after 360 minutes.

This meant that the adsorption of Pb on the crab, oyster and mussel shells were actively occurred and reached its maximum level after exposed for 6 hours and reached its equilibrium state after 540 minutes for almost all cases (figure 1, 2 & 3). It also seemed that the crab shells showed to be a better adsorbent as compared to the oyster and the mussel. In every case (figures 1, 2, & 3), the crab shell could adsorb more than 80% after 180 minutes of exposed although different sizes of particles and mass were used in the experiments.

The mussel shells showed a similar pattern with the crab but with % Pb removal efficiencies slightly lower than those of the crab shells. The definite difference between them was not found and the variation of % Pb removal efficiencies according to time was not so large as compared to those by the oyster shells.

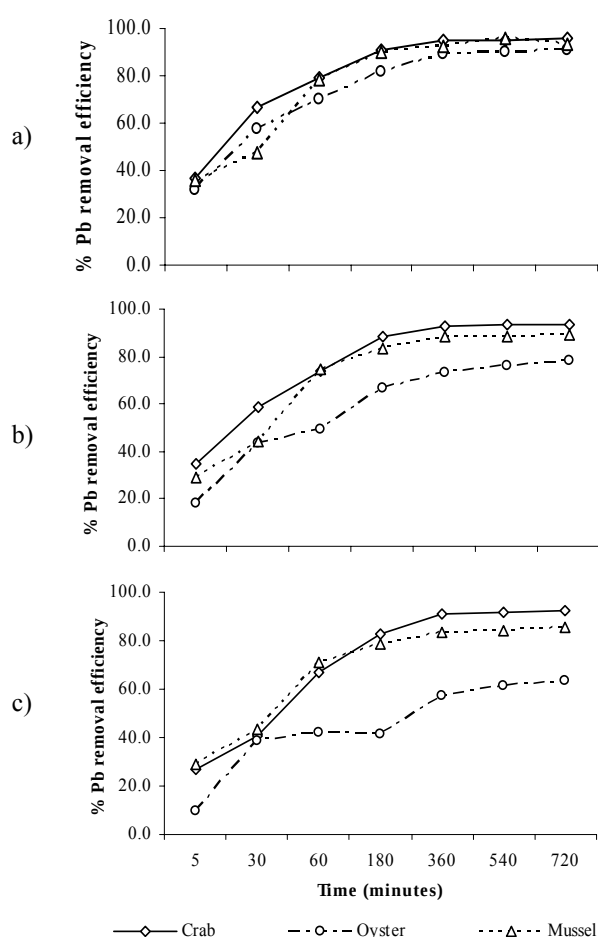


Figure 1. Time courses of % Pb removal efficiencies using 0.25g mass of crab, oyster and mussel with 0.25mm (a), 0.5mm (b) and 1.0mm particle size.

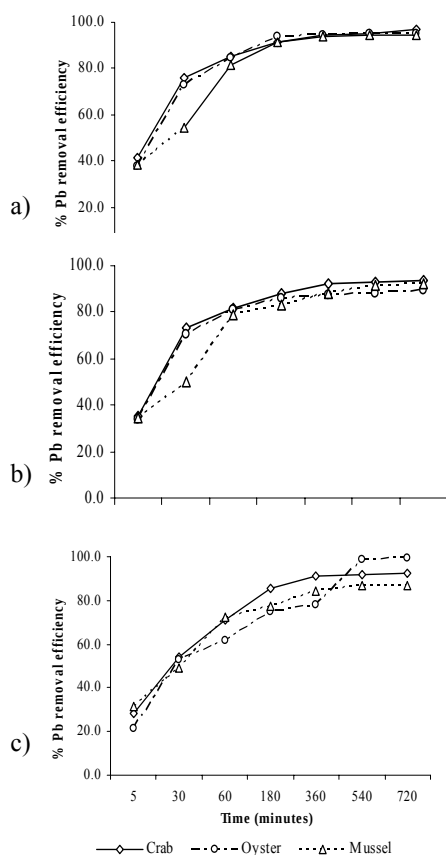


Figure 2. Time courses of % Pb removal efficiencies using 0.50g mass of crab, oyster and mussel with 0.25mm (a), 0.5mm (b) and 1.0mm particle sizes.

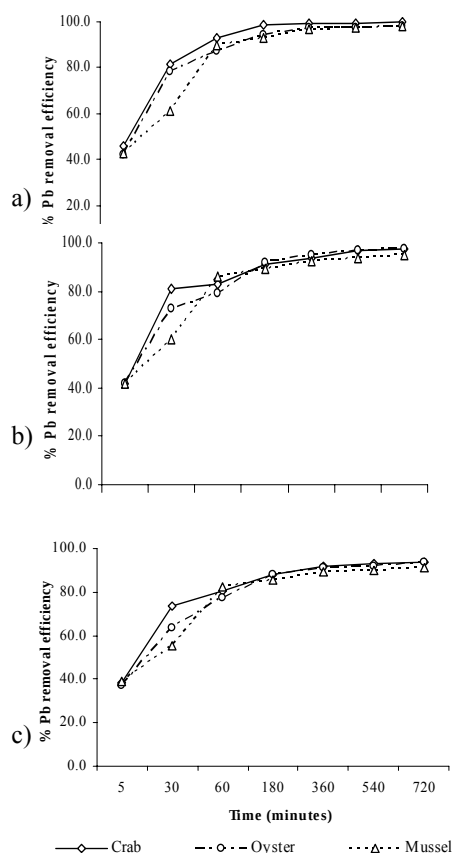


Figure 3. Time courses of % Pb removal efficiencies using 1.0g mass of crab, oyster and mussel with 0.25mm (a), 0.5mm (b) and 1.0mm particle sizes.

3.2 Effects of adsorbent amount

Effects of adsorbent amount on percentage removal of Pb ions were shown in figures 1, 2 & 3. Percentages of lead removal for crab and mussel were slightly increased as increased amount of shells were used during the experiment. Unlike the crab and mussel shells, the % Pb removal efficiency for oyster shells had seemed to be more affected by the adsorbent amount. These were illustrated by figures 1b & 1c, 2b & 2c and 3b & 3c. Figures 1b and 1c showed that the oyster had the lowest % Pb removal efficiency after exposed for 5 minutes (<20%) and had not even reached its maximum absorption at 720 minutes. On the contrary, if increased amounts of 0.5 g and 1.0g were used (figures 2b, 2c, 3b and 3c), the % Pb removal efficiencies could be increased to more than 30% after being exposed for 5 minutes and hence, reached its maximum adsorption within 540 minutes.

This was due to the thickness difference between the three types of shells. Oyster shells were thicker as compared to ones of crab and mussel. Therefore, more oyster shells were needed to increase the number of adsorption sites or surface area and hence, resulted in a higher percentage of metal removal at a

higher dose [5,6,9].

3.3 Effect of particle size

Surface area of the adsorbent would be an important parameter for adsorption. Exposure of adsorbent sites for solid-metal ion interaction was high if the surface area of adsorbent was high too. The smaller the particle size, the higher would be the surface area per unit weight of adsorbent and hence, a higher percentage of metal removal was expected. The results showed that for all types of shell particles of size 0.25mm, the percentage of Pb removal were > 90% within 180 minutes and then reached its equilibrium in 540minutes.

3.4 Statistical analyses

The Skewness-kurtosis coefficient was within the range ± 1 , we assumed that the normality of the data were fulfilled. The statistical analysis using Pearson test verified that the amount of mass and particle size of the adsorbents and the time of exposure had significantly influenced the ability of each type of shell (Table 1). However, it is varies among shell. The adsorption using crab shells were more influence by the particle size of adsorbent used and time of exposure. All three factors were had significantly influence the Pb removal efficiencies using oyster shell but only time were found significant for the mussel shell in the experiment.

A positive correlation was observed between the percentage of adsorption with mass of adsorbent used and the length of exposure in the study. This meant that the ability of the adsorbents to remove Pb from the aqueous solution would increase directly to the amount of adsorbent used and the time period. Basically, it was due to the presence of more adsorbent were available for the Pb adsorption on its surface. The longer time of exposure would give more time for the process to occur. However, a negative correlation occurred between the particle size and the percentage of removal efficiencies of Pb in the aqueous. The smaller the particle sizes used, the better ability of the shells as adsorbents had become because more surface area were available for the Pb adsorption on the surface of the shells.

Table 1: Correlation Coefficients of %Pb removal efficiencies and factors

Factors	% Pb removal efficiencies		
	Crab shell	Oyster shell	Mussel shell
Mass	0.163	0.373**	0.155
Particle size	-0.176*	-0.285**	-0.151
Time	0.685**	0.626**	0.713**

** Correlation is significant at the 0.01 level (2 tailed)

* Correlation is significant at the 0.05 level (2 tailed)

The removal mechanism of lead was proposed to be a micro precipitation of PbCO_3 and $\text{Pb}_3(\text{CO}_3)_2(\text{OH})_2$ which would adsorb on the surface of the shells. Carbonates of (CO_3^{2-}) were derived from calcium carbonates, which formed the major component of the crab, oyster and mussel shell particles and formed complexes of (PbCO_3 and $\text{Pb}_3(\text{CO}_3)_2(\text{OH})_2$) with the lead ions in the flasks [10].

3.5 Linear Regression models

Linear regression models were developed from the experimental data, as the assumptions were compliance with the regression diagnostic as shown in table 2.

Table 2 : Regression model % Pb removal efficiencies for each specimen against mass (g), time (minute), and particle size (mm).

Specimen	Factors	Unstandardized Coefficients		t	Sig.	R ²
		B	Std. Error			
Crab shell	Constant	71.387	3.136	22.761	<0.001	0.500
	Time	0.055	0.005	10.746	<0.001	
	Particle Size	-11.709	4.248	-2.756	0.007	
Oyster shell	Constant	52.93	3.916	13.518	<0.001	0.612
	Mass	27.513	4.161	6.612	<0.001	
	Time	0.056	0.005	11.111	<0.001	
	Particle Size	-21.008	4.161	-5.048	<0.001	
Mussel shell	Constant	58.928	1.962	30.031	<0.001	0.509
	Time	0.060	0.005	11.328	<0.001	

The three linear regression models on the % of Pb removal efficiencies were obtained from the three types of shells (4), (5) and (6):-

$$\% \text{Pb}_{\text{Cr}} = 71.387 + 0.055t - 11.709p \quad (4)$$

$$\% \text{Pb}_{\text{Oy}} = 52.93 + 27.513m + 0.056t - 21.008p \quad (5)$$

$$\% \text{Pb}_{\text{Mu}} = 58.928 + 0.060t \quad (6)$$

where,

% Pb _{Cr}	-	Pb removal efficiency using crab shells as adsorbent
% Pb _{Oy}	-	Pb removal efficiency using oyster shells as adsorbent
% Pb _{Mu}	-	Pb removal efficiency using mussel shells as adsorbent
m	-	mass
t	-	time
p	-	particle size

4. CONCLUDING REMARKS

The removals of lead proceeds were due to the precipitation and adsorption of lead carbonates. In the experiment, most of reaction was completed in 9 hours (540 minutes). These reactions were influenced by the amount and particle sizes of adsorbents used and the time of exposure of lead to the adsorbents during the experiments. The linear regression models which had been developed from the three adsorbents clearly showed that the reaction could be increased further and was directly proportional to its amount but indirectly proportional to the size particle of adsorbents used, i.e., smaller sizes were preferable. Further works would be needed as the model for mussel and crabs showed that only times and particle sizes were the major factors that influenced the percentages of Pb adsorption. We suggested that various of particle sizes and different amount of adsorbent for a longer time of exposure could be taken into account for future experiments.

REFERENCES

1. M.H Yu, *Environmental Toxicology: Biological & Health Effects of Pollutants*. 2nd Edition. CRC Press LLC, New York, 2005.
2. M. Lipmann, *Environmental Toxicants: Human Exposures and Their Health Effects*. Van Nostrand Reinhold, New York, 1992.
3. K.P. Vijayaraghavan, Palanivelu, K., & Velan, M. Biosorption of copper(II) and cobalt(II) from aqueous solutions by crab shell particles, *Biosource Technology*, **97** (12), 411-414, 2006.
4. M. Machida, Yamazaki, Y., Aikawa, M. & Tatsumoto, H., Roles of mineral in carbonaceous adsorbents for removal of Pb(II) ions from aqueous solution. *Separation and Purification Technology*, **46** (1-2), 88-94, 2005.
5. B. Kamsia, Mohd. Harun A., Siti Rahayu M. H., & Noraini A., Removal of lead from aqueous solutions using crab (*Portunus pelagicus*) and clam (*Anadara granosa*) shell waste, Proceedings of the 2nd Southeast Asian Natural Resources and Environmental Management Conference. Kota Kinabalu, Sabah Malaysia, 159 – 163, 2006.
6. D.S. Kim, Pb²⁺ removal from aqueous solution using crab shell treated by acid and alkali. *Biosource Technology*, **94** (3), 345-348, 2004.
7. D. S. Kim, The removal by crab shell of mixed heavy metal ions in aqueous solution. *Bioresource Technology*, **87** (3), 355-357, 2003.
8. J. H Zar, *Biostatistical Analysis*, 4th edition, Prentice-Hall, Inc., 1999.
9. H. K. Ann, Kim, D.S. & Park, B.Y., Crab shell for the removal of heavy metals from aqueous solution. *Water Research*, **35**(15), 3551 – 3556, 2001.
10. M.Y. Lee, Lee, S.H., Shin, H.J., Kajiuchi, T. & Yang, J.W., Characteristics of lead removal by crab shell particles. *Process Biochemistry*, **33** (7), 749-753, 1998.

Kamsia Budin: Environmental Science Programme, School of Science & Technology, UMS, Locked bag 2073, 88999, Kota Kinabalu, Sabah.
Email : bkamsia@ums.edu.my, rahayu@ums.edu.m

Amran Ahmed: Mathematics and Economic Programme, School of Science & Technology, UMS, Locked bag 2073, 88999, Kota Kinabalu, Sabah.

Siti Rahayu Mohd. Hashim: Mathematics and Economic Programme, School of Science & Technology, UMS, Locked bag 2073, 88999, Kota Kinabalu, Sabah.

Noraini Abdullah: Mathematics and Economic Programme, School of Science & Technology, UMS, Locked bag 2073, 88999, Kota Kinabalu, Sabah.

Alice Ngo Chin-Chin: Environmental Science Programme, School of Science & Technology, UMS, Locked bag 2073, 88999, Kota Kinabalu, Sabah.

DUAL TIME SCALES IN EVENT HISTORY ANALYSIS

DANARDONO

Abstract. The connection between *the Cox time-dependent covariate model*, *the piecewise constant hazard Poisson regression* and *the Cox time-dependent strata model* in modeling dual time scales is discussed. Modeling age-period infant mortality model in Indonesia were used as an illustrative application of the methods.

Key words and Phrases: the Cox proportional hazard model, Lexis diagram, piecewise constant hazards, time-dependent strata, time-dependent covariate

1. INTRODUCTION

Event history analysis (EHA) is a statistical analysis of time-to-event data or durational data. Time is indispensable in EHA, since it can be used as a proxy measure for other influences of the events. Additionally, time is the most readily available measurement and easy to utilize for comparison and generalization. For example, in epidemiology, age is the most often used time scale since it reflects cumulative damage that causes mortality, whereas, in clinical studies, time since diagnosis may be more important.

The choice of time scale is driven by the research question of the study. However, in the absence of knowledge about the importance of time scales, we may have to consider all relevant time scales. In an epidemiological surveillance study, it is common to perform an exploratory study to identify new emerging risk factors. One way of exploring the factors is by investigating several relevant time scales.

This paper is motivated by the study of child survival in the Community Health and Nutrition Research Laboratory (CHN-RL) Purworejo, Indonesia [11]. Some extensive studies have been done and reported elsewhere [10, 5], however,

there is a methodological problem left. Survival or infant mortality studies usually consider age as the basic time scale. However, calendar time or period is also important if we assume seasonality influences mortality. When we want to include both age and period in the model, we may choose one of them and use the other as a time dependent covariate [9]. To treat multiple time scales symmetrically, piecewise constant hazards and discrete time models are the usual approaches [9, 1, 6]. Basically the approaches are Poisson regression and Cox's proportional hazards model.

The present paper discusses some solutions and the connection between them. Three solutions are considered here, i.e. *the Cox time-dependent covariate model*, *the piecewise constant hazard Poisson regression* and *the Cox time-dependent strata model*.

2. THE COX MODEL

The Cox proportional hazards model specifies that the hazard of a subject in a certain time t given covariate $\mathbf{Z}$ follows

$$\lambda(t | \mathbf{Z}) = \lambda_0(t) \exp(\boldsymbol{\beta}\mathbf{Z}), \quad (1)$$

where $\lambda_0(t)$ is an unspecified non-negative function called the *baseline hazard* common to all subjects and $\boldsymbol{\beta}$ is a set of unknown regression coefficients.

Cox has proposed a semiparametric approach for the proportional hazards model (1) above [3, 4]. Let $\mathcal{D}$ be the set of indices j of ordered event-times $t_1, t_2, \dots, t_j, \dots$ and R_k be risk set at time t_k the subjects under observation and event-free immediately prior to t_k . The partial likelihood is given by

$$L(\boldsymbol{\beta}) = \prod_{k \in \mathcal{D}} \frac{\exp(\boldsymbol{\beta}\mathbf{Z}_k)}{\sum_{j \in R_k} \exp(\boldsymbol{\beta}\mathbf{Z}_j)}, \quad (2)$$

in which the baseline hazard $\lambda_0(t)$ is canceled out. The $\boldsymbol{\beta}$ can be estimated by finding the value of $\boldsymbol{\beta}$ which maximizes $L(\boldsymbol{\beta})$.

For certain situations, it is often necessary to stratify the subjects into disjoint groups when the proportionality assumptions do not hold for one or several covariates. In this so-called stratified Cox model, the subjects in a certain stratum have a distinct baseline hazard function but common values for the regression coefficients. The partial likelihood for the stratified Cox model is given by

$$L(\boldsymbol{\beta}) = \prod_{s=1}^S L_s(\boldsymbol{\beta}), \quad (3)$$

where S is the number of strata and $L_s(\boldsymbol{\beta})$ is the partial likelihood as in Equations (2) but calculated only for the subjects in stratum s .

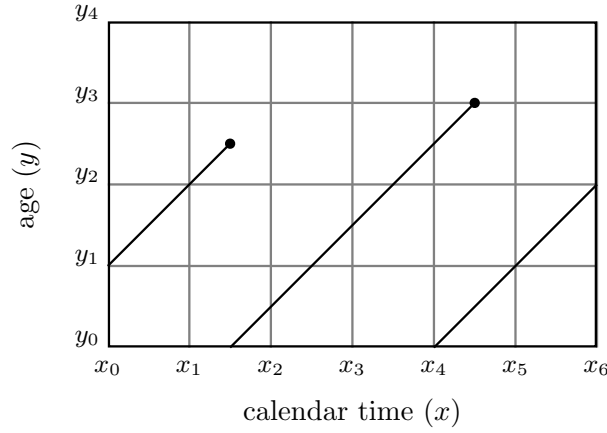


Figure 1: Hypothetical event history data on a Lexis diagram. The lines represent the observed follow-up time by age and calendar time (period); the dots represents the event of interest (deaths, diseased)

It is necessary to perform a test for departures from proportionality in the Cox regression analysis since the method assumes that the hazards are proportional across the time t . The tests are basically tests for nonzero slopes in a generalized linear regression of the rescaled residuals as a function of time [7].

3. MODELING DUAL TIME SCALES

Figure 1 represents typical dual time scales event history data on a *Lexis* diagram [9]. Lexis diagram is a popular diagram used in demography to describe age-period-cohort data.

The general aim is to model the hazards as a function of age y , calendar time x and covariate $\mathbf{Z}$ which may also depend on y and x . Let $\mu(x, y)$ be the hazard function at period x and age y . Generalizing from the single time scale, the Cox proportional hazard model for dual time scales is

$$\mu(x, y | \mathbf{Z}) = \mu_0(x, y) \exp(\beta \mathbf{Z}), \quad (4)$$

where $\mu_0(x, y)$ is the baseline hazard function at period x and age y common to all individuals. Three approaches are considered here to model (4), i.e., the piecewise constant hazards, time-dependent strata and time-dependent covariate methods.

3.1 Piecewise Constant Hazards

In the piecewise constant hazards model we assume that the hazard function $\mu(x, y)$ is piecewise constant across the Lexis plane. Technically, the Lexis plane

is divided into sufficiently small rectangles such that constant hazard function μ can be reasonably assumed in each rectangle. Let u_i be the total exposure time in a rectangle and d_i be the number of events (0 or 1) for individual i , then the contribution of individual i to the likelihood is

$$L_i(\mu) = (\mu)^{d_i} \exp(-\mu u_i), \quad i = 1, \dots, n, \quad (5)$$

in this specific rectangle. To assess other effects on the hazard we may specify $\mu \exp(\mathbf{Z}\boldsymbol{\beta})$ instead of only μ , where $\mathbf{Z}$ is a vector of covariates and $\boldsymbol{\beta}$ is a vector of unknown regression coefficients. Although any functional form of $\mathbf{Z}$ and $\boldsymbol{\beta}$ is possible, the log-linear form $\exp(\boldsymbol{\beta}\mathbf{Z})$ is convenient.

Let the Lexis plane, as in Figure (1), be divided into smaller rectangles

$$\Omega_{(r,s)} = \{(x, y) : x \in [x_{r-1}, x_r) \text{ and } y \in [y_{s-1}, y_s)\},$$

$r = 1, 2, \dots, R$, $s = 1, 2, \dots, S$; $d_{i(r,s)}$ and $u_{i(r,s)}$ be the number of observed events and time spent (exposure time) in each $\Omega_{(r,s)}$ for individual i , $i = 1, \dots, n$. The likelihood for the piecewise constant hazards model (5) for all individuals and over the lexis grid Ω is

$$L(\boldsymbol{\mu}, \boldsymbol{\beta}) = \prod_{r=1}^R \prod_{s=1}^S \prod_{i=1}^n \left[(\mu_{rs} e^{\boldsymbol{\beta}\mathbf{Z}_i})^{d_{i(r,s)}} \exp(-\mu_{rs} e^{\boldsymbol{\beta}\mathbf{Z}_i} u_{i(r,s)}) \right], \quad (6)$$

where μ_{rs} is the baseline hazard in $\Omega(r, s)$.

It is possible to assess the effects of time (age and calendar time) on the hazard by assuming a multiplicative decomposition $\mu_{rs} = \lambda_s \gamma_r$.

3.2 Time-dependent Approaches

In the single time scale situation, the partial likelihood used in the Cox proportional hazards model to estimate the regression coefficients can be interpreted as a profile likelihood obtained from a piecewise constant hazards likelihood maximized to certain nuisance parameters and allowing the width of the time intervals approaching zero [8, 2]. This procedure does not work in the dual time scale situation due to the lack of smoothness of the maximum likelihood baseline rate estimates [9, 1].

An alternative approach is to let the partition of one time scale interval be fixed as the partition in the other direction gets finer and finer. In the limit, we get two different solutions, depending on which partition is kept fixed.

We consider the likelihood for the piecewise constant hazards model of Equation (6). Now, given $\boldsymbol{\beta}$, the μ_{rs} may be separately estimated as follows. Looking at specific values of r and s and suppressing the dependence of them, and taking logs gives

$$\ell_{rs} = \sum_{i=1}^n [d_i \log \mu + d_i \boldsymbol{\beta}\mathbf{Z}_i - \mu \exp(\boldsymbol{\beta}\mathbf{Z}_i) u_i]. \quad (7)$$

By equating the derivative of (7) wrt μ to zero, we get $\hat{\mu}_{rs}(\beta)$:

$$\hat{\mu}_{rs}(\beta) = \frac{\sum_{i=1}^n d_{i(r,s)}}{\sum_{i=1}^n u_{i(r,s)} \exp(\beta \mathbf{Z}_i)}, \quad r = 1, \dots, R; \quad s = 1, \dots, S. \quad (8)$$

By replacing μ_{rs} in (6) by (8), taking logarithms and simplifying, we get the profile log likelihood

$$\ell_p(\beta) \propto \sum_{r=1}^R \sum_{s=1}^S \sum_{i=1}^n d_{i(r,s)} \log \left(\frac{e^{\beta \mathbf{Z}_i}}{\sum_{j=1}^n e^{\beta \mathbf{Z}_j} u_{j(r,s)}} \right). \quad (9)$$

3.3 The Time-dependent Strata Approach

We proceed with the approach with a fixed period (calendar time) x scale, i.e., we keep R in (9) fixed.

Now let $\omega = y_s - y_{s-1}$ be the constant width of the time intervals on the y scale. When $S \rightarrow \infty$ ($\omega \rightarrow 0$), $d_i(r, s)$ and $u_i(r, s)$ will become

$$d_i(r, s) = \begin{cases} 1 & \text{if an event occurs for individual } i \text{ in } \Omega(r, s), \\ 0 & \text{otherwise,} \end{cases}$$

$$u_i(r, s) \approx \begin{cases} \omega & \text{if individual } i \text{ is observed in } \Omega(r, m), \\ 0 & \text{otherwise.} \end{cases}$$

Let

$$Y_i(r, s) = \begin{cases} 1 & \text{if individual } i \text{ is observed in } \Omega(r, m), \\ 0 & \text{otherwise.} \end{cases}$$

The profile likelihood (9) then becomes

$$\begin{aligned} \ell_p(\beta) &\approx \sum_{r=1}^R \sum_{s=1}^S \sum_{i=1}^n d_{i(r,s)} \log \left(\frac{e^{\beta \mathbf{Z}_i}}{\sum_{j=1}^n e^{\beta \mathbf{Z}_j} \omega Y_{j(r,s)}} \right) \\ &= \sum_r \sum_s \sum_i d_{i(r,s)} \log \left(\frac{e^{\beta \mathbf{Z}_i}}{\sum_j e^{\beta \mathbf{Z}_j} Y_{j(r,s)}} \right) \\ &\quad - \sum_r \sum_s \sum_i d_{i(r,s)} \log(\omega) \\ &\propto \sum_r \sum_s \sum_i d_{i(r,s)} \log \left(\frac{e^{\beta \mathbf{Z}_i}}{\sum_j e^{\beta \mathbf{Z}_j} Y_{j(r,s)}} \right) \end{aligned} \quad (10)$$

removing the terms independent of β . Since the $d_i(r, s)$ has values 1 only at the event times, the contributions to the likelihood are only at the event times. Therefore the denominator of the log part is actually a sum over the *risk set* given r .

The profile likelihood can be written as

$$\ell_p(\beta) = \sum_{r=1}^R \sum_{i \in D_r} \log \left(\frac{\exp(\beta \mathbf{Z}_i)}{\sum_{j \in R_r(y_i)} e^{\beta \mathbf{Z}_j}} \right), \quad (11)$$

where D_r is the event set and $R_r(y_i)$ is the risk sets at y_i , given r . In Figure 1, event set is all lines with dots, and the risk set is all lines that intersect the horizontal line crossing each dot (the event times y). The profile likelihood (11) corresponds to the partial likelihood of Cox's proportional hazards model with basic time scale age y and time dependent strata on the time scale x .

3.4 Time-dependent Covariate Approach

Assuming a multiplicative model for the baseline hazard function,

$$\mu_{rs} = \lambda_s \gamma_r, \quad r = 1, \dots, R; \quad s = 1, \dots, S, \quad (12)$$

we get a slightly different profiling procedure, leading to the time-dependent covariate approach. The log likelihood becomes

$$\ell(\gamma, \lambda, \beta) = \sum_{s=1}^S \sum_{r=1}^R \sum_{i=1}^n [d_{i(r,s)} \log(\lambda_s \gamma_r e^{\beta \mathbf{Z}_i}) - \lambda_s \gamma_r e^{\beta \mathbf{Z}_i} u_{i(r,s)}]. \quad (13)$$

Given β and $\gamma_r, r = 1, \dots, R$, maximizing (13) with respect to $\lambda_1, \dots, \lambda_S$ is straightforward. The solution is

$$\hat{\lambda}_s = \frac{\sum_r \sum_i d_{i(r,s)}}{\sum_r \gamma_r \sum_i e^{\beta \mathbf{Z}_i}}, \quad s = 1, \dots, S. \quad (14)$$

Substituting (14) into (13) and simplifying by removing the terms independent of β and γ gives the profile likelihood

$$\ell_p(\gamma, \beta) \propto \sum_{i=1}^n \sum_{r=1}^R \sum_{s=1}^S d_{i(r,s)} \log \left(\frac{\gamma_r \exp \beta \mathbf{Z}_i}{\sum_{t=1}^R \gamma_t \sum_{j=1}^n e^{\beta \mathbf{Z}_j} u_{j(t,s)}} \right). \quad (15)$$

We proceed with this derivation in a similar manner to that of the case with time dependent strata. When $S \rightarrow \infty$ or $\omega \rightarrow 0$, $d_{i(r,s)}$ and $u_{i(r,s)}$ becomes the event indicator and at risk indicator at time s . The summation over all individuals i becomes the summation over the event times $i \in D$. At the denominator of the log part, summation will be determined only at the event times s , since all other terms will vanish by the definition of the event indicator $d_{i(r,s)}$. Similarly, in the denominator the summation will be over the risk set $R(y_i)$. The summation over γ_t will also be completely determined by $j \in R(y_i)$. The profile likelihood becomes

$$\ell_p(\gamma, \beta) = \sum_{i \in D} \sum_{r=1}^R \log \left(\frac{\gamma_r \exp(\beta \mathbf{Z}_i)}{\sum_{j \in R(y_i)} \sum_m \gamma_m e^{\beta \mathbf{Z}_j}} \right). \quad (16)$$

The log profile likelihood (16) is exactly the log of Cox's partial likelihood with a time dependent categorical covariate, where the categories are defined by the time intervals $(x_{r-1}, x_r]$, $r = 1, \dots, R$. Instead of categorical covariate, we may also specify the values of x_r or any function of x_r at the event times.

A similar connection can be derived by letting the age be fixed and period interval lengths approach zero. The result will be the Cox proportional hazards model with (age) entering as time dependent strata or as a time dependent covariate in the model with basic time scale calendar time.

Other pairs of time scales are of course possible. For instance, dual time scales age and time since diagnosis arise frequently in clinical studies, age and time since weaning is another example from childhood life studies.

4 Infant Mortality Age-Period Analysis

We now turn into the practical application which motivates this study. The CHN-RL surveillance database from 1994 to 1998 consists of 3953 subjects were used to model infant mortality. The covariates under interest are the education of the mother and the gender of the infant (Extensive studies of other variables have been reported elsewhere [10, 5].)

Analyses with piecewise constant hazards, Cox's proportional hazards with age time scale and Cox's proportional hazards with period time scale were performed. Two-month grids were applied for both age and period. For the piecewise constant hazards model, the standard Poisson model for the number of events in each grid with log link function was used. The total exposure times in each grid was entered to the model as an *offset*. For the Cox model with age time scale, period time was included as time-dependent strata or time-dependent covariate. Similarly, in the Cox model with period time scale, age was included as time-dependent strata or time-dependent covariate. The time-dependent covariates in both models using age and period as the basic time scale were treated non-parametrically using a penalized smoothing spline.

There is no scientific background suggesting that the two time scales, age and calendar time, are two alternative time scales. However, we can examine this by checking the proportionality assumption of the model using age and period as the basic time scale in separate analyses. No model violates the proportionality assumption with relatively large p -values for the proportionality test of 0.332 and 0.763 for age and period time scale, respectively.

Tables 1 and 2 show the result of likelihood ratio tests for the variables in each model and estimated coefficients, respectively. In this particular data set, in fact, they gave similar results. However, the safe approach is to consider the results from a Cox model with the time-dependent strata or piecewise constant hazards method. The general conclusion is that maternal education is quite important and gives protective effect in the case of infant mortality.

Table 1: Likelihood ratio test (LRT) for variables in the infant mortality models using piecewise constant hazards (pc-hazards), Cox proportional hazards with age time scale (Cox-age), Cox proportional hazards with period time scale (Cox-period)

Variables	pc-hazards	Cox-age		Cox-period	
		td-strata	td-covar	td-strata	td-covar
Age	< .001	—	—	—	< .001
Period	.395	—	.979	—	—
Gender	.080	.100	.075	.104	.132
Maternal educ.	< .001	< .001	< .001	< .001	< .001

Table 2: Estimated coefficients and their standard errors for gender and maternal education in the infant mortality models

Variables	pc-hazards	Cox-age		Cox-period	
		td-strata	td-covar	td-strata	td-covar
Gender					
boy	—	—	—	—	—
girl	-0.31 (.18)	-0.29 (.18)	-0.32 (.18)	-0.29 (.18)	-0.29 (.19)
Maternal educ.					
none	—	—	—	—	—
6 years	-0.76 (.26)	-0.76 (.27)	-0.74 (.26)	-0.72 (.27)	-0.53 (.28)
9 years	-1.17 (.34)	-1.18 (.35)	-1.16 (.34)	-1.11 (.35)	-1.10 (.36)
12 years	-2.74 (.56)	-2.72 (.56)	-2.70 (.56)	-2.67 (.56)	-2.43 (.57)

5 Conclusions

As we have discussed in Section 3, the piecewise constant hazards, the Cox-time dependent covariate, and the Cox-time dependent strata methods are closely related. Which method should we use in practice? The first consideration when we face a multiple time scales problem in event history analysis is to look for the scientific background of the time scales. However, when there is no information, but we believe multiple time scales should be included in the model, we recommend the Cox model with time-dependent strata or the piecewise constant hazard approach. The Cox model with time-dependent covariate is advisable if we have information on the functional form of the time dependent covariate, which is usually difficult to determine in practice.

REFERENCES

1. C. BERZUINI AND D. CLAYTON, Bayesian analysis of survival on multiple time scales, *Statistics in Medicine*, **13**, 823–838, 1994.
2. D. CLAYTON, The analysis of event history data: A review of progress and outstanding problems, *Statistics in Medicine*, **7**, 819–841, 1988.

3. D.R. COX, Regression models and life-tables (with discussion), *Journal of the Royal Statistical Society, Series B, Methodological*, **34**, 187–220, 1972.
4. D.R. COX, Partial likelihood, *Biometrika*, **62**, 269–276, 1975.
5. DANARDONO, *Event history analysis of childhood mortality and morbidity in Purworejo, Indonesia*, Statistical studies 30, Department of Statistics, Umeå University, 2003
6. B. EFRON, The two-way proportional hazards model, *Journal of the Royal Statistical Society, Series B, Methodological*, **64**, 899–909, 2002.
7. P.M. GRAMBSCH AND T.M. THERNEAU, Proportional hazards tests and diagnostics based on weighted residuals, *Biometrika*, **81**, 515–526, 1994.
8. S. JOHANSEN, An extension of Cox's regression model, *International Statistical Review*, **51**, 165–174, 1983.
9. N. KEIDING, Statistical inference in the lexis diagram, *Phil. Trans. R. Soc. London A*, **332**, 487–509, 1990.
10. A. WAHAB, A. WINKVIST, H. STENLUND, AND S.A. WILOPO., Infant mortality among Indonesian boys and girls: Effect of sibling status, *Annals of Tropical Paediatrics*, **21**, 66–71, 2001
11. S.A. WILOPO AND CHN-RL TEAM, *Key Issues on Research Design, Data Collection and Management. Community Health and Nutrition Research Laboratory, Faculty of Medicine, Gadjah Mada University*, Reprint Series No. 2. Community Health and Nutrition Research Laboratory, Yogyakarta, 1997

DANARDONO: Department of Mathematics, Gadjah Mada University, Yogyakarta, Indonesia.
 E-mail: danardono@ugm.ac.id.

DEVELOPMENT OF BOOTSTRAP SIMULATION FOR TRACER MODELING

SUTAWANIR DARWIS, AGUS YODI GUNAWAN,
M. ALI ASHAT, SRI WAHYUNINGSIH, NURTITI SANUSI

Abstract. A simulation of nonlinear regression based on original data for tracer flow model is described. A nonlinear regression of tracer profile is developed, and a simulation based on original data is proposed. The approach can be seen as a variant of Monte Carlo simulation for linear regression. Tracer experiments involve injecting a solute tracer into reservoir system and monitoring its recovery at various production wells and aims to investigate the connectivity between injection and production wells. Partial differential equations are used as a model representation of physical process of tracer flow. Earlier modeling tracer test employed nonlinear regression method to obtain point estimates to match the return profile to physical model, and little research has been done to produce a confidence interval for the tracer return profile. However, the sampling distribution of parameter estimate remains an open research problem. Due to the complexity of estimator dependency on sample values, an asymptotic approach is not an easy task. Bootstrap simulation provides an alternative initial study of sampling distribution of parameter estimate of tracer nonlinear regression, modeling, and it is shown that simulation provides a reasonable result

Keywords : *tracer , simulation of nonlinear regression*

1. INTRODUCTION

This paper focuses on the simulation study, called bootstrap, of the parameters of tracer flow model. It is a method of random resampling of the random components in stochastic models. The tracer profile is modeled as a nonlinear regression. The parameters are estimated using nonlinear least squares method. The distribution of parameter estimate is studied using random sampling from historical data. The methodology will be seen to be a variant of bootstrapping linear regression tailored to nonlinear regression model.

Tracer experiment involves injecting a (chemical) tracer into a reservoir system and monitoring its recovery through time at various well locations. The applications of tracer tests in geothermal reservoir is to predict the cooling of production wells due to long term reinjection of cooler fluid through studying connection between injection and production wells (Axelsson, Bjornsson and Montalvo, 2005). This paper aims to develop a confidence interval for flow velocity and channel dispersion parameters through bootstrap methodology. Traditional Monte Carlo simulation is based on two assumptions: (1). A model which predicts the tracer performance and (2). The distribution of the parameters of the model are known (Jochen, Spivey, 1996). However, the bootstrap approach offers the advantages over Monte Carlo

simulation: there are no prior assumptions of parameter distribution; the sampling distribution of parameter distribution is based on a single sample data set. In recent years, bootstrap gains its popularity in engineering literature due its flexibility to normality assumption (Jochen and Spivey, 1996, Cheng, Wang, McVay, Lee, 2005).

The diffusion equation for tracer transport is given by

$$D \frac{\partial^2 C}{\partial x^2} = u \frac{\partial C}{\partial x} + \frac{\partial C}{\partial t}$$

where $D(\text{m}^2/\text{s})$ is the dispersion coefficient, u (m/s) is the flow velocity $C(x, t)(\text{kg}/\text{m}^3)$ is the tracer concentration in the channel and $x(\text{m})$ is the distance from injection well. The solution is given by a normal probability with mean ut and variance $2Dt$

$$C(x, t) = \frac{1}{\sqrt{4\pi Dt}} e^{-\frac{(x-ut)^2}{4Dt}}$$

Monte Carlo simulation will focus on transformed tracer concentration at the production well

$$(1 + ar) \frac{u}{\sqrt{4\pi Dt}} e^{-\frac{(x-ut)^2}{4Dt}}$$

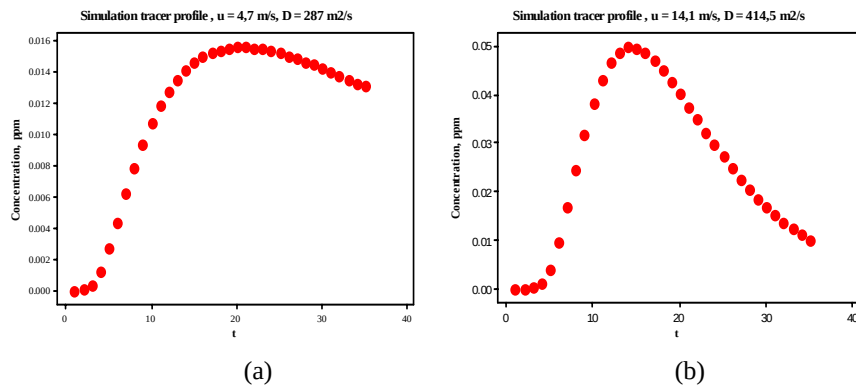
where $r \sim U(0,1)$, a is randomness adjustment, $0 < a < 1$.

The main idea of the simulation, implemented using Monte Carlo method, is that sampling distribution can often be studied by resampling a given set of data and calculating the parameter of interest for each synthetic sample. The resulting bootstrap distribution provides an approximate sampling distribution for the parameter estimate. Based on the unconditional coefficient of variation of the Monte Carlo approximation suggests that 25 resample will give reasonable results (Booth, Sarkar, 1998). The details of the simulation is as follow: the original data $\mathbf{X} = (X_1, \dots, X_n)$ is a sample from an unknown distribution F , and let $\mathbf{X}^* = (X_1^*, \dots, X_n^*)$ be a resample of size n chosen with replacement from empirical distribution $X \sim F : 1/n$. Simulation is used to estimate the distribution of a random variable $\theta = \theta^*(\mathbf{X}^*)$. Let $G(z) = P(\theta^* \leq z)$. The Monte Carlo simulation begins by generating a large number B of samples $\mathbf{X}^{*b}$, $b = 1, \dots, B$ of size n with replacement from the original data $\mathbf{X}$, and compute the statistics θ^{*b} . The bootstrap distribution estimate is $G^*(z) = 1/B \sum_{b=1}^B I(\theta^{*b} \leq z)$, where I is the indicator function. Efron

and Tibshirani (1986) discussed the details of bootstrapping regression, and the nonlinear least squares regression is developed in regression textbooks (Kennedy and Gentle, 1980)

2. MAIN RESULTS

The quantitative interpretation of tracer test analysis and injection simulation based on diffusion process has been implemented on the ICEBOX software (Axelsson, Björnsson, Montalve, 2005). The ICEBOX includes several modules: TRINV for estimation of tracer model, and TRCOOL for prediction cooling of reservoir due to long term injection. However, the previous existing approach focuses on point estimate, and little attention has been done to produce a sampling distribution. This section addresses the issue of sampling distribution of tracer flow parameters. A simulation of nonlinear regression of tracer profile is proposed to study the parameter sampling distribution. For simplicity of presentation, a bootstrap replication $B = 10$ was chosen. Figure 1 shows the simulated tracer profile for parameter flow velocity $u = 4.7, 14.1, 36.2$ m/s, and dispersion $D = 287, 414.5, 3161$ m²/s and noised tracer profile $(1 + .2U(0,1)) \times c(x,t)$. The parameters flow velocity and channel dispersion D clearly influence the tracer profile. The peak represents the number of path that connected the injection and production wells. Single peak suggests that at least one path. Peak arrival time is related to flow velocity. Small peak arrival time indicates a fast low. Figure 1.a has small flow velocity (4.7 m/s), medium channel dispersion (297 m²/s) and shows a medium peak arrival time. Figure 1.b has fast velocity (14.1 m/s), medium dispersion (414.5 m²/s), and shows a fast peak arrival time. Figure 1.c has fast velocity (36.2 m/s), large dispersion (3161 m²/s), and shows $t = 12$ as peak arrival time. A noised tracer profile $(1+.2U(0,1))$ of Fig 1.c will be used for estimation and simulation of nonlinear regression of tracer flow model as proposed in section 1 (Figure 1.d). Figure 2 shows the bootstrap simulation for tracer flow model with flow velocity 36.2 m/s and channel dispersion 3161 m²/s. Using nonlinear least squares procedure, the estimated parameters are 33.805 and 2217.753 with no significant error $sse = .0002$. Ten simulation curves of tracer flow are plotted on the graph with the original data. The curves captures the main characteristic of the original data, except there is some peak arrival time delay from 12 (for the original data) to 13 (for simulated curves). The anomaly at the peak of arrival time of the original data may cause disturbance to the simulated curves. A larger bootstrap replication B is needed to obtain a better result. This small replication simulation experiment provides a acceptable result.



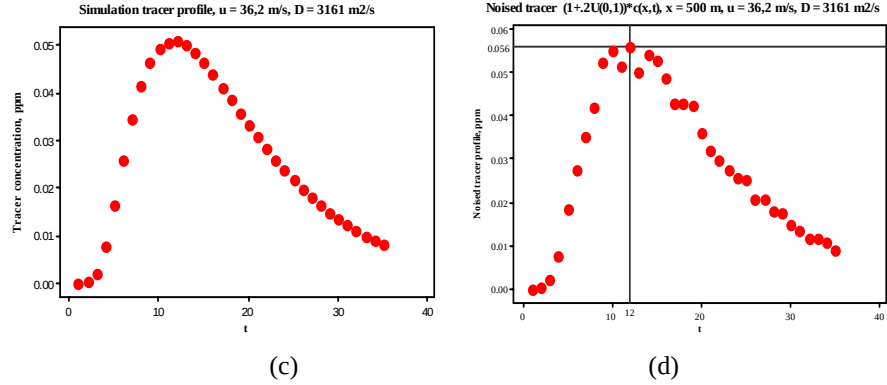


Figure 1. Simulated tracer profile $u = 4.7, 14.1, 36.2$ m/s, $D = 287, 414.5, 3161$ m^2/s and noised tracer profile $(1 + .2U(0,1)) \times c(x,t)$

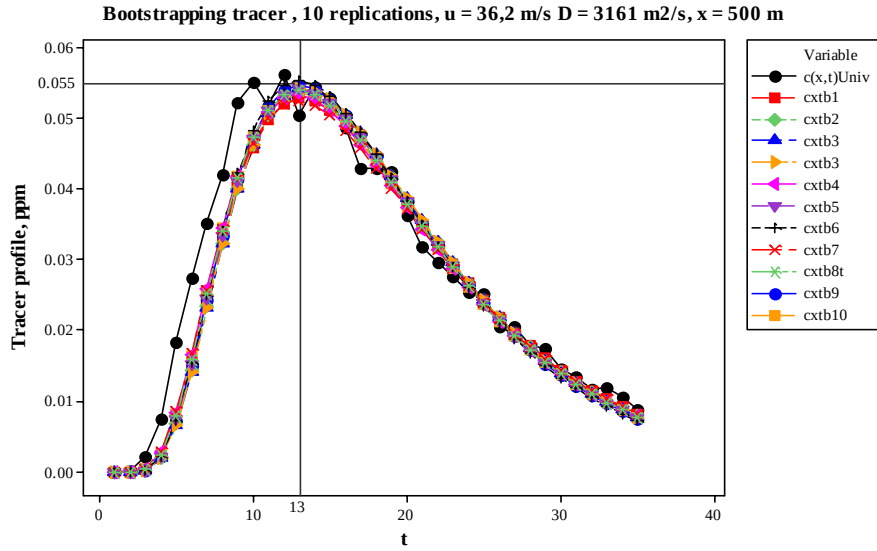


Figure 2. Bootstrap estimate tracer profile based ten replications for noised tracer profile $(1 + .2U(0,1)) \times c(x,t)$, $u = 36.2$ m/s, $D = 3161$ m^2/s . The simulation curves are similar to the original data. The delay of peak arrival time is due to the unusual decline near the original peak arrival time.

3. CONCLUDING REMARKS

This paper addresses the issue of curves estimation confidence interval for tracer flow model to achieve better prediction accuracy using Monte Carlo simulation due to the intractability of nonlinear least squares estimation. The confidence interval is more informative compared to just a single (point estimate) value and gains its popularity in engineering literature. The proposed method consists of a combination of nonlinear regression and simulation from original data. The regression modeling was supported by physical model and analytical solution of process. The methodology can be seen as a variant of Monte Carlo simulation for linear regression. A nonlinear regression was developed using flow velocity and channel dispersion as the model parameters. The simulation curves fits well to the original data and suggests that only small number of resample are required

SUTAWANIR DARWIS, AGUS YODI GUNAWAN, M. ALI ASHAT,
SRI WAHYUNINGSIH, NURTITI SANUSI

to give reasonable result. The future work will possibly be to develop an analytical study to determine the optimum number of resample for a given accuracy such that the simulation error has little impact on the conclusions.

Acknowledgement. The authors would like to thank Institut Teknologi Bandung for financial support of this research. The comments of the reviewers improves the quality of this papers and are highly appreciated.

REFERENCES

1. Axelsson G, Bjornsson G and Montalvo F, 2005, Quantitative Interpretation of Tracer Test Data, *Proceedings World Geothermal Congress*, Turkey, 24-29 April 2005, p 1 – 12
2. Booth J G and Sarkar S, 1998, Monte Carlo Approximation of Bootstrap Variances, *The American Statistician*, vol 52, no 4, 354 - 357
3. Cheng Y, Wang Y, McVay D A and Lee W J, 2005, Practical Application of a Probabilistic Approach to Estimate Reserves Using Production Decline, *SPE95974* presented at the 2005 SPE Annual Technical Conference and Exhibition, Dallas USA, 9 – 12 October, Society of Petroleum Engineer , p 1 – 13
4. Efron B and Tibshirani R J, 1993, *An Introduction to the Bootstrap*, Chapman & Hall, Ch 9
5. Jochen V A and Spivey J P, 1996, Probabilistic Reserve Estimation Using Decline Curve Analysis with the bootstrap, *SPE36633* presented at the 1996 SPE Annual Technical Conference and Exhibition, Denver USA, 6 – 9 October, Society of Petroleum Engineers, p 589-596
6. Kennedy W J and Gentle J E, 1980, *Statistical Computing*, Marcel Dekker, Ch 10

Sutawanir Darwis: Statistics Research Division, Faculty of Mathematics & Natural Sciences, Institut Teknologi Bandung, Indonesia
E-mail: sdarwis@math.itb.ac.id

Agus Yodi Gunawan: Financial & Industrial Mathematics Research Division, Faculty of Mathematics & Natural Sciences, Institut Teknologi Bandung, Indonesia.
E-mail: ayodi@math.itb.ac.id

M. Ali Ashat: Geothermal Research Division, Faculty of Earth Sciences & Mineral Technology, Institut Teknologi Bandung, Indonesia.
E-mail: Ali_Ashat@tm.itb.ac.id

Sri Wahyuningsih: Statistics Dept, Faculty of Mathematics & Natural Sciences, Universitas Mulawarman, Samarinda, East Kalimantan, Indonesia.
E-mail: sri_ws305@students.itb.ac.id

Nurtiti Sanusi: Mathematics Dept, Faculty of Mathematics & Natural Sciences, Universitas Haluoleo, Kendari, South East Sulawesi, Indonesia.
E-mail: ntiti@students.itb.ac.id

NUMERICAL SIMULATION OF STATISTICAL TESTS TO DETERMINE THE BEST SUBINTERVAL IN GLOBAL OPTIMIZATION

YOSZA DASRIL, ISMAIL BIN MOHD, AND SISKI CANDRA
NINGSIH

Abstract This paper will discuss the randomness, normality and unimodality tests. Firstly, the randomness and normality tests must be done to the data which are obtained by splitting the interval $[a, b]$ into several subintervals $[x_i, x_{i+1}]$ ($i = 0, \dots, n-1$) with $x_0 = a$ and $x_n = b$ for verifying that the optimization problem constitutes a Wiener process. From these several subintervals, the best chosen subinterval is obtained by using the optimization probabilistic algorithm which incorporates the ideas of randomness, normality and unimodality tests. Furthermore, numerical results are given to illustrate these tests.

Keywords : *Global Optimization, Probabilistic, Randomness test, Normality test and Unimodality Tests.*

1. INTRODUCTION

The task of global optimization is to find a solution in the solution set for which the objective function obtained its smallest value. The global optimization aims is determining not just a local minimum but the smallest local minimum with respect to the solution set or feasible region.

Global optimization is concerned with finding the global minimum of a function $f: S \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^1$ where f might be have multiple local minima but only the smallest one is of interest here, namely

$$f^* = \inf\{f(x) \mid x \in S\}. \quad (1.1)$$

The value f^* is called the global minimum of f . A point x^* for which $f(x^*) = f^*$, is called a global minimizer of f ([5]). The global maximum of function f can be determined by transforming the maximization to minimization problem using

$$\max f(x) = -\min(-f(x)). \quad (1.2)$$

Extending the class of function to include multimodal functions make the global optimization problems unsolvable in the general. In order to be solvable some smoothness condition on the function in addition to continuity must be known.

In this paper, we describe randomness and normality tests to the data which are obtained by splitting the interval $[a, b]$ into several subintervals $[x_i, x_{i+1}]$ ($i = 0, \dots, n-1$) with $x_0 = a$ and $x_n = b$ for verifying that the optimization problem constitutes a Wiener process. From these several subintervals, the best chosen subinterval is obtained by using the optimization probabilistic algorithm. And then, the unimodality test will be performed in the best subinterval to constitute the subinterval is unimodal function.

We consider the problem of finding a global minimum of the optimization problem

$$z = f(x) \quad (1.3)$$

subject to

$$a \leq x \leq b \quad (1.4)$$

where $f(x) : \mathbb{R}^1 \rightarrow \mathbb{R}^1$ and $x \in \mathbb{R}^1$.

The set of the feasible points which will be considered in this paper, is a closed set defined by

$$F = \{x \mid a \leq x \leq b, x \in \square I\}. \quad (1.5)$$

In Section 2, we will describe the stochastic process for objective function. In Section 3, we will describe the statistical tests. The data, which are supplied to the parameter, will be given in Section 4. Section 5 will consider the stopping criterion for the algorithm to determine the best subinterval. How to the determine the best subinterval will be explained in Section 7. Section 8 will describe the unimodality tests. In addition, numerical results are given to illustrate these tests in Section 9. We will complete this paper with the conclusion as given in Section 10.

2. STOCHASTIC PROCESS

A stochastic model for the objective function $f(x)$ implies that $f(x)$ is considered as a realization of a stochastic process $f(x, \omega)$ where ω belongs to some probability space. As the stochastic model, we shall consider the Wiener Process ([1]) $f(x)$ for $x \in X = [x_0, \bar{x}]$ with

$$f(x_0) = \mu \text{ and } f(x) - f(y) \sim N(0, \sigma^2 |x - y|), \text{ for } x, y \in X. \quad (2.1)$$

Given n function evaluations $f_i = f(x_i)$ ($i = 1, \dots, n$) the distribution of $f(x)$, conditioned by $z_n = (x_1, f_1, x_2, f_2, \dots, x_n, f_n)$, is normal with some expected

value $\mu(x)$ and variance $\sigma^2(x)$, where for any subinterval $\Delta_i = [x_i, x_{i+1}]$, ($i = 0, 1, \dots, n-1$) we are given

$$\mu(x) = E(f(x) | zn) = \int_i \frac{x_{i+1} - x}{x_{i+1} - x_i} + \int_{i+1} \frac{x - x_i}{x_{i+1} - x_i}, \quad (2.2)$$

$$\sigma^2(x) = \text{var}(f(x) | zn) = \sigma^2 \frac{(x - x_i)(x_{i+1} - x)}{x_{i+1} - x_i}, \quad (2.3)$$

$x \in \Delta_i$, $i = 0, \dots, n-1$, and $f_0 = \mu$.

If $i = n$, and $\Delta_n = [x_n, x]$ then for $x \in \Delta_n$,

$$\mu(x) = E(f(x) | zn) = f(x_n) \quad (2.4)$$

and

$$\sigma^2(x) = \text{var}(f(x) | zn) = \sigma^2(x - x_n). \quad (2.5)$$

The following theorem ([2]), which mentions about the distribution of the global minimum of $f(x)$, is designed for algorithm described in the next section.

Theorem 2.1

If $f(x)$, $x \in [a, b]$, is a Wiener process such that $f(a) = f_a$, and by considering the distribution

$$f(z) = P\{\min_{a \leq t \leq b} f(t) \leq z | f(b) = h\},$$

then

$$f(z) = \begin{cases} 1, & (z \geq \min(f_a, f_b)) \\ \exp\left(-2 \frac{(f_a - z)(f_b - z)}{\sigma^2(b - a)}\right), & (z < \min(f_a, f_b)). \end{cases} =$$

The statistical tests must be performed to check whether the objective function is Wiener process or not before we proceed with the probabilistic algorithm.

3. STATISTICAL TESTS

Given an objective function $f: S \subseteq \mathbb{R}^1 \rightarrow \mathbb{R}^1$ and the points x_i ($i = 1, \dots, n$) split the search interval X in equal parts. Let

$$z_i = f(x_i) - f(x_{i-1}) \quad (i = 1, \dots, n). \quad (3.1)$$

The goodness of fit of the Wiener process for f can be checked in two steps, testing whether z_i ($i = 1, \dots, n$) is a random sample drawn from a normal distribution.

3.1 Randomness Test

The randomness of sample z_i ($i = 1, \dots, n$) can be checked by the sample correlation coefficient

$$R = \frac{\frac{1}{n} \sum_{i=1}^n i z_i - \bar{z} \bar{I}}{S_z S_1}, \quad (3.2)$$

where

$$\bar{z} = \frac{1}{n} \sum_{i=1}^n z_i, \quad (3.3)$$

$$S_z^2 = \frac{1}{n} \sum_{i=1}^n (z_i - \bar{z})^2, \quad (3.4)$$

$$\bar{I} = \frac{n+1}{2}, \quad (3.5)$$

and

$$S_1 = \frac{n^2 - 1}{12}. \quad (3.6)$$

If z_i ($i = 1, \dots, n$) are independent on i , then the random variable

$$T_n = R \sqrt{\frac{(n-2)}{(1-R^2)}} \quad (3.7)$$

follows a t -distribution with $\nu = n - 2$ degrees of freedom. The variables are random variables if $|T_n| < T_{\alpha, n}$.

3.2 Normality Test

The normality of the sample z_i can be examined as follows. Firstly, we compute

$$\sigma_i = \max \left\{ F_0(z_{(i)}; \bar{z}; S_z) - \frac{i-1}{n}, \frac{i}{n}, \frac{i}{n} - F_0(z_{(i)}; \bar{z}; S_z) \right\} \quad (3.8)$$

where z_i ($i = 1, \dots, n$) are values z_i ($i = 1, \dots, n$) rearranged in increasing order, and $F_o(x; \bar{x}; s)$ is the standard normal distribution function given by

$$F_o(x; \bar{x}; s) = \int_{-\infty}^{\frac{x-\bar{x}}{s}} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du \quad (3.9)$$

and secondly, we compute

$$D_n = \max_{i=1, \dots, n} (\sigma_i) \quad (3.10)$$

The variables are drawn from a normal distribution if $D_n < D_{\alpha, n}$.

4. THE PARAMETERS

We need the following parameters for using with the algorithm

$$\mu = f(x_0), \quad (4.1)$$

$$\Delta = \frac{\bar{x} - x_0}{n-1}, \quad x_i = x_0 + i\Delta, \quad (i = 1, \dots, n-1) \quad (4.2)$$

and $\hat{\sigma}$ the maximum likelihood of $\hat{\sigma}$ as

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n-1} \frac{(f(x_i) - f(x_{i-1}))^2}{x_i - x_{i-1}} \quad (4.3)$$

5. STOPPING CRITERIA

The stopping criteria to determine the best interval explained in Section 8 for an initial approximation point can be obtained by the Theorem 2.1. Let L_n be the set of subintervals of local inadequacy of the model, after n function evaluations, performed in $x_0 \leq x_1 \leq \dots \leq x_{n-1}$. At the next step, $f(x)$ is assumed a unimodal in any of the subinterval of L_n and a sample path of the Wiener process for $x \in X - L_n$ ([6]).

Suppose that

$$P_i = P \left\{ \min_{x \in [x_i, x_{i+1}]} f(x) \leq f_n^* - \varepsilon \mid f(x_i), f(x_{i+1}) \right\}. \quad (5.1)$$

If $[x_i, x_{i+1}] \not\subset L_n$ then

$$P_i = \exp \left(-2 \frac{(f_n^* - \varepsilon - f(x_{i+1}))(f_n^* - \varepsilon - f(x_i))}{\sigma^2(x_{i+1} - x_i)} \right), \quad (5.2)$$

where $f_n^* = \min \{f(x_0), f(x_1), \dots, f(x_{n-1})\}$ and ε is the prefixed accuracy level. If $[x_i, x_{i+1}] \subset L_n$, then $P_i = 0$. Finally, we compute

$$PF_n = P \left\{ \min_{x \in X} f(x) > f_n^* - \varepsilon \mid z_n \right\} = \prod_{i=0}^{n-2} (1 - P_i) \quad (5.3)$$

Algorithm will terminate when $PF_n > PT$ where PT is a prefixed probability level and as example $PT = 0.99$. If $PF_n \leq PT$, then change the value of n .

6. THE CONTROL OF PARAMETER Γ

The sequence $\{\gamma_j\}$ can be determined from Theorem 2.1 given by

$$P_{\gamma_j} = P \left\{ \min_{x \in X} f(x) > f_n^* - \gamma_j \mid z_n \right\}. \quad (6.1)$$

γ_j is kept constant if $P_{\gamma_j} \leq P_T$ with P_T is a given fixed probability level. If $P_{\gamma_j} > P_T$ then $\gamma_{j+1} = 2/\gamma_j$.

7. THE BEST SUBINTERVAL

The probabilistic algorithm for obtaining the “best” subinterval from which we can perform the Newton’s method to find the optimal solution, is as follows:

A subinterval Δ_p is chosen such that

$$\begin{aligned} P \left\{ \min_{x \in \Delta_p} f(x) < f_{n-1}^* - \gamma \mid z_{n-1} \right\} &= \max_i P \left\{ \min_{x \in \Delta_i} f(x) < f_{n-1}^* - \gamma \mid z_{n-1} \right\} \\ &= \max_i P \left\{ \min_{x \in \Delta_i} f(x) < f_{n-1}^* - \gamma \mid f_i, f_{i+1} \right\} \\ &= \\ \max_i \exp \left(-2 \frac{(f_{n-1}^* - \gamma - f_{i+1})(f_{n-1}^* - \gamma - f_i)}{\sigma^2(x_{i+1} - x_i)} \right) \\ &= \text{size}(\Delta_p) \end{aligned} \quad (7.1)$$

where $f_{n-1}^* = \min(\mu, f_1, f_2, \dots, f_{n-1})$, and γ is some positive value.

Now, we must divide the size interval X by $\text{size}(\Delta_p)$ to obtain subinterval $\Delta_j = [x_j, x_{j+1}]$ ($j = 0, \dots, m$) where

$$m = \left\lceil \frac{\bar{x} - x_0}{\text{size}(\Delta_p)} \right\rceil. \quad (7.2)$$

Suppose that the function evaluations have been performed, such that

$$f(x_s) = \min \{f(x_0), f(x_1), \dots, f(x_m)\} \quad (7.3)$$

where s is the index of a point with minimum value. Therefore, the best subinterval is an interval which contains x_s and can be expressed by

$$I = [x_{s-1}, x_{s+1}] \quad (7.4)$$

8. UNIMODALITY TEST

If a function $f: [a, b] \subset \mathbb{R}^1 \rightarrow \mathbb{R}^1$ which attain its minimum value in the interval $[a, b] \subset \mathbb{R}^1$ at x^* is strictly decreasing for $x < x^*$ and is strictly increasing for $x > x^*$, then f is said to be unimodal in $[a, b]$ ([3],[4],[8]).

Suppose that $f: \mathbb{R}^1 \rightarrow \mathbb{R}^1$ unimodal in the best subinterval $I = [x_{s-1}, x_{s+1}]$ chosen as described in Section 7. Denote $x_1^{(0)} = x_{s-1}$ and $x_2^{(0)} = x_{s+1}$ and then evaluate the function value at two intermediate points $x_3^{(0)}$ and $x_4^{(0)}$ such that $x_1^{(0)} < x_3^{(0)} < x_4^{(0)} < x_2^{(0)}$. We can reduce the width of the interval that contains the minimizer, based on values of $f(x_3^{(0)})$ and $f(x_4^{(0)})$ according to three cases as follows.

For the case $f(x_3^{(0)}) < f(x_4^{(0)})$, the minimum value should lie to the left of $x_4^{(0)}$, and hence the subinterval $[x_1^{(1)}, x_2^{(1)}] = [x_1^{(0)}, x_4^{(0)}]$ is retained.

For the case $f(x_3^{(0)}) > f(x_4^{(0)})$, the minimum value should lie to the right of $x_3^{(0)}$, and hence the subinterval $[x_1^{(1)}, x_2^{(1)}] = [x_3^{(0)}, x_2^{(0)}]$ is retained.

For the case $f(x_3^{(0)}) = f(x_4^{(0)})$, the minimum will lie in $[x_3^{(0)}, x_4^{(0)}]$, and hence the subintervals $[x_1^{(1)}, x_2^{(1)}] = [x_3^{(0)}, x_4^{(0)}]$ is retained.

However, we do not know about the modality property of f in the subinterval $[x_1^{(0)}, x_2^{(0)}] = [x_{s-1}, x_{s+1}]$. Therefore in our implementation, if the sequence of points $x_1^{(0)}, x_1^{(1)}, x_1^{(2)}, \dots$ or $x_2^{(0)}, x_2^{(1)}, x_2^{(2)}, \dots$ such that

$f(x_1^{(i)}) > f(x_1^{(i+1)})$ or $f(x_2^{(i)}) < f(x_2^{(i+1)})$ ($i = 0, 1, 2, \dots$), then f is unimodal in that subinterval. This criterion can be achieved since the subinterval $[x_1^{(0)}, x_2^{(0)}] = [x_{s-1}, x_{s+1}]$ has been tested by the statistical tests as described in the previous section.

9. NUMERICAL RESULTS

In this section, we will give the results of statistical and unimodality tests performed to minimize

$$f(x) = \frac{1}{2} \cos(x) + \frac{1}{3} \sin(2x)$$

subject to

$$0 \leq x \leq 5.$$

The results of the statistical tests are reported in Table 1, where n denote the number of samples, T_n is the values of equation (3.7) and D_n is the value of equation (3.10). The values of $T_{0.05}$, $T_{0.01}$, $D_{0.05,n}$ and $D_{0.01,n}$ are obtained by using the table of t-distribution and the table of Critical Value for the Liliefors Test for Normality([7]).

Table 1 : Results of Randomness and Normality Tests

n	T_n	$T_{0.05}$	$T_{0.01}$	D_n	$D_{0.05,n}$	$D_{0.01,n}$
10	0.318	2.306	3.355	0.139	0.258	0.294
20	0.228	2.101	2.878	0.118	0.190	0.231
30	0.188	2.048	2.763	0.102	0.161	0.187
40	0.164	2.024	2.711	0.103	0.140	0.163
50	0.147	2.011	2.682	0.104	0.125	0.146
100	0.105	1.980	2.682	0.110	0.089	0.103

The results of the unimodality test are reported in the following table for four iterations.

Table 2 : Results of Unimodality Test

Iteration	j	x_j	$f(x_j)$
1	1	2.3970	-0.6999
	2	3.0819	-0.5388
	3	2.6253	-0.7210
	4	2.8536	-0.6610
2	1	2.3970	-0.6999
	2	2.5836	-0.6610
	3	2.5492	-0.7236
	4	2.7014	-0.7093
3	1	2.3970	-0.6999
	2	2.7014	-0.7093
	3	2.4985	-0.7200
	4	2.5999	-0.7229
4	1	2.4985	-0.7200
	2	2.7014	-0.7093
	3	2.5661	-0.7239
	4	2.6338	-0.7202

10. CONCLUSIONS

From the Table 9.1, we can see that the value of $|T_n| < T_{\alpha,n}$, says that the variables of the function are random variables, and the values of $D_n < D_{\alpha,n}$, suggests that the variables are drawn from a normal distribution. Actually, the global minimizer for the above problem is 2.5663 and the corresponding value is -0.7239.

From the Table 9.2, it confirms that the best subinterval determined by probabilistic algorithm, is unimodal.

11. ACKNOWLEDGEMENTS

The authors wish to express their appreciation to Universiti Teknikal Malaysia Melaka (UTeM) for their financial support to attend and present this paper.

REFERENCES

1. Archetti, F and Betro, B. *A Probabilistic Algorithm For Global Optimization* (to appear in *Calcolo*, 1978).
2. Archetti, F. 1980, *Analysis of Stochastic Strategies For The Numerical Solution of The Global Optimization Problem*. Article on Numerical Techniques For Stochastic System, North-Holland.
3. Avriel, M. 1976. *Nonlinear Programming*. New jersey; Prentice –hall, Inc.
4. Mohan, C.J and Kannan, M.M. 2004, *Optimization; Theory and Practice*, Harraow, U.K
5. Rao, S. S. 2002, *Optimization Theory and Application*, Wiley Eastern
6. Sheldon M. Ross. 1993, *Introduction To Probability Models*, San Diego.
7. Sheskin, J. D. 2000, *Handbook of Parametric and Nonparametrik Statistical Procedures*, London.
8. Wismer, D. A dan Chattergy, R. *Introduction To Nonlinear Optimization A Problem Solving Approach* (New York, 1978).

Yosza Dasril: Department of Industrial Electronics, Faculty of Electronics and Computer Engineering, Universiti Teknikal Malaysia Melaka (Utem),
Locked Bag 1200, Ayer Keroh, Hang Tuah Jaya, 75450 Melaka
E-mail : yosza@utem.edu.my

Ismail Bin Mohd: Department of Mathematics, Faculty of Science and Technology, Kolej Universiti Sains dan Teknologi Malaysia
21030 Mengabang Telipot, Kuala Terengganu, Malaysia
E-mail: ismailmd@kustem.edu.my

Siska Candra Ningsih: Department of Mathematics, Faculty of Science and Technology, Kolej Universiti Sains dan Teknologi Malaysia
21030 Mengabang Telipot, Kuala Terengganu, Malaysia
E-mail: siskaningsih@yahoo.com.

GENERATING SQUARE-ROOT DIFFUSIONS PROCESS BY MONTE CARLO SIMULATION WITH APPLICATION IN SYAREEA INSURANCE

ADHITYA RONNIE EFFENDIE, SUBANAR, DANARDONO

Abstract. In this paper we describe method of Monte Carlo Simulation to generate square-root diffusions process such as Cox-Ingersoll-Ross (CIR) model. This CIR model can be useful to price financial instrument such as Option or Net Single Premium in Actuarial calculation. Because the Stochastic Differential Equation (SDE) of CIR model is not explicitly solvable, we use the fact that the distribution of the solution, given its initial value and up to a scale factor, will be a noncentral chi-square distribution. This property can be used to simulate square-root diffusions process.

Key words and Phrases : Monte Carlo method, Square-Root Diffusions Process, Stochastic Differential Equation

1. INTRODUCTION

The interest rate has an important place in all transactions that include simply lending and borrowing. However, its importance has risen with developing and slight stationary economic conditions of the world. The desire to have foresight for the level of future interest rate has become crucial in the sense that to know the lending and borrowing rate and not to missprice the interest rate instruments. As a result, the modelling of interest rate has risen as a problem. As solution to this problem there have been many models proposed.

Their common feature is that they all model a stochastic problem, i.e. they deal with uncertainty. The first model in 1973 was proposed by Merton. However, the pioneering one was suggested by Vasicek in 1977 and in the following years many other models that are much more analytically untractable has come out. These stochastic models are mainly classified with respect to the number of factors which are assumed to have a stochastic evolution in the model.

In Islam, there are several rules which have to obeyed in our daily financial transactions. One of them is we have to keep all possible return of financial transactions in the future, including the interest rate, uncertain. Thus, locking the rate of return into a particular fixed interest rate for long period of time is prohibited. As a solution of this problem we can use a stochastic model of interest rate for describing the uncertain evolution of financial transactions return in the future.

In this paper, we want to show that after we can unlock the fixed interest rate for some periods of time then we can directly put this stochastic interest rate on the calculation of future benefit of a life insurance product with a fixed premium. We call this insurance Syareea.

Thus, the main difference between a conventional life insurance is that the Syareea insured do not know how much exactly he will be paid for the benefit because it depends on financial situation of the insurance company. This method would be slightly different from “defined contribution” method in Pension fund analysis.

2. INSTANTANEOUS FORWARD RATE

Suppose that at time t we have two spot rates with maturity T_1 and T_2 respectively, i.e. $R(t, T_1)$ and $R(t, T_2)$, with $t < T_1 < T_2$. Then we will have a forward rate between T_2 and T_1 at time t , $f(t, T_1, T_2)$ as follows:

$$e^{(T_2-t)R(t, T_2)} = e^{(T_1-t)R(t, T_1)} \times e^{(T_2-T_1)f(t, T_1, T_2)}. \quad (1)$$

Define a zero-coupon bond price $b(t, T) = e^{-R(t, T)}$ and instantaneous rate of

$$\text{return } r(t) = \lim_{\Delta t \rightarrow 0} \frac{R(t, t + \Delta t) - R(t, t)}{\Delta t} = -\frac{\partial}{\partial T} \ln b(t, T) \Big|_{T=t} = -\frac{\partial}{\partial T} \ln b(t, t) \quad (2)$$

then we have the following instantaneous forward rate and current short rate respectively:

$$f(t, T) = \lim_{\Delta t \rightarrow 0} \frac{R(t, T + \Delta t) - R(t, T)}{\Delta t} = -\frac{\partial}{\partial T} \ln b(t, T) \quad (3)$$

$$f(t, t) = R(t, t) = r(t). \quad (4)$$

Under Heath-Jarrow-Morton (HJM) framework, for given T , (3) are stochastic processes in t . Thus we will begin our analysis of syareea insurance by choosing a stochastic differential equation (SDE) for the forward rates.

3. THE COX-INGERSOLL-ROSS MODEL

Given an initial forward rate curve, $f(0, T)$, (which is integrable as a function of T), the forward curve evolves as

$$df(t, T) = \alpha(t, T)dt + \sigma(t, T)dW(t). \quad (5)$$

The integral form of (5) can be written as

$$f(t, T) = f(0, T) + \int_0^t \alpha(s, T)ds + \int_0^t \sigma(s, T)dW(s). \quad (6)$$

Hence, its corresponding current short rate is

$$r(t) = f(t, t) = f(0, t) + \int_0^t \alpha(s, t) ds + \int_0^t \sigma(s, t) dW(s). \quad (7)$$

In this paper we choose Cox-Ingersoll-Ross (CIR) model of interest rate. Under this model, the interest rate has the following stochastic differential equation:

$$dr(t) = a(b - r(t))dt + \sigma\sqrt{r(t)}dW(t). \quad (8)$$

This SDE (a square root diffusion process) is not explicitly solvable and we will use Monte Carlo simulation to overcome this problem. Fortunately, unlike its current short rate, the bond price of CIR model is analytically solvable. Now, we are calculating the bond price of the CIR model. As in [7] the bond price function

$$B(t, T) = E\left(\exp\left(-\int_t^T r(u)du\right) \middle| r(t)\right) \quad (9)$$

is in the form $B(t, T) = \exp(-A(t, T)r(t) + C(t, T))$, another example of an affine yield model, which has solutions:

$$A(t, T) = \frac{2(\exp(c(T - t)) - 1)}{(a + c)(\exp(c(T - t)) - 1) + 2c}$$

and

$$C(t, T) = \frac{2ab}{\sigma^2} \ln\left(\frac{2\gamma \exp((a + c)(T - t) / 2)}{(a + c)(\exp(c(T - t)) - 1) + 2c}\right)$$

with $c = \sqrt{a^2 + 2\sigma^2}$.

4. THE MONTE CARLO SIMULATION

To simulate (8) we will need knowledge of transition density for the process. Based on results of Feller [4], Cox et al [3] noted that the distribution of $r(t)$ given $r(u)$ for some $u < t$ is, up to a scale factor, a noncentral chi-square distribution. The transition law of $r(t)$ in (8) can be expressed as

$$r(t) = \frac{\sigma^2(1 - \exp(-a(t - u)))}{4a} \chi_d^2\left(\frac{4a \exp(-a(t - u))}{\sigma^2(1 - \exp(-a(t - u)))} r(u)\right), t > u$$

where $d = \frac{4ab}{\sigma^2}$. This say that, given $r(u)$, $r(t)$ is distributed as

$\sigma^2(1 - \exp(-a(t - u))) / 4a$ times a noncentral chi-square random variable with d degree of freedom and noncentrality parameter

$$\lambda = \frac{4a \exp(-a(t-u))}{\sigma^2 (1 - \exp(-a(t-u)))} r(u)$$

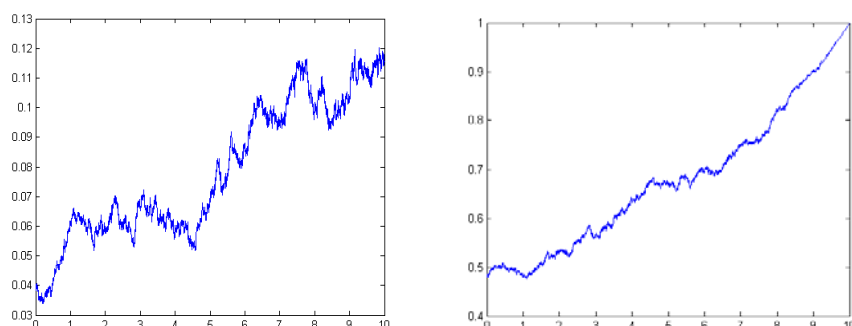


Figure 1. $r(t)$ sample path with $a=0.2, b=5\%, \sigma=0.1, r(0)=0.4$ (left) $B(t,10)$ sample path with $a=0.2, b=5\%, \sigma=0.1, r(0)=0.4$ (right)

5. APPLICATION: SYAREEA INSURANCE

Suppose that A (25 years old) goes to Insurance company giving amount of money, say 10 million rupiahs, and under specific circumstances he will promised a benefit that are not exactly known, but he will get information, on average, how much he will be paid if the claim occur. Furthermore, assume that this man enroll a 10 years term insurance, providing stochastic benefit when he die up to 10 years from the contract. The objective is to determine the stochastic benefit of the contract.

Suppose that the Insurance company uses the US 1979-81 Life Table which has been extrapolated using Gompertz survival function. Another assumptions macro condition (CIR parameters) just like the example ($a=0.2, b=5\%, \sigma=0.1, r(0)=0.4$).

After simulation take a place, we have ten possible outcomes of the benefit, (in million rupiahs):

$$\text{inv} \left(A_{\frac{1}{25:10}} \right) = \begin{bmatrix} 25.63 & 18.67 & 21.28 & 23.78 & 20.51 \\ 21.46 & 23.25 & 22.53 & 22.35 & 20.45 \end{bmatrix}$$

So this man will have around 180 million rupiahs for the benefit, if claim occurs up to 10 years from the contract, but he has a chance to get more than that.

REFERENCES

1. Fabozzi, Frank j., 2002, *Interest Rate, term structure, and valuation modeling*, John Wiley & Sons, Inc., Hoboken, New Jersey.
2. Cox, John C., Ingersoll, Jonathan E., Ross, Stephen A., *A Theory of the Term Structure of Interest Rates*, *Econometrica*, Vol. 53 No.2 (Mar., 1985).pp.385-408
3. Colwell, David, 2007, *Financial Mathematics*, Diktat kuliah ICE-AMSI Summer School, University of Sydney, Sydney
4. Feller, W. 1951, Two singular difussion problems, *Annals of Mathematics* 54:173-182.

5. Glasserman, Paul, 2004, *Monte Carlo Methods in Financial Engineering*, Springer, New York.
6. Shreve, E. Shreve, 2004, *Stochastic Calculus for Finance I: The Binomial Asset Pricing Model*, Springer, New York
7. Shreve, E. Shreve, 2004, *Stochastic Calculus for Finance II: Continuous Time Model*, Springer, New York
8. Hoek, John Van der, 2007, *Stochastic Calculus*, Diktat kuliah ICE-AMSI Summer School, University of Sydney, Sydney

Adhitya Ronnie Effendie: Department of Mathematics, Gadjah Mada University, Yogyakarta, Indonesia.
E-mail: adhityaronnie@ugm.ac.id.

BINOMIAL METHOD FOR INDONESIAN OPTION PRICING

GUNARDI, SUBANAR, J.A.M. VAN DER WEIDE AND SRI HARYATMI

Abstract. The Jakarta Stock Exchange (JSX) of Indonesia has started to trade an option at September 9th, 2004. The option can be considered as an American style barrier option with automatic (immediate) exercise if the price hits or crosses the barrier before maturity. The payoff of the option is based on a Weighted Moving Average (WMA) of the price of the underlying stock. The barrier is fixed at the strike price plus or minus a 10 percent. The option is automatically exercised when the underlying stock hits or crosses the barrier and the difference between the strike and barrier is paid immediately. We will refer to type of this option as an Indonesian option. Based on the exercise time, there are two types of option: an Indonesian option with American type and an Indonesian option with European type. So the options which are traded in the JSX are called the American Indonesian options.

This paper deals with the pricing of the Indonesian option for both exercise time style when WMA is replaced by stock price. As the model of stock price, we use Black-Scholes (BS) model. The price of European Indonesian option can be found analytically by Expected payoff method [2] and numerically by Simulation Monte Carlo method [4] and Partial Differential Equation method [3]. The price of American Indonesian option can be found numerically by Partial Differential Equation method [3]. In this paper, the price of Indonesian option for both exercise time style are calculated by Binomial method.

Key words and Phrases: Indonesian option, WMA, Binomial

1. INTRODUCTION

In Indonesia, September 9, 2004, the Jakarta Stock Exchange (JSX) ¹ introduced the trading of options on stocks of five companies: Telekomunikasi Indonesia Tbk (TLKM), Astra International Tbk (ASII), HM Sampoerna Tbk (HMSP),

¹For more information, see the JSX home page: <http://www.jsx.co.id>

Bank Central Asia Tbk (BBCA) and Indofood Sukses Makmur Tbk (INDF). The regulations for trading in options are available on the JSX web site. These regulations contain, among others, descriptions of put and call option contracts with the stock of one of the above-mentioned companies as underlying. Since these contracts are rather special, we will refer to them as **Indonesian options**. As usual, a strike price K and a maturity date T are specified in these contracts. For example an AHMSP4100 contract is a call option contract with maturity date January (A = January, B=February,...,L=December) and with strike 4100 written on HMSP stocks (HM Sampoerna Tbk). The letters O,P,...,Z are used to code the maturity date for put option contracts. So the put option contract with maturity January on the same stock and with the same strike is coded by OHMSP4100.

Every 15 minutes, a Weighted Moving Average Price (WMA) of the stock-price over the last 30 minutes is published: Monday, starting at 10.00.01 the WMA over the time-period 9.30-10.00

$$P_{(10.01-10.15)} = \frac{\sum_{t=9.30}^{t=10.00} S_t Q_t}{\sum_{t=9.30}^{t=10.00} Q_t},$$

where S_t is the stock price at time t , and Q_t the transaction volume. At 10.15.01 the next WMA price over the time-period 9.45-10.15 is published, and so on. The last WMA price of the morning (section 1) is published at 12.00.01, except on Friday. On Friday, the last WMA price publication in the morning is at 11.30.01. In the afternoon (section 2) the first publication is at 14.15.01 and the last one at 16.00.01. So Monday-Thursday are divided in 17 periods of 15 minutes between consecutive publications of WMA prices. These periods are numbered 1,2,...,17. Friday has 15 periods.

The option contract can be exercised anytime during the period of stock option trading, till maturity T or till the WMA price hits or crosses the barrier $B = 0.9K$ for a put, $B = 1.1K$ for a call. If the option contract is exercised at time t and if the WMA price has not hit or crossed the barrier, then the payoff is equal to the difference between the strike and the last published WMA price, $K - \text{WMA}$ for a put, $\text{WMA} - K$ for a call. This payoff is payed the next day. If the WMA price hits or crosses the barrier, then the option contract is automatically exercised and the owner of option contract gets immediately $K - B$ in case of a put and $B - K$ in case of a call. The option contract can then be considered as an American barrier option with automatic exercise.

As we know that Indonesian option is a modification of American option with a barrier and an automatic exercise when the stock price hit the barrier. The price of Indonesian option is divided in two terms: the intrinsic value and the time value, see [7] and [6]. This price is not fair because the behavior of the stock price is not included on the calculation.

There are some important result that are related to calculate the fair price of Indonesian options. Higham [5] shows how to price European options and American options by Expected payoff method, Binomial methods, Monte Carlo simulation methods and Partial Differential Equation approach. Wilmott [8] shows how to

use finite difference method to solve PDE numerically. Broadie and Glasserman [1] use a continuous correction for discrete barrier options to get better result in simulation.

This paper deals with the pricing of the Indonesian option for both exercise time style when WMA is replaced by stock price. As the model of stock price, we use Black-Scholes (BS) model. The price of European Indonesian option can be found analytically by Expected payoff method [2] and numerically by Monte Carlo Simulation method [4] and Partial Differential Equation method [3]. The price of American Indonesian option can be found numerically by Partial Differential Equation method [3].

In this paper we will apply Binomial method to calculate the price of Indonesian call option. The derivation of this method is based on Higham [5]. The binomial method uses a discrete model for the asset price.

- Time-grid. Let T be the expiry time and $\delta t = T/M$ the spacing between successive time points. So asset prices will considered at times $t_i = i\delta t$, for $0 \leq i \leq M$.
- Price movement. Between successive time points the asset price moves either up by a factor u or down by a factor d . An upward move occurs with probability p and a downward move occurs with probability $1 - p$.

At time t_i there are $i + 1$ possible asset prices, which we label by

$$S_n^i = d^{i-n} u^n S_0, \quad n = 0, 1, \dots, i. \quad (1)$$

Deriving the parameters

Let the random variable $R_i = 1$ if the asset price goes up from time t_{i-1} to time t_i and $R_i = 0$ if the asset price goes down from time t_{i-1} to time t_i . Hence

$$R_i = \begin{cases} 0 & \text{with probability } 1 - p, \\ 1 & \text{with probability } p. \end{cases} \quad (2)$$

It follows that $E(R_i) = p$ and $Var(R_i) = p(1 - p)$. After n time steps the asset has undergone $\sum_{i=1}^n R_i$ upward movements and $n - \sum_{i=1}^n R_i$ downward movements. So the asset price $S(t_n)$ at time $t_n = n\delta t$ is given by

$$S(t_n) = S_0 u^{\sum_{i=1}^n R_i} d^{n - \sum_{i=1}^n R_i} = S_0 d^n \left(\frac{u}{d}\right)^{\sum_{i=1}^n R_i}. \quad (3)$$

Taking $\ln$ gives

$$\ln \left(\frac{S(t_n)}{S_0} \right) = n \ln d + \ln \left(\frac{u}{d} \right) \sum_{i=1}^n R_i. \quad (4)$$

By the Central Limit Theorem

$$\sum_{i=1}^n R_i \approx N(np, np(1 - p)), \quad (5)$$

so $\ln(S(t_n)/S_0)$ will be approximately normal with mean

$$n \ln d + np \ln \left(\frac{u}{d} \right) = n(1-p) \ln d + np \ln u, \quad (6)$$

and variance

$$\ln^2 \left(\frac{u}{d} \right) np(1-p). \quad (7)$$

To match the continuous asset price model

$$S(t) = S_0 e^{(\mu - \frac{1}{2}\sigma^2)t + \sigma\sqrt{t}Z}, \quad (8)$$

is used in the Black-Scholes model, we require

$$(1-p) \ln d + p \ln u = (r - \frac{1}{2}\sigma^2)\delta t, \quad (9)$$

$$\ln^2 \left(\frac{u}{d} \right) p(1-p) = \sigma^2 \delta t. \quad (10)$$

Here we imposed the risk neutrality assumption, since the binomial method works with expected values. Regarding $\delta = T/M$ as pre-specified, we now have two equations in the three unknowns p, u and d . Solving equation 9 and 10 for $p = \frac{1}{2}$, we get

$$u = e^{\sigma\sqrt{\delta t} + (r - \frac{1}{2}\sigma^2)\delta t} \quad \text{and} \quad d = e^{-\sigma\sqrt{\delta t} + (r - \frac{1}{2}\sigma^2)\delta t} \quad (11)$$

2. MAIN RESULTS

Based on the exercise time, there are two types of option: an Indonesian option with American type and an Indonesian option with European type. So the options which are traded in the JSX are called the American Indonesian options.

This section deals with the pricing of the Indonesian option for both exercise time style when WMA is replaced by stock price. As the model of stock price, we use Black-Scholes (BS) model. The price of Indonesian option is calculated by Binomial method.

2.1. Indonesian Call Option

In order to price these Indonesian call options with early-exercise feature we use a binomial tree method. It is common practice to use the recombining tree, which assures that if we have M layers we have i stock prices $S_n^i, 1 \leq n \leq i$ at layer i . The calculation procedure is backwards, starting from the pay-off at layer M prices $\Lambda(S_n^i), 1 \leq n \leq M-1$ are calculated for layer $M-1$ and so on to the beginning of the tree. In order to adapt the calculation method for the Indonesian option, we have to check in every node n at layer i if the stock price process is bigger than B . As hitting B means automatic exercise of the call we have to replace in each step all the calculated option prices belonging to stock price levels bigger than B .

European type

In order to price the Indonesian call options with automatic exercise, asset prices in (1) are modified when the prices hit the barrier, then we have

$$S_n^M = \begin{cases} S_n^M, & \text{if } S_n^M < B \\ B, & \text{otherwise} \end{cases} \quad \text{for } n = 0, 1, \dots, M. \quad (12)$$

and for $i = 0, 1, \dots, M-1$,

$$\Lambda(S_n^i) = \begin{cases} 0, & \text{if } S_n^i < B \\ B - K, & \text{otherwise} \end{cases} \quad \text{for } n = 0, 1, \dots, i. \quad (13)$$

Hence, choosing the best of the two possibilities leads to the relation

$$V_n^i = \max [\Lambda(S_n^i), e^{-r\delta t} (pV_{n+1}^{i+1} + (1-p)V_n^{i+1})]. \quad (14)$$

where $\Lambda(S_n^i) = \max(S_n^i - K, 0)$.

Based on (12), (13) and (14), we compute the time-zero option price V_0^0 , so $V_{eic} = V_0^0$.

M	V_{eic}	T(seconds)
100	57.3768	0.005
1000	59.5302	0.180
10000	60.1571	7.342
20000	60.2363	30.053
30000	60.2628	71.423
40000	60.2855	133.327
60000	60.3111	323.847
80000	60.3259	656.174
100000	60.3357	1081.860

Table 1: The price of European Indonesian call option

American type

In order to price the Indonesian call options with early-exercise feature, asset prices in (1) are modified when the prices hit the barrier, then we have

$$S_n^i = \begin{cases} S_n^i, & \text{if } S_n^i < B \\ B, & \text{otherwise} \end{cases} \quad \text{for } n = 0, 1, \dots, i, \quad i = 0, 1, \dots, M. \quad (15)$$

Hence, choosing the best of the two possibilities leads to the relation

$$V_n^i = \max [\Lambda(S_n^i), e^{-r\delta t} (pV_{n+1}^{i+1} + (1-p)V_n^{i+1})]. \quad (16)$$

where $\Lambda(S_n^i) = \max(S_n^i - K, 0)$.

Based on (15) and (16), we compute the time-zero option price V_0^0 , so $V_{aic} = V_0^0$.

Numerical result

The price of the Indonesian call option contract INDF1000 on stocks of Indofood Sukses makmur Tbk with current stock price $S_0 = 1000$, strike price $K = 1000$, maturity time $T = 0.25$, interest rate $r = 0.09$, volatility $\sigma = 0.34$, by Expected payoff method (analytic formula)[2] is $V_{eic} = 60.4221$.

The price of Indonesian call option with European exercise type (Table 1) and American exercise type (Table 2) are calculated by Binomial method for M time steps.

From table 1, The price of European Indonesian call option for Binomial method and expected method (analytic formula) are very closed for $M = 100000$. It takes 1081.860 seconds or 18.031 minutes.

From Table 2, The price of American Indonesian call option for Binomial method (V_{aic}) and expected method (analytic formula) (V_{eic}) are very closed for $M = 100000$. It takes 1075.395 seconds or 17.9 minutes.

M	V_{aic}	T(seconds)
100	58.5513	0.007
1000	59.9255	0.152
10000	60.2883	7.329
20000	60.3318	32.702
30000	60.3461	73.413
40000	60.3569	139.279
60000	60.3693	324.205
80000	60.3769	645.126
100000	60.3814	1075.395

Table 2: The price of American Indonesian call option

2.2. Indonesian Put Option

In order to price these Indonesian put options with early-exercise feature we use a binomial tree method. It is common practice to use the recombining tree, which assures that if we have M layers we have i stock prices $S_n^i, 1 \leq n \leq i$ at layer i . The calculation procedure is backwards, starting from the pay-off at layer M prices $\Lambda(S_n^i), 1 \leq n \leq M-1$ are calculated for layer $M-1$ and so on to the beginning of the tree. In order to adapt the calculation method for the Indonesian option, we have to check in every node n at layer i if the stock price process is below B . As hitting B means automatic exercise of the put we have to replace in each step all the calculated option prices belonging to stock price levels below B .

European type

This replacement of calculated prices $\Lambda(S_n^i), 1 \leq n \leq i$ at layer i can be done in two ways:

$$S_n^M = \begin{cases} S_n^M, & \text{if } S_n^M > B \\ B, & \text{otherwise} \end{cases} \quad n = 0, 1, \dots, M. \quad (17)$$

and for $i = 0, 1, \dots, M-1$,

$$\Lambda(S_n^i) = \begin{cases} 0, & \text{if } S_n^i > B \\ K - B, & \text{otherwise} \end{cases} \quad n = 0, 1, \dots, i. \quad (18)$$

Hence, choosing the best of the two possibilities leads to the relation

$$V_n^i = \max [\Lambda(S_n^i), e^{-r\delta t} (pV_{n+1}^{i+1} + (1-p)V_n^{i+1})]. \quad (19)$$

where $\Lambda(S_n^i) = \max(K - S_n^i, 0)$.

Based on (17), (18) and (19), we compute the time-zero option price V_0^0 , so $V_{eip} = V_0^0$.

M	V_{eip}	T (seconds)
100	47.9626	0.034
1000	53.3021	0.135
10000	53.7297	7.555
20000	53.7993	30.949
30000	53.8350	72.860
40000	53.8569	137.046
60000	53.8800	313.984
80000	53.8936	649.855
100000	53.9042	1111.580

Table 3: The price of European Indonesian put option

American type

The contract specification of the options traded on the JSX dictate that it is possible to exercise the option at any moment. Therefore the Indonesian option value is equivalent to the value of an American knock-out option, which you would exercise before the stock hits the barrier, giving you $B - K$. In order to price the Indonesian put options with early-exercise feature, asset prices in (1) are modified when the prices hit the barrier, then we have

$$S_n^i = \begin{cases} S_n^i, & \text{if } S_n^i > B \\ B, & \text{otherwise} \end{cases} \quad n = 0, 1, \dots, i, \quad i = 0, 1, \dots, M. \quad (20)$$

Hence, choosing the best of the two possibilities leads to the relation

$$V_n^i = \max [\Lambda(S_n^i), e^{-r\delta t} (pV_{n+1}^{i+1} + (1-p)V_n^{i+1})]. \quad (21)$$

where $\Lambda(S_n^i) = \max(K - S_n^i, 0)$.

Based on (20) and (21), we compute the time-zero option price V_0^0 , so $V_{aip} = V_0^0$.

Numerical result

The price of the Indonesian put option contract VINDF1000 on stocks of Indofood Sukses makmur Tbk with current stock price $S_0 = 1000$, strike price $K = 1000$, maturity time $T = 0.25$, interest rate $r = 0.09$, volatility $\sigma = 0.34$, by analytic method [2] is $V_{eip} = IDR53.9859$.

Base on Table 3 and 4, The price of Indonesian put option by Binomial method for European and American type are not different and closed to the analytic price when $M = 100000$, It means that for $B = 0.9 * K$ the European type and the American type are not different.

M	Option price (IDR)	T (seconds)
100	53.1195	0.008
1000	53.8134	0.227
10000	53.9225	9.273
20000	53.9510	39.529
30000	53.9623	89.811
40000	53.9675	164.388
60000	53.9749	396.408
80000	53.9797	781.285
100000	53.9827	1314.295

Table 4: The price of American Indonesian put option

It is now interesting to compare the American Indonesian Put with the European Indonesian Put, the standard European put and the standard American put. Figure 1 show the different prices.

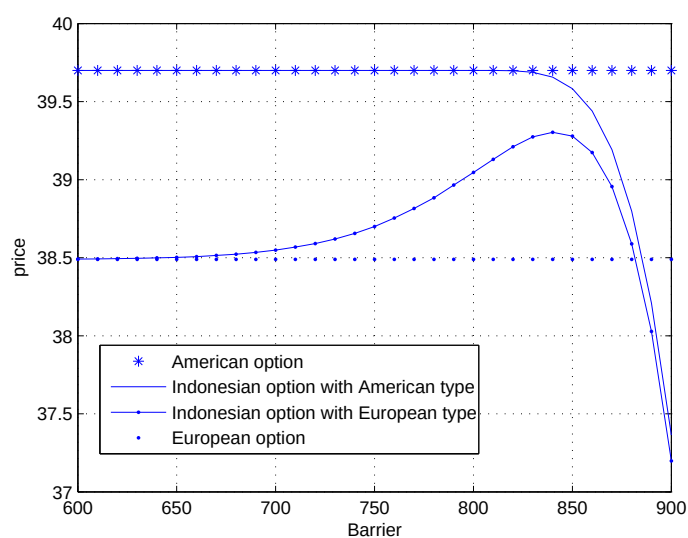


Figure 1: Numerical Example for Indonesian Put Option

3. CONCLUDING REMARKS

The Indonesian American put value is not so different from the Indonesian European put if the barrier is not too low. For some points the European Indonesian put is decreasing if the barrier is decreasing, its American counterpart is not, because by decreasing the barrier, the number of possible exercise strategies increases and therefore its value.

Acknowledgement. I wish to thank J.H.M. Anderluh, C.W. Oosterlee and Ferry Jaya Permana for their support to my research during my stay at The Delft University of Technology. In this occasion I also would like to thank H.P.S. Althuis, the director of Center for International Cooperation in Applied Technology for his help to get financial support from the Royal Netherlands Academy of Arts and Science (KNAW).

I also thank the Royal Netherlands Academy of Arts and Science (KNAW) for the financial support during my stay for doing research in the Netherlands.

REFERENCES

1. BROADIE, M. AND GLASSERMAN, P., A Continuity Correction for Discrete Barrier Options, *Mathematical Finance* **7**, 325-348, 1997.
2. GUNARDI, ANDERLUH, J. H. M., SUBANAR, VANDER WEIDE, J. A. M. AND HARYATMI, S., Indonesian Options, *Technical report, Delft University of Technology, The Netherlands*, 2006.
3. GUNARDI, SUBANAR, VANDER WEIDE, J. A. M. AND HARYATMI, S., P(I)DE approach for Indonesian Option pricing, *The 5th National Symposium on Financial mathematics ICE-EM*, University of Melbourne, Australia, 2006.
4. GUNARDI, SUBANAR, VANDER WEIDE, J. A. M. AND HARYATMI, S., Indonesian Call Option pricing by Monte Carlo Simulation Methods, *International Conference on Science and Technology* University of Technology Mara, Malaysia, 2006.
5. HIGHAM, D. J., *An Introduction to Financial Option Valuation*, Cambridge University Press, Cambridge, 2004.
6. HULU, E., Kontrak Opsi Saham (KOS) Instrumen Investasi, *Seminar Nasional Matematika V FMIPA Universitas Gadjah Mada*, 2006.
7. VEALE, S. T., *Stocks Bonds options futures*, New York Institute of Finance, 2001.
8. WILMOTT, P., *Derivatives*, Wiley, New-York, 1998.

GUNARDI: Department of Mathematics, Gadjah Mada University, Yogyakarta, Indonesia.
E-mail: gunardi@ugm.ac.id.

REGIME SWITCHING GARCH (AN APPLICATION TO DOWJONES INDEX RETURN)

M. ARBI HADIYAT, NUR IRIAWAN, SLAMET MULYONO

Abstract. Since firstly proposed by Engle (1982) and Bollerslev (1986), ARCH-GARCH models have been used to describe volatility behaviors of time series, especially in stock market analysis. One of the weaknesses of ARCH-GARCH is its inability to model behavior transition between high volatilities and low volatilities. In this research, markov switching GARCH is investigated and applied to capture the presence of different volatility regimes, i.e. low volatilities regime and high volatility regime in Dowjones index return. However, there is no information to decide which observations belong to each of the regimes, and to account this difficulty, EM algorithm is applied for parameter estimation. The result shows that Dowjones index return includes two volatility regimes. The transition matrix of the model yields that low volatility regime is often happened than the high one.

Key words and Phrases : *GARCH, Markov Chain, Switching model, Mixture Normal, EM-Algorithm.*

1. INTRODUCTION

Over past ten years since it has been proposed by Engle (1982) and Bollerslev (1986), ARCH and GARCH were known as advance and accurate methods to describe the volatility behavior of asset's return data. However, considering the transition between the change of low volatility and high volatility regimes, common GARCH model needs to be developed to capture such regime switching. Proposed firstly by Hamilton (1989), regime switching time series model has been developed to analyze such jumps between regimes and then followed by regime switching ARCH as in Hamilton (1994). Finally, Gray (1996) has succeeded to implement regime switching GARCH to weekly Treasury Bill rates data, and followed by Marcucci (2005) with *t*-student and generalized error innovations included in model.

2. REGIME SWITCHING GARCH

The main feature of regime switching models is the possibility for some, or all, the parameter to switch across different regimes according to a Markov chain process, governed by unobservable state variable S_t , with transition probability

$$\Pr(S_t = j | S_{t-1} = i) = p_{ij}, \quad i, j = 1, 2$$

and transition matrix

$$\begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} = \begin{bmatrix} P & (1-P) \\ (1-Q) & Q \end{bmatrix}$$

The regime switching GARCH consist of four elements, conditional mean, conditional variance, regime process and conditional distributions. The specific form of regime switching GARCH(1,1) has conditional mean

$$z_t = \delta_i + \varepsilon_{it}, \quad i = 1, 2$$

with the error term follows the mixture normal distribution

$$\varepsilon_{it} | \zeta_{t-1} \sim \begin{cases} N(0, h_{1t}) & \text{with probability } p_{1t} \\ N(0, h_{2t}) & \text{with probability } p_{2t} \end{cases},$$

then the conditional mean will be as follows

$$z_t | \zeta_{t-1} \sim \begin{cases} N(\delta_1, h_{1t}) & \text{with probability } p_{1t} \\ N(\delta_2, h_{2t}) & \text{with probability } p_{2t} \end{cases}, \quad (1)$$

and $\zeta_{t-1} = \{z_{t-1}, z_{t-2}, \dots, z_1\}$ represent the information set to time $t-1$. Conditional variance of the model is

$$h_{it} = \omega_i + \alpha_i \varepsilon_{t-1}^2 + \beta_i h_{t-1}, \quad i = 1, 2$$

In order to avoid the regime path dependence of past error ε_{t-1}^2 and conditional variance h_{t-1} , Gray (1996) proposed to replace them by

$$\begin{aligned} \varepsilon_t^2 &= z_t - E[z_t | \zeta_{t-1}], \\ h_t &= E[z_t^2 | \zeta_{t-1}] - [E(z_t | \zeta_{t-1})]^2, \end{aligned}$$

Where

$$\begin{aligned} E[z_t | \zeta_{t-1}] &= \sum_{i=1}^2 p_{it} \mu_{it}, \quad i, j = 1, 2 \\ E[z_t^2 | \zeta_{t-1}] &= \sum_{i=1}^2 p_{it} (\mu_{it}^2 + h_{it}), \quad i, j = 1, 2 \end{aligned} \quad (2)$$

then the model doesn't involve the regime path dependence.

The next element is the switching process. The *ex-ante probability* p_{it} (Gray,1996) represents the proportions of mixture distribution which can be reformulated by

$$\begin{aligned} p_{it} &= \Pr(S_t = i | \zeta_{t-1}) = \sum_{j=1}^2 \Pr(S_t = i, S_{t-1} = j | \zeta_{t-1}) \\ &= \sum_{j=1}^2 \Pr(S_t = i | S_{t-1} = j, \zeta_{t-1}) \cdot \Pr(S_{t-1} = j | \zeta_{t-1}), \end{aligned} \quad (3)$$

The first term in (3) is the way the markov chain transition probability involved in regime switching models. Further, the second probability in (3) can be derived according to Bayes formula

$$\begin{aligned} \Pr(S_{t-1} = i | \zeta_{t-1}) &= \Pr(S_{t-1} = i | \zeta_{t-1}, \zeta_{t-2}) = \frac{f(z_{t-1}, S_{t-1} | \zeta_{t-2})}{f(z_{t-1} | \zeta_{t-2})} \\ &= \frac{f(z_{t-1} | S_{t-1} = i, \zeta_{t-2}) \cdot \Pr(S_{t-1} = i | \zeta_{t-2})}{\sum_{i=1}^2 f(z_{t-1} | S_{t-1} = i, \zeta_{t-2}) \cdot \Pr(S_{t-1} = i | \zeta_{t-2})}. \end{aligned} \quad (4)$$

The probability in (4) is commonly reported in regime switching literature as *filtered probability*, i.e. the probability of being in regime i based on information set to time t .

The likelihood can be constructed in common way based on (1) and yields

$$\ln L(\theta | z_t, \zeta_{t-1}) = \sum_{t=1}^T \ln \left[\sum_{i=1}^2 f(z_t | S_t = i, \zeta_{t-1}) \cdot p_{it} \right],$$

with p_{it} follows (3). The unobserved S_t affect the likelihood optimization to become unfulfilled. An alternative way to optimize the likelihood is using the Expectation – Maximization (EM) algorithm first proposed by Dempster, Laird, and Rubin (1977). The expectation step yields

$$\begin{aligned} \mathbb{Q} &= E \left[\ln L(\theta | z_t, S_t, \zeta_{t-1}) \mid z_t, \zeta_{t-1} \right], \\ &= \sum_{t=1}^T \sum_{j=1}^2 \varphi_{it} \left[\ln p_{it} - \frac{1}{2} \ln(2\pi h_{it}) - \frac{(z_t - \delta_i)^2}{2h_{it}} \right]. \end{aligned} \quad (5)$$

to be solved numerically using standard optimization algorithm such as Newton's method.

3. REGIME SWITCHING GARCH FORECASTING

Regime switching is assumed to be exist in times forward, and the probability of being in certain regime at time $t+f$ is computed using markov chain property, that is

$$\begin{bmatrix} \Pr(S_{t+f} = 1 | \zeta_{t-1}) \\ \Pr(S_{t+f} = 2 | \zeta_{t-1}) \end{bmatrix} = \begin{bmatrix} p_{1t} \\ 1 - p_{1t} \end{bmatrix} \begin{bmatrix} P & 1-P \\ 1-Q & Q \end{bmatrix}^f = \begin{bmatrix} p_{1t} \\ 1 - p_{1t} \end{bmatrix} \mathbf{P}^f$$

and prediction for conditional mean is based on the expectation in (2) for time $t+f$

$$\begin{aligned} E[z_{t+f} | \zeta_{t-1}] &= \begin{bmatrix} \mu_{1t+f} & \mu_{2t+f} \end{bmatrix} \begin{bmatrix} p_{1t+f} \\ 1 - p_{1t+f} \end{bmatrix} \\ &= \begin{bmatrix} \hat{\delta}_1 & \hat{\delta}_2 \end{bmatrix} \left\{ \begin{bmatrix} p_{1t} \\ 1 - p_{1t} \end{bmatrix} \begin{bmatrix} P & 1-P \\ 1-Q & Q \end{bmatrix}^f \right\}' \end{aligned} \quad (6)$$

The prediction for conditional variance in each regime, h_{1t+f} and h_{2t+f} are $h_{it+f} = \hat{\omega}_i + \hat{\alpha}_i \varepsilon_{t+f-1}^2 + \hat{\beta}_i h_{t+f-1}$ for $f=1$ and $h_{it+f} = \hat{\omega}_i + (\hat{\alpha}_i + \hat{\beta}_i) h_{t+f-1}$ for $f \geq 2$, whereas h_{t+f} is given by :

$$h_{t+f} = E[z_{t+f}^2 | \zeta_{t-1}] - [E(z_{t+f} | \zeta_{t-1})]^2$$

with $E[z_{t+f} | \zeta_{t-1}]$ as (6) and

$$\begin{aligned} E[z_{t+m}^2 | \zeta_{t-1}] &= \begin{bmatrix} \mu_{1t+m}^2 + h_{1t+m} \\ \mu_{2t+m}^2 + h_{2t+m} \end{bmatrix}' \begin{bmatrix} p_{1t+m} \\ p_{2t+m} \end{bmatrix} \\ &= \begin{bmatrix} \hat{\delta}_1^2 + h_{1t+m} \\ \hat{\delta}_2^2 + h_{2t+m} \end{bmatrix}' \left\{ \begin{bmatrix} p_{1t} \\ 1 - p_{1t} \end{bmatrix} \begin{bmatrix} P & 1-P \\ 1-Q & Q \end{bmatrix}^f \right\}' \end{aligned}$$

4. APPLICATION TO DOWJONES INDEX RETURN

As an empirical implementation, the regime switching GARCH(1,1) is applied to capture the regime volatility switching of the daily closing price Dowjones stock market return index. Regime switching GARCH gives more capability if there's high difference between low and high volatility. The plot of Dowjones index transformed in return series is shown in figure 1.

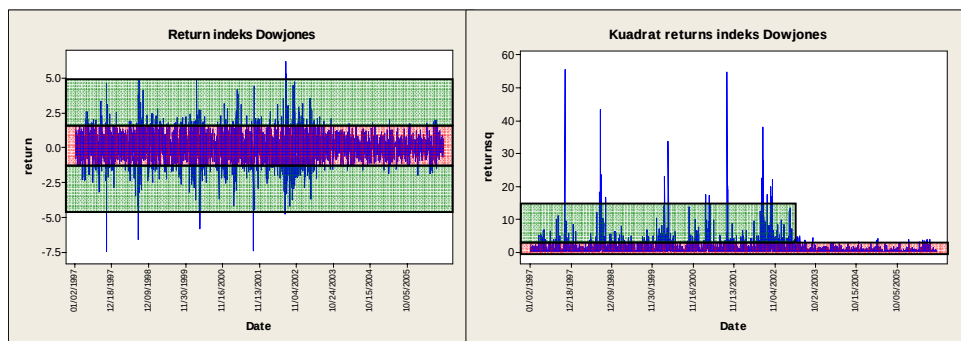


Figure 1. Time series plot of Dowjones index

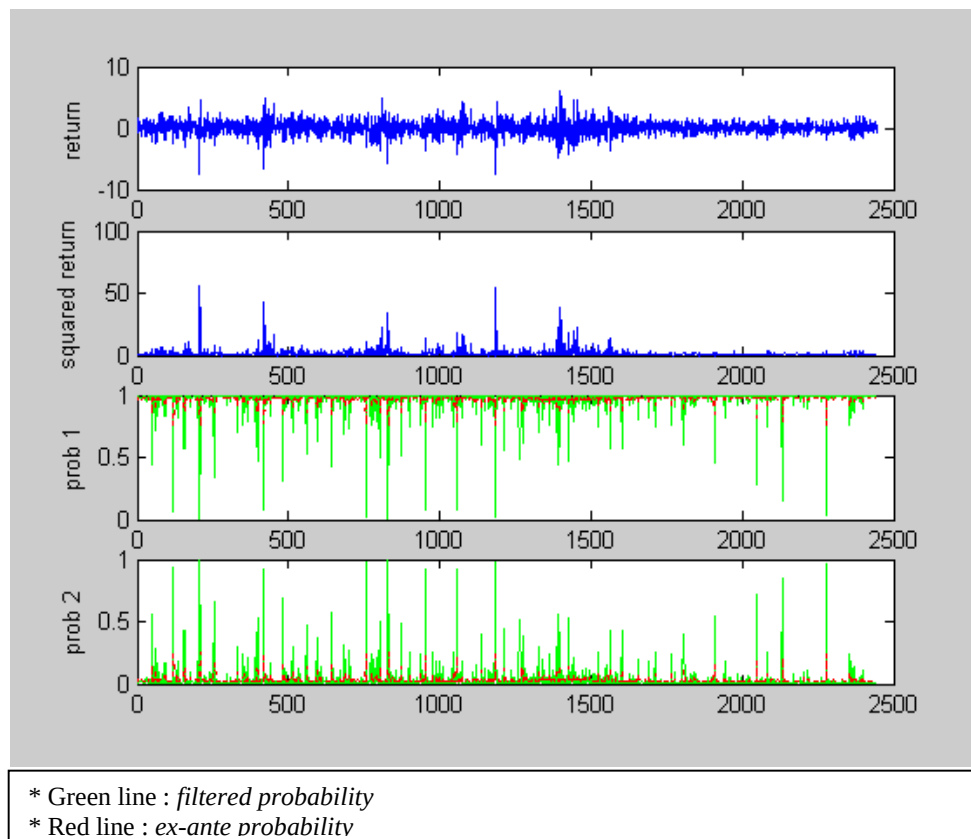


Figure 2. Time series plot of index and return series, ex-ante and filtered probability

Figure 1 shows simply that there is indication for volatility regimes occurred in the series. Low volatility gives little fluctuation (red highlighted) rather than the high one (green highlighted). The probability (ex-ante and filtered) of being in certain regime at time t given in figure 2, plotted together with return series. Using EM algorithm as in (5), parameter estimation result shown on table 1 as follows.

Table 1. Parameter Estimation

	first regime				Transition prob.
Parameter	δ_1	ω_1	α_1	β_1	P
Estimation	0.048036	0.016778	0.032885	0.893320	0.97579
Std.err	0.02998	0.0080294	0.011077	0.017997	0.0042901
t-stat	1.6023**	2.0895*	2.9686*	49.636*	227.45*
	second regime				Transition prob
Parameter	δ_2	ω_2	α_2	β_2	Q
Estimation	-1.76900	0.0000000082	0.156920	0.828470	0.24845
Std.err	0.084882	0.000000056	0.07764	0.086676	0.081103
t-stat	-20.841*	0.17855**	2.0211*	9.5582*	3.0634*

* Significance at $\alpha = 10\%$

** Not significance at $\alpha = 10\%$

Almost all of parameters are significance at $\alpha = 10\%$ except δ_1 and ω_2 , but the both insignificance parameters cannot be removed from the model because in another regime they are both significance. However, the markov chain transition matrix can be formed based on estimation result at table 1.

$$\mathbf{P} = \begin{bmatrix} P & 1-P \\ 1-Q & Q \end{bmatrix} = \begin{bmatrix} 0.976 & 0.024 \\ 0.752 & 0.248 \end{bmatrix}$$

The transition matrix informs that the transition process frequently occurred in first regime ($P = 0.976$) rather than the second regime ($Q = 0.248$), and if the second regime is occurred then probability to switch at the first regime is big enough. Figure 2 shows that first regime represents the high volatility and the second one captures high volatility, by considering the probability p_{1t} and φ_{1t} .

5. CONCLUDING REMARKS

Based on the analysis and empirical example in the previous part of the paper, the conclusion can be drawn as follows :

- a. Regime switching GARCH gives more informative model to describe switching behaviour between high and low volatilities rather than usual GARCH model.
- b. EM algorithm is used in estimation process since there's no information about which regime occurred at time t .
- c. The forecasting process is done by first computing markov chain probabilities at $t+f$.
- d. The Dowjones index return does follows the switching GARCH process based on empirical studies above.

Acknowledgements.

The first author gives big appreciate to Professor Nur Iriawan and Mr. Slamet Mulyono for all their support in finishing the paper and graduating from ITS.

REFERENCES

1. A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via EM algorithm", *Journal of the Royal Statistical Society Series B*, **39**, 1-38, 1977.
2. J.D. Hamilton, *Time Series Analysis*, Princeton University Press, New Jersey, 1994.
3. J.D. Hamilton, and R. Susmel, Autoregressive conditional heteroscedasticity and changes in regime, *Journal of Econometrics*, **64**, 307-333, 1994.
4. J. Marcucci, *Regime Switching GARCH in The Analysis and Forecasting of Stock Market Volatility and The Effects on Option Evaluation*, Technical Report, San't Anna School of Advanced Studies, Pisa, Italy, and University of California, San Diego, 2001.
5. J. Marcucci, Forecasting stock market volatility with regime switching GARCH models, *Studies in Nonlinear Dynamics & Econometrics*, **9**, 1-55, 2005.
6. R.F. Engle, Autoregressive conditional heteroscedasticity with estimates of U.K inflation, *Econometrica*, **50**, 987-1008, 1982.
7. S. Gray, Modeling the conditional distribution of interest rates as a regime switching process, *Journal of Financial Economics*, **42**, 27-62. 1996.
8. T. Bollerslev, Generalized autoregressive conditional heteroscedasticity, *Journal of Econometrics*, **31**, 307-327, 1996.

M. Arbi Hadiyat : Master Student at Department of Statistics - Institut Teknologi Sepuluh Nopember (ITS) Surabaya. Lecturer at Universitas Surabaya (Ubaya).
E-mail: arbi@ubaya.ac.id

Nur Iriawan : Lecturer at Department of Statistics Institut Teknologi Sepuluh Nopember (ITS) Surabaya

Slamet Mulyono : Lecturer at Department of Statistics Institut Teknologi Sepuluh Nopember (ITS) Surabaya

AN APPLICATION OF THE GSTAR-KRIGING MODEL

BUDI NURANI RUCHJANA AND SUTAWANIR DARWIS

Abstract. The Generalized Space Time Autoregressive-Kriging (GSTAR-Kriging) model as a combination of the GSTAR model and Kriging technique still has remain study both of theoretical and application. In the GSTAR model, we have varying autoregressive parameter

$\phi_{10}^{(i)}$ and space time parameter $\phi_{11}^{(i)}$, because we assume that the parameter depends on locations. But, the GSTAR model still restriction, because it is only used to predict observation at sample locations.

In this paper we propose and apply the GSTAR-Kriging model to real data. We use the least squares estimator of the GSTAR as an input in the Kriging technique to get the observation at the new location as unsample location. We apply the GSTAR-Kriging model to oil production data using the spatial weights. We also propose the GSTAR-Kriging software using visual basic excel.

Keywords and Phrases: GSTAR-Kriging, least squares, the spatial weights, unsample location

1. INTRODUCTION

In daily life the space time models play an important, especially for modeling the combination of time series and spatial data. Pfeifer (1979) propose the Space Time Autoregressive (STAR) model, but it is only applicable for the homogenous locations. Ruchjana (2002) extend the STAR model to be the Generalized Space Time Autoregressive (GSTAR) model with varying autoregressive parameter $\phi_{10}^{(i)}$ and space time parameter $\phi_{11}^{(i)}$.

Lack of the GSTAR model is only applicable to predict observation at sample locations. But in real phenomenon, usually we want to try to predict observation at unsample location. Kriging technique from Krige (1963) can be used to predict observation at unsample locations based on the information from the nearest neighbourhood of sample locations.

In this paper, we propose the Generalized Space Time AutoRegressive (GSTAR)-Kriging as an extension of the GSTAR model from Ruchjana (2002). We apply the GSTAR-Kriging model to real data and propose the GSTAR-Kriging software using visual basic excel to make computation more easier and faster.

2. MAIN RESULTS

THE GENERALIZED SPACE TIME AUTOREGRESSIVE MODEL

The GSTAR(p, λ_k) model as proposed by Ruchjana (2002) is written as:

$$\mathbf{z}(t) = \sum_{k=1}^p \sum_{\ell=0}^{\lambda_k} \Phi_{kl} \mathbf{W}^{(\ell)} \mathbf{z}(t-k) + \mathbf{e}(t) \quad (1)$$

where:

λ_k : spatial order of the k^{th} autoregressive term

Φ_{kl} : the diagonal matrices with the diagonal elements as the autoregressive and the space time parameters of time lag k and spatial lag l for each location $(\phi_{kl}^{(1)}, \dots, \phi_{kl}^{(N)})$.

$\mathbf{W}^{(l)}$: $(N \times N)$ the weight matrix for spatial lag $l = 0, 1, 2, \dots, \lambda_k$

$\mathbf{e}(t)$: the white noise with mean vector $\mathbf{0}$ and variance-covariance matrix $\sigma^2 \mathbf{I}_N$

$\mathbf{z}(t)$: $(N \times 1)$ of observation vector at time t

The GSTAR(1;1) model which both of parameters $\phi_{10}^{(i)}$ and $\phi_{11}^{(i)}$ are changing is presented:

$$\mathbf{z}(t) = \Phi_{10} \mathbf{z}(t-1) + \Phi_{11} \mathbf{W}^{(1)} \mathbf{z}(t-1) + \mathbf{e}(t) \quad (2)$$

where the Φ_{10} is the diagonal matrices with the diagonal elements as the autoregressive parameters of lag time 1 for each location $(\phi_{10}^{(1)}, \phi_{10}^{(2)}, \dots, \phi_{10}^{(N)})$. The Φ_{11} is the diagonal matrices with the diagonal elements as the space time parameters in the lag spatial 1 and lag time 1 for each location $(\phi_{11}^{(1)}, \phi_{11}^{(2)}, \dots, \phi_{11}^{(N)})$.

The GSTAR(1;1) can be written as the vector autoregressive order one model or VAR(1):

$$\mathbf{z}(t) = \Phi \mathbf{z}(t-1) + \mathbf{e}(t) \quad (3)$$

where $\Phi = [\Phi_{10} + \Phi_{11} \mathbf{W}^{(1)}]$.

GSTAR(1;1) in (2) model has a linear model in form:

$$\mathbf{z} = \mathbf{X} \vec{\phi} + \mathbf{e} \quad (4)$$

The least squares estimator can be written as:

$$\hat{\phi} = \begin{bmatrix} \hat{\phi}_{10}^{(i)} \\ \hat{\phi}_{11}^{(i)} \end{bmatrix} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{z} \quad (5)$$

2.2 KRIGING TECHNIQUE

Kriging is a method of calculating estimates of a regionalized variable at a point, over an area, or within a volume, and uses as a criterion the minimization of an estimation variance. Kriging interpolation involves the generation of images of the reservoir properties and commonly used to visualize reservoir heterogeneities. Therefore, kriging techniques not well suited for reproducing geological reservoir patterns where the number of data are very limited. Using kriging technique, we can predict the observation at unsample location (Armstrong (1998)).

A kriged estimator Z_V^* is a linear combination of n values of the regionalized variable:

$$Z_V^* = \sum_{i=1}^n \lambda_i Z_i \quad (6)$$

Weights λ_i are calculated according to these criteria:

1. Unbiased

$$E[Z_V^* - Z_V] = 0$$

$$E[Z(x)] = m = E[Z_V]$$

2. The estimation variance is minimized

$$Var[Z_V^* - Z_V]$$

Error variance $[Z_V^* - Z_V]$ can be expressed as part of semivariogram

$$= 2 \sum \lambda_i \bar{\gamma}(x_i, V) - \sum \sum \lambda_i \lambda_j \gamma(x_i, x_j) - \bar{\gamma}(V, V)$$

where $\bar{\gamma}(x_i, V)$ is mean of variogram between location x_i and unsampled location V .

The constraint of total λ_i is 1, is approached by Lagrange multiplier, so we have the kriging system equations:

$$\begin{aligned} \sum \lambda_j \gamma(x_i - x_j) + \mu &= \frac{1}{V} \int_V \gamma(x - x_i) dx \text{ or} \\ \sum_{j=1}^n \lambda_j \gamma(S_i - S_j) + \mu &= \gamma(s_i, V) \text{ and } \sum \lambda_i = 1 \end{aligned} \quad (7)$$

From these equations system we can have λ_j and μ which give the minimum kriging:

$$\sigma_k^2 = \sum \lambda_i \bar{\gamma}(x_i, V) - \bar{\gamma}(V, V) + \mu$$

For two locations, we have the minimum variance of kriging (Armstrong, 1998):

$$\begin{aligned} \phi &= 2(\lambda_1 \bar{\gamma}(x_1, V) + \lambda_2 \bar{\gamma}(x_2, V)) - (\lambda_1 \lambda_1 \gamma(x_1, x_1) + \lambda_1 \lambda_2 \gamma(x_1, x_2) \\ &\quad + \lambda_2 \lambda_2 \gamma(x_2, x_2)) - \bar{\gamma}(V, V) + 2\mu(1 - (\lambda_1 + \lambda_2)) \end{aligned} \quad (8)$$

Suppose $\gamma_{ij} = \gamma(x_i, x_j)$ and $\gamma_{iV} = \gamma(x_i, V)$ so, we have:

$$\phi = 2(\lambda_1 \bar{\gamma}_{1V} + \lambda_2 \bar{\gamma}_{2V}) - (\lambda_1 \lambda_1 \gamma_{11} + \lambda_1 \lambda_2 \gamma_{12} + \lambda_2 \lambda_1 \gamma_{21} + \lambda_2 \lambda_2 \gamma_{22})$$

$$-\gamma(V, V) + 2\mu(1 - \lambda_1 - \lambda_2) \quad (9)$$

or

$$\begin{aligned} \phi &= 2\lambda_1\bar{\gamma}_{1V} + 2\lambda_2\bar{\gamma}_{2V} - \lambda_1^2\gamma_{11} + 2\lambda_1\lambda_2\gamma_{12} + \lambda_2^2\gamma_{22} \\ &\quad - \gamma(V, V) + 2\mu(1 - \lambda_1 - \lambda_2) \end{aligned}$$

Using differentiation to λ_1 , λ_2 , μ we have:

$$\frac{\partial \phi}{\partial \lambda_1} = 2\gamma_{1V} - 2\lambda_1\gamma_{11} - 2\lambda_2\gamma_{21} - 2\mu = 0$$

$$\frac{\partial \phi}{\partial \lambda_2} = 2\gamma_{2V} - 2\lambda_1\gamma_{12} - 2\lambda_2\gamma_{22} - 2\mu = 0$$

$$\frac{\partial \phi}{\partial \mu} = 2 - 2\lambda_1 - 2\lambda_2 = 0$$

$$\text{so that } \gamma_{1V} = \lambda_1\gamma_{11} + \lambda_2\gamma_{12} + \mu \quad (10)$$

$$\gamma_{2V} = \lambda_1\gamma_{12} + \lambda_2\gamma_{22} + \mu$$

Furthermore, the differentiation ϕ to μ gives $\lambda_1 + \lambda_2 = 1$ or $\lambda_2 = 1 - \lambda_1$

Because the semivariogram at distance zero is zero, so that:

$$\gamma_{1V} = \lambda_2\gamma_{12} + \mu$$

$$\gamma_{2V} = \lambda_1\gamma_{12} + \mu$$

Elimination of μ gives: $\gamma_{1V} - \gamma_{2V} = \lambda_2\gamma_{12} - \lambda_1\gamma_{12}$.

Because of $\lambda_2 = 1 - \lambda_1$, so we can write : $\gamma_{1V} - \gamma_{2V} = (1 - \lambda_1)\gamma_{12} - \lambda_1\gamma_{12}$

$$\text{Finally we get : } \lambda_1 = \frac{1}{2} \left(\frac{\gamma_{2V} + \gamma_{1V}}{\gamma_{12}} + 1 \right) \text{ or } \lambda_1 = \frac{1}{2} + \left(\frac{\gamma_{2V} - \gamma_{1V}}{2\gamma_{12}} \right) \quad (11)$$

And for λ_2 we get :

$$\lambda_2 = 1 - \lambda_1$$

$$\lambda_2 = 1 - \left(\frac{1}{2} + \left(\frac{\gamma_{2V} - \gamma_{1V}}{2\gamma_{12}} \right) \right)$$

$$\lambda_2 = \frac{1}{2} + \left(\frac{\gamma_{1V} - \gamma_{2V}}{2\gamma_{12}} \right) \quad (12)$$

To get the value of λ_1 and λ_2 using ordinary kriging method we should have the values of γ_{1V} , γ_{2V} and γ_{12} . The value of γ_{12} is semivariogram experimental from two sample points and γ_{1V} is the semivariogram of the first sample point and the unsample point which will be predicted.

For case study we use the linear semivariogram for two locations

$$\gamma(h) = \begin{cases} \frac{\hat{\gamma}(r)}{r} h & , h \leq r \\ \hat{\gamma}(r) & , h > r \end{cases} \quad (13)$$

2.3 CASE STUDY

For case study, we choose two variables from oil reservoir volcanic in West

Java, i.e.: (1) the oil production, Z for the GSTAR model's observation at 3 wells (each 80 monthly observation) and (2) thickness data, Y for the spatial weight. We apply the spatial weights matrix in the GSTAR(1;1) models to oil production data in 3 wells. Each of them has a similar pattern as AR(1) model.

Table 1. Sub Surface Coordinates

Well	x(m)	y(m)
Well-1	14357.27	24.33
Well-2	14055.55	351.38
Well-3	14592.43	424.60

Using the non linear method, we fitted the spherical model for the experimental semivariogram of thickness, as presented:

$$\hat{\gamma}(h_{ij}) = \begin{cases} 6416.872 [1.5(h_{ij}/2100) - 0.5(h_{ij}/a)^3] & , h_{ij} \leq 2100\text{m} \\ 6416.872 & , h_{ij} > 2100\text{m} \end{cases} \quad (14)$$

The variance (sill) of thickness is 6416.872 m^2 and range is 2100 m. Furthermore, using isotropic semivariogram, we have the standardized spatially weight matrix $\mathbf{W}_{ss}$:

$$\mathbf{W}_{ss} = \begin{bmatrix} 0 & 0.605 & 0.395 \\ 0.657 & 0 & 0.343 \\ 0.556 & 0.444 & 0 \end{bmatrix}$$

Through the least squares method, parameters of GSTAR(1;1) models with the standardized spatial weight and Residual Sum of Squares (RSS) are shown in Table 3.

Table 2. The Least Squares Estimator of the GSTAR(1;1)

Weight	$\hat{\phi}_{10}^{(i)}$	$\phi_{11}^{(i)}$	RSS in each well	RSS Total
$\mathbf{W}_{ss}$ W68	0.9032	0.0583	0.8287×10^6	2.9132×10^6
W72	0.8787	0.1818	0.1531×10^6	
W120	0.8015	0.1298	0.5533×10^6	

Future prediction of oil production using GSTAR(1;1) are obtained using:

$$\begin{pmatrix} \hat{z}_1(t) \\ \hat{z}_2(t) \\ \hat{z}_3(t) \end{pmatrix} = \begin{pmatrix} 0.9032 & 0 & 0 \\ 0 & 0.8787 & 0 \\ 0 & 0 & 0.8015 \end{pmatrix} \begin{pmatrix} z_1(t-1) \\ z_2(t-1) \\ z_3(t-1) \end{pmatrix} + \begin{pmatrix} 0.0583 & 0 & 0 \\ 0 & 0.1818 & 0 \\ 0 & 0 & 0.1298 \end{pmatrix} \begin{pmatrix} 0 & 0.605 & 0.395 \\ 0.657 & 0 & 0.343 \\ 0.556 & 0.444 & 0 \end{pmatrix} \begin{pmatrix} z_1(t-1) \\ z_2(t-1) \\ z_3(t-1) \end{pmatrix} \quad (15)$$

In the context of this paper, we want to combine the GSTAR and the kriging technique, is called the GSTAR-kriging. The objective is to reproduce the pattern of spatial correlation of the model parameters based on experience production data modeled as GSTAR models. In this case, we use the parameters GSTAR(1;1) as an input of kriging technique.

The application proposes a combination of vector time series process and spatial simulation to forecast the GSTAR parameters of new well locations. It is supposed that the experience wells locations are fixed, while the new well varied in the xy-plane. Let $s = (x, y)$ denotes the location of new well. GSTAR and kriging are used to describe production profile of a new well. Table 4 shows the parameters of GSTAR from three production wells.

Table 3. Input for kriging, the GSTAR parameters $\phi_{10}(s), \phi_{11}(s)$ from three well locations

Well s_i	Coordinate (x_i, y_i)	$\phi_{10}(s_i)$	$\phi_{11}(s_i)$
1	(14357.27, 24.33)	.9032	.0583
2	(14055.55, 351.38)	.8787	.1818
3	(14592.43, 424.60)	.8015	.1298

Using the parameter $\phi_{10}(s), \phi_{11}(s)$ as an input in the kriging technique, we will have the parameters of the GSTAR(1;1) at unsample location. For case study, we choose the ten new wells between two wells Well-1 and Well-2. We have the $\phi_{11}^{(s_i)}$ parameter for ten unsample locations such as Table 4.

Table 4. Parameter $\phi_{11}^{(s_i)}$ for 10 Wells

Well	x(m)	y(m)	$\phi_{11}^{(s_i)}$
W1	14050.00	350.00	0.18
W2	14075.00	340.00	0.18
W3	14100.00	300.00	0.16
W4	14075.00	290.00	0.16
W5	14100.00	275.00	0.16
W6	14125.00	280.00	0.15
W7	14200.00	200.00	0.12
W8	14300.00	100.00	0.08
W9	14325.00	40.00	0.07
W10	14350.00	30.00	0.06

The coordinate's position of ten unsample locations between two wells are shown at Figure 2.1.

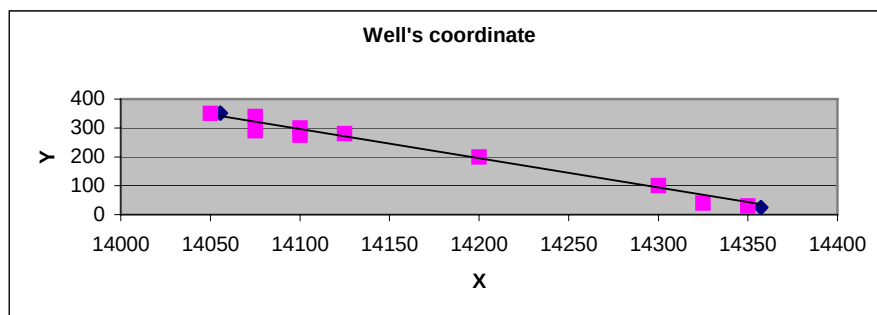


Figure 2.1 coordinate's position of ten unsample locations

Using the $\phi_{11}^{(s_i)}$ parameters, we can simulate the GSTAR(1;1) for ten unsample locations, and we have the model for prediction of oil production for each new well. We still study more in this case, we hope this method will give the recommendation for oil field, especially in oil well placement using the space time model.

3. CONCLUDING REMARKS

This paper discusses performance of GSTAR-Kriging, as an extension of the GSTAR model, and the idea of its application in the determination of a new oil well placement. A GSTAR model was proposed as a model for oil production time series data. Lack of the GSTAR model is only applicable for oil prediction of sample locations. The combination of the GSTAR model and kriging technique seems a suitable representation for modeling oil production to predict oil production at unsample locations. So, the production profile of a new well can be described by the GSTAR-Kriging model.

Acknowledgement

The authors thanks to the General Manager of Pertamina DOH JBB Jatibarang, for providing data for case study. Also the author thanks to Prof. Dr. Phillip E. Pfeifer, Dr. H. P. Lopuhaa and Dr. J. Bruining for a nice discussion in the space time models.

REFERENCES

1. Armstrong, M., *Basic Linear Geostatistics*, Germany: Springer-Verlag Berlin Heidelberg, 1998.
2. Box, G.E.P and Gwilym M. Jenkins, *Time Series Analysis, Forecasting and Control*, San Fransisco : Holden-Day, Inc., 1976.
3. Pfeifer, P.E. *Spatial Dynamic Modeling*. unpublished Ph.D Dissertation. Georgia Institute of Technology, Georgia. 1979.
4. Ruchjana, B. N. *A Generalized Space Time Autoregressive Order One Model and Its Application to Oil Production*. unpublished Ph.D Dissertation, Department of Mathematics, ITB. Bandung. 2002.
5. Stoffer, D. Estimation and Identification of Space-Time ARMAX Models in the Presence of Missing Data. *Journal of the American Statistical Association*, **81**, 395, 762-772. 1986.

Budi Nurani Ruchjana: Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Bandung, Indonesia.
E-mail: bnurani@yahoo.com

Sutawanir Darwis: Faculty of Mathematics and Natural Sciences, Institut Teknologi Bandung, Bandung, Indonesia
E-mail: sdarwis@math.itb.ac.id

STATISTICAL SIGNIFICANCE TEST FOR NEURAL NETWORK CLASSIFICATION

SRI REZEKI, SUBANAR, SURYO GURITNO

Abstract. Model selection in neural networks can be guided by statistical procedures, such as hypothesis tests, information criteria and cross validation. Taking a statistical perspective is especially important for nonparametric models like neural networks, because the reason for applying them is the lack of knowledge about an adequate functional form. Many researchers have developed model selection strategies for neural networks which are based on statistical concepts. In this paper, we focused on the model evaluation by implementing statistical significance test. We used Wald-test to evaluate the relevance of parameters in the networks for classification problem. Parameters with no significance influence on any of the network outputs have to be removed. In general, the results show that Wald-test work properly to determine significance of each weight from the selected model. An empirical study by using Iris data yields all parameters in the network are significance, except bias at the first output neuron.

Key words: parameter significance, Wald-test, neural networks, classification.

1. INTRODUCTION

One of the most unresolved questions in the literature on neural networks (NN) is what architecture should be used for a given problem. Architecture selection requires choosing both the appropriate number of hidden units and the connection thereof [10]. A desirable network architecture contains as few hidden units and connections as necessary for a good approximation of the true function, taking into account the trade-off between estimation bias and variability due to estimation errors. It is therefore necessary to develop a methodology to select an appropriate network model for a given problem.

Reed [6] provides a survey about the usual approaches pursued in the network literature. The approaches, for example, are regularization, pruning, and stopped training. In regularization methods, the network weights are chosen such that they minimize the network error function (e.g. sum of squared errors) plus a penalty term for the networks complexity. Another way to justify the regularization term is to formalize and interpret the method in a Bayesian framework. This was reviewed for example in Bishop [3] and Ripley [9].

Pruning methods identify those parameters that do not ‘significantly’ contribute to the overall network performance. However, this ‘significance’ is usually not judged on the basis of test statistics (an exception is Burgess [4], who uses F-tests to identify insignificant parts of a network model). Pruning methods use the so-called ‘saliency’ as a measure of a weight’s importance. The saliency of a weight is defined as the increase in network model error (e.g. sum

of squared errors) incurred by setting this weight to zero. The idea is to remove the weights with low saliency; however, the method does not provide any guidelines as to whether or not a saliency should be judged as low. It is shown by Anders [1] that the computation of the saliency is generalized by a corresponding Wald-test statistics. An alternative to conventional pruning methods is developed and tested in Kingdon [5]. The basic idea of the approach, called ‘network regression pruning’, is to remove network weights while retaining the network’s mapping. A weight is seen as redundant if, after its removal, the original mapping can be (approximately) recovered by adjusting the remaining weights of the effected node.

In the application of stopped training, the data set is split into a training and a validation set. If the model errors in the validation set begin to grow during the training process, the training algorithm is stopped. In statistical terms, the method tries to make up the model being over parameterized by stopping the estimation algorithm before the minimum of the network error function is reached.

Anders and Korn [2] proposed some model selection strategies based on different statistical procedures. Those are the strategies which combine the LM and Wald-tests, strategies based on information criteria, and cross validation. As the results of the study, they suggest that a statistical analysis should become an integral part of neural network modeling.

In this paper, we use the results of our previous researches about model selection based on resampling procedures [7, 8]. The selected model is the lowest complex model from the subset models that are not significantly different with minimum error mean. Furthermore, the model is evaluated by using Wald-test for testing significance of each weights of the selected model.

2. HYPOTHESIS TESTING IN NEURAL NETWORK

Neural Networks build a class of very flexible models that can be used for a variety of different application, for example nonlinear regression or discriminant analysis. The term neural networks is not uniquely defined. Instead, it comprises many different network types and models. In this article, we will deal with so called feed forward neural networks (FFNN), an example of which is shown in Figure 1.

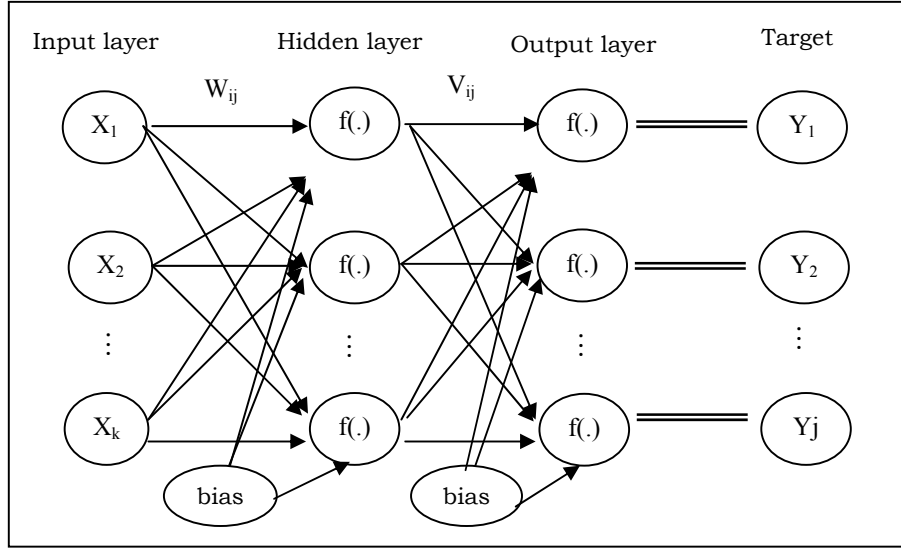


Figure 1. FFNN with single hidden layer

Since multilayer perceptron neural networks are non-linear regression models, the standard procedure for testing parameter significance, like wald test, apply in principle. However, to perform this test the (asymptotic) distribution of the network parameters is needed. This distribution was derived by White [13]. If there exists a unique set of optimum parameters, w^* , that leads to the best approximation of the true function by a certain network model, White shows that w^* can be consistently estimated through a set of parameters, $\hat{w}$, obtained by quasi maximum likelihood methods under quite general conditions. Moreover, $\hat{w}$ is asymptotically normal with mean w^* and covariance matrix $(1/n)C^*$, or:

$$\sqrt{n}(\hat{w} - w^*) \sim N(0, C^*) \quad (1)$$

where n is the number of observations. Since neural network only approximations to the true underlying functions, the derivation of the covariance matrices relies on the theory of misspecified models developed by White [11, 12, 14]. If a neural network is estimated via maximum likelihood method with log likelihood contribution $\ln L_t$, the covariance matrix of the parameters is given by

$$(1/n)C^* = (1/n)A^{*-1}B^*A^{*-1}. \quad (2)$$

The matrices A^* and B^* are defined as

$$A^* \equiv -E[\nabla^2 \ln L_t] \text{ and } B^* \equiv E[\nabla \ln L_t \nabla \ln L_t]$$

where ∇ denotes the gradient and ∇^2 is the Hessian. Note that the covariance matrix, C^* , accounts for misspecification in mean as well as in variance, i.e. a consistent estimate of C^* converges to the covariance matrix of the limiting distribution of $\hat{w}$. This holds, for example, if the network does not encompass the true underlying function and a possibly heteroscedastic error term has not been taken account of in the specification of the likelihood function. As can be

seen, the covariance matrix, $\mathbf{C}^*$, is a generalization of the covariance formula in standard maximum likelihood theory. If the model is correctly classified, the asymptotic covariance matrix of the limiting distribution can be estimated by both the inverse of the matrix $\mathbf{A}^*$ and the inverse of the matrix $\mathbf{B}^*$. In this case, $\mathbf{A}^*$ asymptotically equals $\mathbf{B}^*$ and $\mathbf{B}^* \mathbf{A}^{*-1}$ converges to the identity matrix $\mathbf{I}$.

The following corollary, proposition and theorems are needed to derive and proof the Wald test statistics for parameters of neural networks model.

Theorem 2.1. (White [13]). Let $(\Omega, \mathcal{F}, P)$ be a complete probability space on which is defined the sequence of independent identically distributed (i.i.d.) random variables

$$\{Z_t\} = (Z_t : \Omega \rightarrow \mathbb{R}^\nu, t = 1, 2, \dots), \quad \nu \in \mathbb{N} \equiv \{1, 2, \dots\}$$

Let $l : \mathbb{R}^\nu \times W \rightarrow \mathbb{R}$ be a function such that for each w in W , a compact subset of $\mathbb{R}^s$, $s \in \mathbb{N}$, $l(\cdot, w)$ is measurable- $\mathcal{B}^\nu$ (where $\mathcal{B}^\nu$ is the Borel σ -field generated by the open sets of $\mathbb{R}^\nu$), and for each z in $\mathbb{R}^\nu$, $l(z, \cdot)$ is continuous on W . Suppose further that there exists $d : \mathbb{R}^\nu \rightarrow \mathbb{R}^+$ such that for all w in W $|l(z, w)| \leq d(z)$ and $E(d(Z_t)) < \infty$ (i.e., l is dominated on W by an integrable function).

Then for each $n = 1, 2, \dots$ there exists a solution $\hat{w}_n$ to the problem

$$\min \hat{\lambda}_n(w) = n^{-1} \sum_{t=1}^n l(Z_t, w) \text{ and } \hat{w}_n \xrightarrow{a.s.} w^*,$$

where $w^* \equiv \{w^* \in W : \lambda(w^*) \leq \lambda(w)\}$, $\lambda(w) \equiv E(l(Z_t, w))$

Theorem 2.2. (White [13]). Let $(\Omega, \mathcal{F}, P)$, $\{Z_t\}$, W and l be as in Theorem 2.1, and suppose that $\hat{w}_n \xrightarrow{a.s.} w^*$ where w^* is an isolated element of W^* interior to W .

Suppose in addition that for each z in $\mathbb{R}^\nu$, $l(z, \cdot)$ is continuously differentiable of order 2 on W ; that $E(\nabla l(Z_t, w^*)' \nabla l(Z_t, w^*)) < \infty$; that each element of $\nabla^2 l$ is dominated on W by an integrable function; and that $\mathbf{A}^* \equiv E(\nabla^2 l(Z_t, w^*))$ and $\mathbf{B}^* \equiv E(\nabla l(Z_t, w^*)' \nabla l(Z_t, w^*))$ are nonsingular $(s \times s)$ matrices, where ∇ and ∇^2 denote the $(s \times 1)$ gradient and $(s \times s)$ Hessian operators with respect to w .

Then $\sqrt{n}(\hat{w}_n - w^*) \xrightarrow{d} N(0, \mathbf{C}^*)$, where $\mathbf{C}^* = \mathbf{A}^{*-1} \mathbf{B}^* \mathbf{A}^{*-1}$. If in addition each element of $\nabla l \nabla l'$ is dominated on W by an integrable function, then $\hat{\mathbf{C}}_n \xrightarrow{a.s.} \mathbf{C}^*$, where $\hat{\mathbf{C}}_n = \hat{\mathbf{A}}_n^{-1} \hat{\mathbf{B}}_n \hat{\mathbf{A}}_n^{-1}$, and

$$\hat{\mathbf{A}}_n = \frac{\sum_{t=1}^n \nabla^2 l(Z_t, \hat{w}_n)}{n}, \quad \hat{\mathbf{B}}_n = \frac{\sum_{t=1}^n \nabla l(Z_t, \hat{w}_n) \nabla l(Z_t, \hat{w}_n)'}{n}.$$

Corollary 2.3. (White [15]) Let $\{\mathbf{b}_n\}$ be a sequence of random $k \times 1$ vectors such that $\mathbf{V}_n^{-1/2} \mathbf{b}_n \xrightarrow{d} N(0, \mathbf{I})$, where $\{\mathbf{V}_n\}$ and $\{\mathbf{V}_n^{-1}\}$ are $O(1)$. Let $\{\mathbf{A}_n\}$ be a $O(1)$ sequence of (nonstochastic) $q \times k$ matrices with full row rank q for all n sufficiently large, uniformly in n . Then the sequence $\{\mathbf{A}_n \mathbf{b}_n\}$ is such that

$$\mathbf{\Gamma}_n^{-1/2} \mathbf{A}_n \mathbf{b}_n \xrightarrow{d} N(0, \mathbf{I}),$$

where $\mathbf{\Gamma}_n \equiv \mathbf{A}_n \mathbf{V}_n \mathbf{A}_n'$ and $\mathbf{\Gamma}_n$ and $\mathbf{\Gamma}_n^{-1}$ are $O(1)$.

Proposition 2.4. (White [15]) Let $g: \mathfrak{R}^k \rightarrow \mathfrak{R}^l$ be continuous on a compact set $C \subset \mathfrak{R}^k$. Suppose that $\{\mathbf{b}_n\}$ is a sequence of random vector $k \times 1$ and $\{\mathbf{c}_n\}$ is a sequence of $k \times 1$ vector such that $\mathbf{b}_n - \mathbf{c}_n \xrightarrow{p} 0$, and for all n sufficiently large, $\mathbf{c}_n$ is interior to C , uniformly in n . Then $g(\mathbf{b}_n) - g(\mathbf{c}_n) \xrightarrow{p} 0$.

Theorem 2.5. (White [15]) Let $\mathbf{V}_n^{-1/2} \mathbf{b}_n \xrightarrow{d} N(0, \mathbf{I}_k)$, and suppose there exists $\hat{\mathbf{V}}_n$ positive semidefinite and symmetric such that $\hat{\mathbf{V}}_n - \mathbf{V}_n \xrightarrow{p} 0$, where $\mathbf{V}_n$ is $O(1)$, and for all n sufficiently large, $\det(\mathbf{V}_n) > \delta > 0$. Then $\mathbf{b}_n' \hat{\mathbf{V}}_n^{-1} \mathbf{b}_n \xrightarrow{d} \chi_k^2$.

The Wald statistic allows the simplest analysis, although it may or may not be the easiest statistic to compute in a given situation. The motivation for the Wald statistic is that when the null hypothesis is correct $\mathbf{S}\hat{\mathbf{w}}_n$ should be close to $\mathbf{S}\mathbf{w}^* = \mathbf{s}$, so a value of $\mathbf{S}\hat{\mathbf{w}}_n - \mathbf{s}$ far from zero is evidence against the null hypothesis.

The theorem about Wald statistic that be used for hypothesis testing of parameters in NN model is adapted from White [15] Theorem 4.31. The result of adaptation for our specific purpose is as follow:

Theorem 2.6. Let the conditions of Theorem 2.2 hold, i.e.:

- (i) $\mathbf{C}^{*-1/2} \sqrt{n}(\hat{\mathbf{w}}_n - \mathbf{w}^*) \xrightarrow{d} N(\mathbf{0}, \mathbf{I})$, where $\mathbf{C}^* \equiv \mathbf{A}^{*-1} \mathbf{B}^* \mathbf{A}^{*-1'}$, and $\mathbf{C}^{*-1}$ is $O(1)$,
- (ii) there exists a matrix $\hat{\mathbf{B}}_n$ positive semidefinite and symmetric such that $\hat{\mathbf{B}}_n - \mathbf{B}^* \xrightarrow{p} 0$. Then $\hat{\mathbf{C}}_n - \mathbf{C}^* \xrightarrow{p} 0$, where $\hat{\mathbf{C}}_n = \hat{\mathbf{A}}_n^{-1} \hat{\mathbf{B}}_n \hat{\mathbf{A}}_n^{-1'}$, and

$$\hat{\mathbf{A}}_n = \frac{\sum_{t=1}^n \nabla^2 l(Z_t, \hat{\mathbf{w}}_n)}{n}, \quad \hat{\mathbf{B}}_n = \frac{\sum_{t=1}^n \nabla l(Z_t, \hat{\mathbf{w}}_n) \nabla l(Z_t, \hat{\mathbf{w}}_n)'}{n}.$$

And, let $\text{rank}(\mathbf{S}) = q \leq k$. Then under $H_0: \mathbf{S}\mathbf{w}^* = \mathbf{s}$,

- (i) $\mathbf{\Gamma}_n^{-1/2} \sqrt{n}(\mathbf{S}\hat{\mathbf{w}}_n - \mathbf{s}) \xrightarrow{d} N(\mathbf{0}, \mathbf{I})$, where

$$\mathbf{\Gamma}_n \equiv \mathbf{S}\mathbf{C}^* \mathbf{S}' = \mathbf{S}\mathbf{A}^{*-1} \mathbf{B}^* \mathbf{A}^{*-1'} \mathbf{S}'.$$

- (ii) The Wald statistic

$$W_n \equiv n(\mathbf{S}\hat{\mathbf{w}}_n - \mathbf{s})' \hat{\mathbf{\Gamma}}_n^{-1} (\mathbf{S}\hat{\mathbf{w}}_n - \mathbf{s}) \xrightarrow{d} \chi_q^2, \quad (3)$$

where $\hat{\mathbf{\Gamma}}_n \equiv \mathbf{S}\hat{\mathbf{C}}_n \mathbf{S}'$.

Proof: We use Corollary 2.3, Proposition 2.4 and Theorem 2.5 to prove Theorem 2.6 and the results are as follow.

- (i) Under H_0 , $\mathbf{S}\hat{\mathbf{w}}_n - \mathbf{s} = \mathbf{S}(\hat{\mathbf{w}}_n - \mathbf{w}^*)$, so

$$\mathbf{\Gamma}_n^{-1/2} \sqrt{n}(\mathbf{S}\hat{\mathbf{w}}_n - \mathbf{s}) = \mathbf{\Gamma}_n^{-1/2} \mathbf{S} \mathbf{C}^{*1/2} \mathbf{C}^{*-1/2} \sqrt{n}(\hat{\mathbf{w}}_n - \mathbf{w}^*).$$

It follows from Corollary 2.3 that $\mathbf{A}_n = \mathbf{S}$ and $\mathbf{b}_n = \sqrt{n}(\hat{\mathbf{w}}_n - \mathbf{w}^*)$, so that $\Gamma_n^{-1/2} \sqrt{n}(\mathbf{S}\hat{\mathbf{w}}_n - \mathbf{s}) \xrightarrow{d} N(\mathbf{0}, \mathbf{I})$.

- (ii) Based on Theorem 2.1 we have that $\hat{\mathbf{C}}_n - \mathbf{C}^* \xrightarrow{a.s.} 0$, so it imply that $\hat{\mathbf{C}}_n - \mathbf{C}^* \xrightarrow{p} 0$. By using Proposition 2.4, where $\hat{\Gamma}_n = g(\hat{\mathbf{C}}_n)$ and $\Gamma_n = g(\mathbf{C}^*)$, so that $\hat{\Gamma}_n - \Gamma_n \xrightarrow{p} 0$. Given the result in (i), i.e. $\Gamma_n^{-1/2} \sqrt{n}(\mathbf{S}\hat{\mathbf{w}}_n - \mathbf{s}) \xrightarrow{d} N(\mathbf{0}, \mathbf{I})$, so by implementing Theorem 2.5, this yields

$$W_n \equiv n(\mathbf{S}\hat{\mathbf{w}}_n - \mathbf{s})' \hat{\Gamma}_n^{-1} (\mathbf{S}\hat{\mathbf{w}}_n - \mathbf{s}) \xrightarrow{d} \chi_q^2. \quad \blacksquare$$

Thus, a test about the relevance (significance) of input where the hypothesis can be stated as

$$H_0 : \mathbf{S}\mathbf{w}^* = 0 \text{ against } H_1 : \mathbf{S}\mathbf{w}^* \neq 0$$

can be evaluated by applying Theorem 2.6. As an example, Wald statistic to evaluate this hypothesis testing is

$$\hat{W}_n = n\hat{\mathbf{w}}_n' \mathbf{S}' (\mathbf{S}\mathbf{C}^* \mathbf{S}')^{-1} \mathbf{S}\hat{\mathbf{w}}_n,$$

where $\mathbf{C}^*$ as defined in earlier section.

3. RESEARCH METHODOLOGY

We employ the result of our previous researchs about model selection for neural network classification [7, 8]. The data used in this study were Fisher's iris data. The strategy of model selection is as follow:

1. Take the whole data set and created $b = 1000$ training sets by a resampling technique taking the rest of the instances for testing. Both training and testing sets are the same size.
2. Train the model (FFNN) from the lowest to the highest degree of complexity (the degree of complexity ranging from 1 to m) by using the same sample, then generate m misclassification errors measure for training and testing sets.
3. Repeate step 2 for another samples which have been created so that there are 1000 misclassification errors for each model.
4. Calculate the difference of misclassification error between two models, then determine the difference mean for each two models.
5. Apply multiple matched pair comparison and select the least complex model from the subset models that are not significantly different from the model with minimum error mean.

The testing set is a set of observations, not previously shown to the neural network, and only used to asses the performance (generalization) of a fully-specified classifiers. Therefore we focus on misclassification error from testing set. Afterwards, we applied Wald test to determine significance of each weights

from the selected model.

4. EMPIRICAL RESULTS

The architecture of neural network model obtained is NN (4-1-3). It means 4 neurons in input layer, 1 neuron in hidden layer and 3 neurons in output layer. The result can be illustrated completely with the weights, that is:

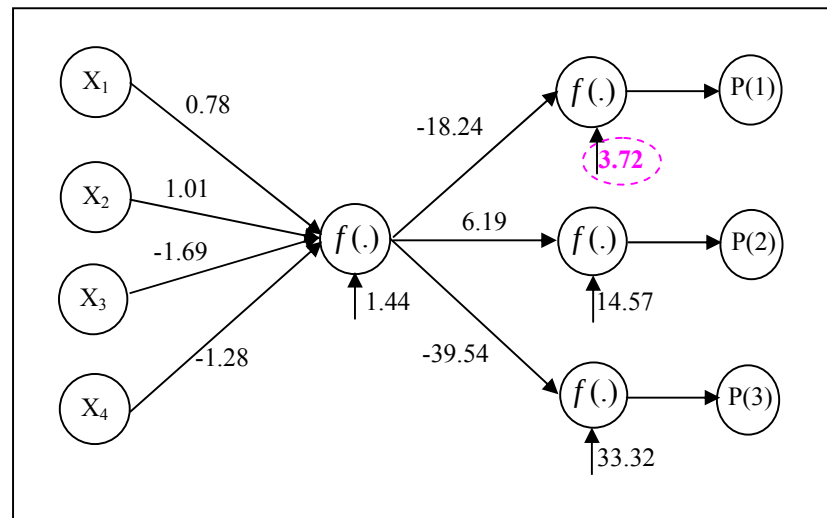


Figure 1. Architecture of NN (4-1-3) with the weights

The results of analysis can be summarized as follow:

Table 1. Weights Estimation of NN (4-1-3)

Weight	Beta	Var Beta	z	p-value
b→h1	1.44	0.11220	4.29889	0.00002
i1→h1	0.78	0.00635	9.78603	0.00000
i2→h1	1.01	0.00701	12.06526	0.00000
i3→h1	-1.69	0.01466	-13.95797	0.00000
i4→h1	-1.28	0.01733	-9.72337	0.00000
b→o1	3.72	6.32407	1.47926	0.13907
h1→o1	-18.24	7.62251	-6.60657	0.00000
b→o2	14.57	7.51995	5.31315	0.00000
h1→o2	6.19	7.74649	2.22402	0.02615
b→o3	33.32	6.76628	12.80943	0.00000
h1→o3	-39.54	9.59952	-12.76180	0.00000

Usually, neurons in NN are always full connectivity. By using this Wald test for NN classification, however, it is possible that there are weights which are not significance. Based on Table 1, all of the weights are significance except bias at the first output neuron. Therefore, the connection should be removed and the best model is not fully connected.

5. CONCLUSION

In general, neural networks are applied to problems where slight is known about the correct functional form. Therefore, a statistical approach to the model selection seems particularly important and should become an integral part of neural networks modeling. Wald-test is work properly in theoretically because $\hat{w}$ is asymptotically normal with mean w^* and covariance matrix $(1/n)C^*$, or $\sqrt{n}(\hat{w} - w^*) \sim N(0, C^*)$. In empirically, Wald-test is also applicable for determining parameter significance of the selected neural networks model. Although the empirical study presented is encouraging, this is only a first step. More experience has to be gained through new algorithm to construct the neural networks models that are not fully connected. For example, there are neurons in input layer which are not linked to the neuron in hidden layers and so on, or there are neurons in hidden and output layer without bias.

REFERENCES

1. U. Anders, Neural network pruning and statistical hypotheses tests, In N. Kasabov, R. Kozma, K. Ko, R. O'Shea, G. Coghill, & T Gedeon (Eds), *Progress in Connectionist-Based Information Systems* (Addendum pp. 1-4), Proceeding of ICONIP '97, Berlin, Springer, 1997.
2. U. Anders and O. Korn, Model selection in neural network. *Neural Networks*, **12**, 309-323, 1996.
3. C.M. Bishop, *Neural Network for Pattern Recognition*, Oxford: Clarendon Press, 1995.
4. A.N. Burgess, Nonlinear model identification and statistical significance test and their application to financial modeling, In *IEE Proceeding of the Fourth International Conference on Artificial Neural Networks* 409, pp. 312-317, 1995
5. J. Kingdon, *Intellegent systems and financial forecasting*, Berlin, Springer, 1997
6. R. Reed, Pruning algorithms - A survey, *IEEE Transactions on Neural Networks*, **4**, 740-747, 1993.
7. S. Rezeki, B. Sutijo, Subanar and S. Guritno, Statistical Model Selection Based on Resampling Procedure for Neural Networks Classification, to appear in *Proceeding of The First International Conference on Mathematics and Statistics*, (ICOMS), Bandung Islamic University, Bandung, 2006.
8. S. Rezeki, Subanar and S. Guritno, Model Selection For Neural Networks Classification By Using Matched Pair Comparison Test, *Proceeding of The 2nd Indonesia Malaysia Thailand Growth Triangle (IMT-GT) Regional Conference*, Penang, 2006
9. B.D. Ripley, *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge, 1996.
10. W.S. Sarle, (), Neural Network and Statistical Model, *Proceeding of Nineteenth Annual SAS Users Group International Conference*, pp. 1538-1550. Cary: SAS Institute, 1994
11. H. White, Consequences and detection of misspecified nonlinear regression models, *Journal of the American Statistical Association*, **76** (374), 419-433, 1981.
12. H. White, Maximum likelihood estimation of misspecified models, *Econometrica*, **50** (1), 1-25, 1982.
13. H. White, Learning in neural networks: a statistical perspective, *Neural Computation*, **1**, 425-464, 1989.
14. H. White, Estimation, inference, and specification analysis, Cambridge: Cambridge University Press, 1994.
15. H. White, *Asymptotic Theory for Econometricians*, Academic Press Inc., New York, 1999.

SRI REZEKI, SUBANAR, SURYO GURITNO

Sri Rezeki: Department of Mathematics Education, Universitas Islam Riau, Pekanbaru, Indonesia PhD Student, Department of Mathematics, Universitas Gadjah Mada, Indonesia.

E-mail: sri_rezeki_uir@yahoo.com

Subanar : Department of Mathematics, Universitas Gadjah Mada, Yogyakarta, Indonesia

E-mail: subanar@yahoo.com

Suryo Guritno: Department of Mathematics, Universitas Gadjah Mada, Yogyakarta, Indonesia

E-mail: guritno0@mailcity.ac.id

THE TECHNIQUE OF DATA ROBUSTNESS IN LOGIT MODELING

MOHD SAIFULLAH BIN RUSIMAN

Abstract. The new technique in data robustness, the least quartile difference (LQD) regression modeling and the widely used model, logit modeling have been studied in analyzing the nonlinear modeling. Research of these nonlinear models covered its discoveries, functions of models, assumption of residuals, probability distributions and the estimation techniques. A case study using the two methods of modeling as mentioned above was carried out. The comparison between the least quartile difference (LQD) regression modeling and logit modeling were done to find the better model by using the percentage of accuracy. After comparing the two models, it was found that the least quartile difference (LQD) regression modeling appeared to be the better model, having the highest probability of accuracy and can be proposed as the new effective model.

Key words and Phrases : data robustness, least quartile difference (LQD) regression modeling, logit modeling

1. INTRODUCTION

Regression analysis has become one of standard tools in analysis data. Its popularity comes from several sources. The mathematical equation gained from its analysis could explain relationship between the dependent and independent variables. It provides much explanatory power, especially due to its multivariate nature. It is widely available in computer packages, easy to interpret, and there is a widespread belief that it remains a reasonable procedure even if some of the assumptions underlying it are not met in the data. It has been widely used in applied sciences, economic, engineering, computer, social sciences and other fields (Agresti, A., 1996).

While estimates derived from regression analysis may be robust against errors in some assumptions, other assumptions are crucial, and their failure will lead to unreasonable estimates. Such is the case when the dependent variable is a qualitative measure rather than a continuous. The functional form obtained should be approximately near to the real data. If the functional form is far away from the real data, its mean the estimation will be inconsistency, bias and so on. However, other difficulties may arise with a non-linearity approach (Rousseeuw, et. al, 2003). In our daily lives, many data are observed quantitative and qualitatively. In this study, we are concern in qualitative data, mainly in dichotomous decisions. The dependent random variable is assumed to be binary, 0 and 1.

2. LITERATURE REVIEW

Intensive care for the critically ill patients is an essential component of acute hospital care. Although intensive care unit patients account for only 5% of inpatients, it consumes a significant amount of health care resources. In the United States, it accounts for 1% of the GNP and 15-20% of total hospital cost. In United Kingdom, the NHS spent £700 each year on critical care services. The first ICU in Malaysia was established in 1968. Since then, intensive care has developed rapidly and it is now available in all tertiary care hospitals and selected secondary care hospitals in the Ministry of Health.

Thus in 1994, the Intensive Care National Audit & Research Centre (ICNARC) was established in UK. It was co-funded by the Intensive Care Society and the Department of Health. Its main task was to conduct a review on intensive care practice using a Case Mix Programme (CMP) and to make recommendations to the relevant health authorities. The findings of ICNARC and that of the National Expert Group for the Department of Health prompted the British government to release £142.5 million in year 2000 to further improve intensive care unit throughout the country. The National Audit on Adult Intensive Care Units in Malaysia is modeled on the UK experience (The Committee for National Audit on Adult Intensive Care Units, 2002).

The first research on mortality rates in ICU has been done by Dr. Ng Siew Hian in 2002 at Hospital Ipoh, involving only a logit modeling. The second research is continued by Mohd Saifullah Rusiman in 2003. This second research is about the analysis of logit, probit and linear probability model. Finally, logit model has been appeared to be the best model with $\hat{y}_i = \frac{e^A}{1 + e^A}$, where $A = -$

$4.545 + 0.489 \text{ mec\_vent} + 0.094 \text{ s2s\_disc}$ with 91.91% accuracy of predicting. These results achieve the conclusion that logit modeling is one of the famous modeling if the dependent variable is in the categorical form, although logit modeling have been discovered long time ago by Joseph Berkson in 1944.

A robust estimation of the logit modeling may be done based on the central region of the q-q plot by considering data between the 1st and the 3rd quartile. The LMS estimator introduced by Rousseeuw (1984) is a special case of the LQS estimator, as it minimizes the median of the absolute residuals (Bernholt, T., et al., 2007). Other way is to use the range between the 1st and the 3rd quartile of the quantile-quantile (q-q) plot to check if the heating losses and the outdoor temperature have the same distribution and, if yes, to perform the regression in this range of the q-q plot. (Cristian Ghiaus, 2006).

A common problem in linear regression is that largely aberrant values can strongly influence the results. The least quartile difference (LQD) regression estimator is highly robust, since it can resist up to almost 50% largely deviant data values without becoming extremely biased. Additionally, it shows good behavior on Gaussian data in contrast to many other robust regression methods. However, the LQD is not widely used yet due to the high computational effort needed when using common algorithms (Bernholt, T., et al., 2007).

Least square (LS) is one of the most popular regression methods since it is computationally simple and it has minimal variance for Gaussian distributed data. However, the LS estimator can be strongly influenced by outlying values. The aim of robust regression in the plane is to fit a straight line through a set of two-dimensional points in such a way that outliers do not affect the fit. To quantify the robustness of an estimator, Donoho and Huber (1983) define the (finite sample) breakdown point as the smallest fraction of data points that needs to be changed to have an unbounded effect on the estimate. Thus here, the term

‘robust’ stands for a high breakdown value. The LS estimator is not robust, as its breakdown value is $1/n$, i.e. a single outlier can have arbitrarily large effects on the estimation.

The least quartile difference (LQD) estimator, introduced by Croux et al. (1994), has a breakdown point of $(n/2)/n$ if the data fulfill certain requirements. This means that up to 50% of the data can be contaminated without ruining the fit. Also, 50% represents an upper bound for the breakdown point in the class of regression-equivariant estimators. Further, the LQD estimator shows a much better performance at Gaussian distributed errors than other maximum breakdown methods such as least median of squares (LMS) (Rousseeuw, 1984) or least trimmed squares (LTS) regression (Rousseeuw and Leroy, 1987). Deepest regression (DR) (Rousseeuw and Hubert, 1999) shows similar behavior as the LQD for normally distributed samples but does not have a maximum breakdown value. As a drawback, robust regression methods generally need more computation time than non-robust methods. (Bernholt, T., et al., 2007).

3. METHODOLOGY

Before the calculation of estimator, multicollinearity diagnostic should be done to the data. The assumption of residuals should be fulfilled using residual analysis. Then, the influential variables are selected using forward stepwise method. For logit model, the function of logistic distribution is (Kleinbaum, 1994),

$$\hat{Y}_i = P(Y_i = 1 | X_i) = p_i = \frac{e^{\sum_{j=0}^k \beta_j X_{ij}}}{1 + e^{\sum_{j=0}^k \beta_j X_{ij}}}$$

The maximum likelihood estimation used to maximize the function (Mohd Saifullah Rusiman, 2003),

$$L(\mathbf{Y} | \mathbf{X}, \boldsymbol{\beta}) = \prod_{i=1}^n \left(\frac{e^{\sum_{j=0}^k \beta_j X_{ij}}}{1 + e^{\sum_{j=0}^k \beta_j X_{ij}}} \right)^{Y_i} \left(\frac{1}{1 + e^{\sum_{j=0}^k \beta_j X_{ij}}} \right)^{1-Y_i}$$

The first and second derivative given by,

$$\begin{aligned} \frac{\partial \ln L(\mathbf{Y} | \mathbf{X}, \boldsymbol{\beta})}{\partial \beta_j} &= \sum_{i=0}^n \left(Y_i - \frac{1}{1 + e^{-\sum_{j=0}^k \beta_j X_{ij}}} \right) X_{ij} = 0 \\ \frac{\partial^2 L}{\partial \beta_j \partial \beta_k} &= \sum_{i=0}^n X_{ij} \left(1 + e^{-\sum_{j=0}^k \beta_j X_{ij}} \right)^{-2} (-X_{ik} e^{-\sum_{j=0}^k \beta_j X_{ij}}) \\ &= -\sum_{i=0}^n X_{ij} p_i (1 - p_i) X_{ik} \end{aligned}$$

The estimation of $\hat{\beta}_k$ discovered by using maximum likelihood method where the numerical iteration, Newton-Raphson should be used,

$$\mathbf{b}^{(m)} = \mathbf{b}^{(m-1)} + (\mathbf{X}^T \mathbf{V}^{(m-1)} \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{y} - \mathbf{q}^{(m-1)})$$

For LQD modeling, design the q-q plot for each X_i and Y_i in order to find the data with the range of 1st and 3rd quartile. The q-q plot for X_i and Y_i is not suitable if the data are binary data. The logit model should be done to this data and use Microsoft Excel to find the percentage of accuracy. Other methods are get the range for 1st and 3rd quartile and 1st and 50th percentile based on the absolute of residual in order to make another logit model. The comparison among 4 methods will be done to find the better model by using the percentage of accuracy. After comparing the four models, the better model chosen having the highest probability of accuracy and can be proposed as the new effective model.

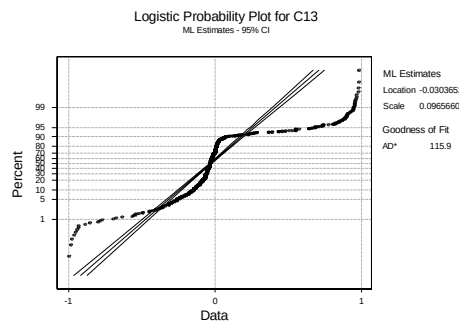
4. ANALYSIS

In this research, data are collected from intensive care unit (ICU), general hospital in Johor. Data collected involving 1311 patients in ICU within the interval 1.1.2001 and 25.8.2002. Dependent variable is patient status with 0 and 1 code. 0 is coded when patient alive in hospital or ICU, whereas 1 is coded when patient died in hospital or ICU. Actually, there are 8 independent variables i.e. sex, race, days in ICU, organ failure, comorbid diseases, mechanical ventilator, score of SAPS II admit and score of SAPS II discharge from hospital.

The correlation analysis has shown that no value of correlation exceed 0.99 between X_i . The value of VIF doesn't exceed 10. The value of conditional index is 490.714 and less than 1000. The value of variance inflation is also less than 0.99. Its indicate that multicollinearity doesn't exist among variable of X .

For logit model, the initial assumption that should be fulfilled is the residual should be distributed as logistic distribution. Referring to the Figure 1, logistic probability plot are nearly linear. Hence, the residual is logistic distributed.

Figure 1 : Logistic Probability Plot for Logit Model



Analysis data in Table 1 has shown that only 2 variable of X are significant. The forward stepwise is used in choosing the influential variable with 91.91% accuracy. The logit model can be stated as,

$$\hat{Y}_i = P(Y_i=1) = \frac{e^{-4.545 + 0.489 \text{ mec\_vent} + 0.094 \text{ s2s\_disc}}}{1 + e^{-4.545 + 0.489 \text{ mec\_vent} + 0.094 \text{ s2s\_disc}}}$$

or $\text{logit}(Y_i) = -4.545 + 0.489 \text{ mec\_vent} + 0.094 \text{ s2s\_disc}$

Table 1 : Output SPSS for Logit Model

Block 0: Beginning Block

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)
Step 0 Constant	-.838	.060	193.862	1	.000	.433

Block 1: Method = Forward Stepwise (Conditional)

Omnibus Tests of Model Coefficients

	Chi-square	df	Sig.
Step 2 Step	10.579	1	.001
Block	1004.532	2	.000
Model	1004.532	2	.000

Classification Table

		Predicted		
		STA	TUS	Percentage
Observed		.00	1.00	Correct
Step 2 STATUS	.00	894	21	97.7
	1.00	85	311	78.5
Overall	Percentage			91.9

a. The cut value is .500

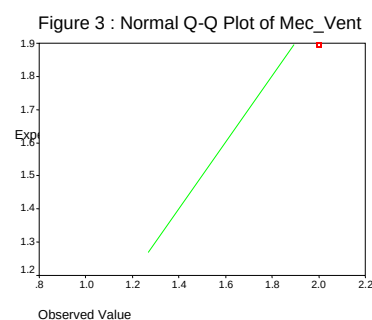
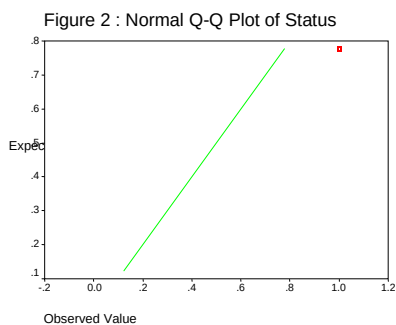
Variables in the Equation

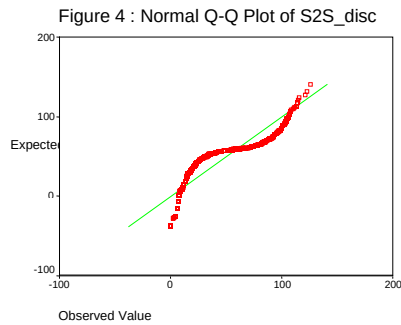
	B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 S2S_DISC	.097	.006	271.114	1	.000	1.102
Constant	-4.316	.210	421.773	1	.000	.013
Step 2 MEC_VENT(1)	.489	.163	9.029	1	.003	1.631
S2S_DISC	.094	.006	255.823	1	.000	1.098
Constant	-4.545	.235	372.782	1	.000	.011

a. Variable(s) entered on step 1: S2S_DISC.

b. Variable(s) entered on step 2: MEC_VENT.

According to the Figure 2-4, only the q-q plot for variable s2s-disc (score of SAPS II discharge) shown the variation of data since the data are continuous, whereas the data for variable status and mec_vent are binary. In applying the concept of least quartile difference modeling, only the 1st and 3rd quartile of data used in order to find another logit model.





Another logit modeling was made based on the range of 1st and 50th percentile and 1st and 3rd quartile for the absolute value of residual. The comparisons among 4 methods are summarized as in Table 2.

Table 2 : The Comparisons among 4 Methods in Logit Modeling

Method	Logit modeling with initial data	1 st and 50 th percentile for residual	1 st and 3 rd quartile for residual	Least Quartile Difference modeling
Model	$\hat{Y}_i = \frac{e^A}{1 + e^A}$ where $A = -4.545$ $+0.489mec\_vent$ $+0.094s2s\_disc$	$\hat{Y}_i = \frac{e^A}{1 + e^A}$ where $A = -43.192$ $+4.566mec\_vent$ $+0.851s2s\_disc$	$\hat{Y}_i = \frac{e^A}{1 + e^A}$ where $A = -63.513$ $+6.781mec\_vent$ $+1.312s2s\_disc$	$\hat{Y}_i = \frac{e^A}{1 + e^A}$ where $A = -6.781$ $+1.053mec\_vent$ $+0.118s2s\_disc$
% accuracy	91.91%	92.51%	92.05%	92.70%

The comparison as shown in the Table 2 indicates that the least quartile difference modeling has been appeared to be the best model in analyzing binary data.

5. CONCLUSION

The least quartile difference estimator, introduced by Croux has a breakdown point of 1/2 if the data fulfill certain requirements. This means that up to 50% of the data can be contaminated without ruining the fit. Also, 50% represents an upper bound for the breakdown point in the class of regression-equivariant estimators. A robust estimation of the logit modeling was done based on the central region of the q-q plot by considering data between the 1st and the 3rd quartile. The comparison between logit modeling with initial data and the least quartile difference modeling indicates that the least quartile difference modeling has been appeared to be the better model and more robust in analyzing binary data with the higher percent of accuracy. This new technique modeling could be proposed as one of the best modeling in analyzing the data. This

modeling as stated $\hat{Y} = P(Y_i = 1) = \frac{e^{-6.781 + 1.053mec\_vent + 0.118s2s\_disc}}{1 + e^{-6.781 + 1.053mec\_vent + 0.118s2s\_disc}}$ indicate

that the mortality rates will increase if the variable of *mec_vent* and *s2s_disc* increase. Hence, the expected of mortality rate at ICU could be predicted based on the variable *mec_vent* and *s2s_disc*.

REFERENCES

1. J. Agulló, *An exchange algorithm for computing the least quartile difference estimator*. Journal of Metrika 55, 3–16, 2002.
2. A. Agresti, *An Introduction to Categorical Data Analysis*. New York : John Wiley & Sons, Inc, 1996.
3. J. H. Aldrich, and F.D. Nelson, *Linear Probability, Logit and Probit Models*. California : Sage Publications, Inc., 1985.
4. E. B. Andersen, *Introduction to the Statistical Analysis of Categorical Data*. Germany : Springer-Verlag Berlin, Heidelberg, 1997.
5. H. Barreto, Maharry, D., *Least median of squares and regression through the origin*. Journal of Computational Statistical Data Analysis. 50, 1391–1397, 2006.
6. T. Bernholt, et al., *Computing the least quartile difference estimator in the plane*. Journal of Computational Statistical Data Analysis, 2007.
7. B.J.T. Morgan, *Analysis of Quantal Response Data*. London : Chapman & Hall, 1992.
8. Cristian Ghiaus, *Experimental estimation of building energy performance by robust regression*. Journal of Energy and Buildings 38, 582–587, 2005.
9. A. J. Dobson, *An Introduction to Generalized Linear Models*. London : Chapman & Hall, 1991.
10. J.L. Horowitz, and N.E. Savin, *Binary Response Models : Logits, Probits and Semiparametrics*. Journal of Economic Perspectives, Vol 15, Issue 4 : 43-56, 2001.
11. D. G. Kleinbaum, *Logistic Regression : A Self-Learning Text*. New York : Springer-Verlag, New York, Inc, 1994.

Mohd Saifullah Rusiman: Department of Mathematics, Centre for Science Studies, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Johor, Malaysia.
E-mail: saifulah@uthm.edu.my, msr9339@yahoo.com

STABILITY OF A NEURONAL MODEL WITH DISTRIBUTED DELAY AND TIME-DEPENDENT PARAMETERS

SARIYASA

Abstract. The convergence dynamic of a neural network model with time-dependent parameters is studied. Distributed delays represented by delay kernels are incorporated in the model. An easily verifiable delay-independent sufficient condition for the exponential stability of the network is established.

Key words and Phrases: exponential stability, distributed delay, time-dependent parameters.

1. INTRODUCTION

To a large extent, the existing literature on theoretical studies of artificial neural networks is predominantly concerned with autonomous systems containing temporally uniform network parameters. Literature dealing with time-varying network parameters seems to be scarce; such studies are however important to understand the dynamical characteristics of neuron behaviour in time-varying environments. Ott, et al. [5] have shown that one can convert a chaotic attractor to any one of a large number of possible time periodic motions by the introduction of time-dependent perturbations to the network parameters. Thus it is worthwhile to consider time-dependent parameters in the equations governing the dynamics of neurons.

In addition to the time-dependent parameters, we incorporate time delays in the processing part of the neuron's architecture. The incorporation of delays in the model is due to the fact that delays are naturally present in biological networks

2000 Mathematics Subject Classification: 34K20, 45J05, 92B20

through synaptic transmission, finite conduction velocity, and neural processing of input stimuli. It is also important to incorporate delays in some networks such as cellular neural networks which can be used to solve some moving image processing problem (Chua and Roska [1]).

Owing to the spatial nature of the neural networks, it is necessary to introduce continuously distributed delays over a certain duration of time such that the distant past has less influence compared with the recent behaviour of the state. The duration over which the past effects affect the current state can extend over a finite or infinite interval (Gopalsamy and He [2, 3]). Thus, we consider a neural network model represented by a system of nonautonomous integrodifferential equations.

Applications of neural networks depend on the dynamics of the governing system of neural networks. However, a common requirement for the applications is to improve the rate of convergence to equilibrium state of the networks and hence to reduce the computation time. Consequently, it is required to design a network that possesses exponential stability. However, the presence of the delays in the networks may affect the convergence rate. Accordingly, it is necessary to eliminate the effects of delay on the convergence rate. Thus, in this article we shall derive a delay-independent sufficient condition for the model to be exponentially stable.

2. THE NEURAL NETWORK MODEL

Consider a neural network model with m neurons given by the following system of integrodifferential equations

$$\frac{dx_i(t)}{dt} = -a_i(t)x_i(t) + \sum_{j=1}^m b_{ij}(t) \tanh \left(\int_0^\infty K_{ij}(s)x_j(t-s)ds \right), \quad t > 0 \quad (1)$$

in which $i \in \{1, 2, \dots, m\}$; $a_i(\cdot)$ and $b_{ij}(\cdot)$ denote continuous real-valued functions defined on $(-\infty, \infty)$; $x_i(t)$ correspond to the membrane potential of the neurons i at time t ; $a_i(\cdot)$ denote quantitative measures of the neuronal dissipation or negative feedback terms; $b_{ij}(\cdot)$ denotes the neuron gains.

The nonlinear activation functions are chosen to be continuous and differentiable nonlinear sigmoidal functions. The choice of sigmoidal activation functions is natural since biological neurons are characterized by a sigmoidal response function (Marcus, et al. [4]). For the purpose of this article we have chosen an activation function of the form $f_j(x) = \tanh(x)$ which has been widely adopted for neural network models.

The equation (1) is supplemented with an initial condition of the form $x_i(s) = \varphi_i(s)$, $s \in (-\infty, 0]$, and $\varphi_i : (-\infty, 0] \rightarrow R$ are continuous and bounded. A solution of (1) is denoted by a vector $\mathbf{x}(t) = (x_1(t), x_2(t), \dots, x_m(t))$.

The delay kernels $K_{ij}(\cdot)$ are assumed to satisfy the following conditions:

- (i) $K_{ij} : [0, \infty) \mapsto [0, \infty)$;

- (ii) $K_{ij}(\cdot)$ are bounded and continuous on $[0, \infty)$;
- (iii) $\int_0^\infty K_{ij}(s)ds = 1$;
- (iv) there exists a positive number μ such that $\int_0^\infty K_{ij}(s)e^{\mu s}ds < \infty$.

3. EXPONENTIAL STABILITY OF THE NETWORK

The following theorem guarantees the exponential stability of the equation (1).

Theorem 3.1. *Suppose the coefficients $a_i(\cdot)$ and $b_{ij}(\cdot)$ are bounded and continuous on $(-\infty, \infty)$ such that*

$$0 < \underline{a}_i > \sum_{j=1}^m \hat{b}_{ji} \quad (2)$$

where $\underline{a}_i = \inf_{t \in \mathbb{R}} a_i(t)$ and $\hat{b}_{ij} = \sup_{t \in \mathbb{R}} |b_{ij}(t)|$. Then the equation (1) is exponentially stable in the sense that there exist constants $\alpha \geq 1$ and $\epsilon > 0$ such that

$$\sum_{i=1}^m |x_i(t) - y_i(t)| \leq \alpha e^{-\epsilon t} \sum_{i=1}^m \left(\sup_{s \in (-\infty, 0]} |x_i(s) - y_i(s)| \right), \quad t > 0.$$

Proof. Let $x_i(t)$ and $y_i(t)$ be two arbitrary solutions of (1). Then

$$\begin{aligned} \frac{d}{dt}[x_i(t) - y_i(t)] &= -a_i(t)[x_i(t) - y_i(t)] + \\ &\quad \sum_{j=1}^m b_{ij}(t) \left[\tanh \left(\int_0^\infty K_{ij}(s)x_j(t-s)ds \right) - \tanh \left(\int_0^\infty K_{ij}(s)y_j(t-s)ds \right) \right] \end{aligned}$$

and using upper right derivatives we obtain

$$\frac{d^+}{dt}|x_i(t) - y_i(t)| \leq -\underline{a}_i|x_i(t) - y_i(t)| + \sum_{j=1}^m \hat{b}_{ij} \int_0^\infty K_{ij}(s)|x_j(t-s) - y_j(t-s)|ds. \quad (3)$$

Consider the function $G_i : [0, \mu] \rightarrow (-\infty, \infty)$ defined by

$$G_i(\epsilon_i) = \underline{a}_i - \epsilon_i - \sum_{j=1}^m \hat{b}_{ji} \int_0^\infty K_{ji}(s)e^{\epsilon_i s}ds, \quad \epsilon_i \in [0, \mu]. \quad (4)$$

Due to the assumption on the delay kernels, each $G_i(\cdot)$ is well-defined on $[0, \mu]$. From (2) and (4) we have

$$G_i(0) = \underline{a}_i - \sum_{j=1}^m \hat{b}_{ji} > 0.$$

Since $G_i(\cdot)$ is continuous and decreasing on $[0, \mu]$, it follows that there exists an $\epsilon \in (0, \mu]$ such that

$$G_i(\epsilon) = \underline{a}_i - \epsilon - \sum_{j=1}^m \hat{b}_{ji} \int_0^\infty K_{ji}(s) e^{\epsilon s} ds \geq 0. \quad (5)$$

Now, let

$$z_i(t) = e^{\epsilon t} |x_i(t) - y_i(t)|, \quad t \in R \quad (6)$$

so that from (3) and (6) we derive

$$\frac{d^+ z_i(t)}{dt} \leq -(\underline{a}_i - \epsilon) z_i(t) + \sum_{j=1}^m \hat{b}_{ij} \int_0^\infty K_{ij}(s) e^{\epsilon s} z_j(t-s) ds. \quad (7)$$

Define a Lyapunov functional $V(\cdot)$ by

$$V(t) = \sum_{i=1}^m \left\{ z_i(t) + \sum_{j=1}^m \hat{b}_{ij} \int_0^\infty K_{ij}(s) e^{\epsilon s} \left(\int_{t-s}^t z_j(w) dw \right) ds \right\}.$$

Note that $V(t) > 0$ for $t > 0$ and that $V(0) < \infty$. Now calculate the rate of change of $V(t)$ along the solutions of (7) to yield

$$\frac{d^+ V(t)}{dt} \leq - \sum_{i=1}^m \left\{ \underline{a}_i - \epsilon - \sum_{j=1}^m \hat{b}_{ji} \int_0^\infty K_{ji}(s) e^{\epsilon s} ds \right\} z_i(t), \quad t > 0. \quad (8)$$

Applying (5) in (8) gives $\frac{d^+ V(t)}{dt} \leq 0$ for $t > 0$ and so $V(t) \leq V(0)$ for $t > 0$. This implies that

$$\sum_{i=1}^m z_i(t) \leq \sum_{i=1}^m \left\{ 1 + \sum_{j=1}^m \hat{b}_{ji} \int_0^\infty s K_{ji}(s) e^{\epsilon s} ds \right\} \left(\sup_{w \in (-\infty, 0]} z_i(w) \right), \quad t > 0. \quad (9)$$

Using (6) in (9) we obtain

$$\sum_{i=1}^m |x_i(t) - y_i(t)| \leq \alpha e^{-\epsilon t} \sum_{i=1}^m \left(\sup_{w \in (-\infty, 0]} |x_i(w) - y_i(w)| \right), \quad t > 0$$

where

$$\alpha = \max_i \left(1 + \sum_{j=1}^m \hat{b}_{ji} \int_0^\infty s K_{ji}(s) e^{\epsilon s} ds \right) \geq 1.$$

The exponential stability of the equation (1) follows. This completes the proof.

4. CONCLUDING REMARKS

In this article we have studied a neural network model with distributed delay. The parameters of the model are assumed to be time-dependent. Using the Lyapunov functional approach, we investigate the convergence dynamic of the model. In particular, we have derived a sufficient condition for the exponential stability of the model. The condition is delay-independent and is easily verifiable since it is expressed in terms of the parameters of the model.

Acknowledgement. The author wishes to thank the referee(s) for the suggestions.

REFERENCES

1. L. O. CHUA AND T. ROSKA, *CNN: A paradigm of nonlinear dynamics in space*, *World Congress of Nonlinear Analysis '92, Tampa, FL, 1992*, Berlin: de Gruyter, pp. 2979–2990, 1996.
2. K. GOPALSAMY AND X. Z. HE, Delay-independent stability in bidirectional associative memory networks, *IEEE Trans. Neural Networks*, **5**, 998–1002, 1994a.
3. K. GOPALSAMY AND X. Z. HE, Stability in asymmetric Hopfield nets with transmission delays, *Physica D*, **76**, 344–358, 1994b.
4. C. M. MARCUS, F. R. WAUGH, AND R. M. WESTERVELT, Nonlinear dynamics and stability of analog neural networks, *Physica D*, **51**, 234–247, 1991.
5. E. OTT, C. GREBOGI, AND J. A. YORKE, Controlling chaos, *Phys. Rev. Lett.*, **64**, 1196–1199, 1990.

SARIYASA: Jurusan Pendidikan Matematika, Universitas Pendidikan Ganesha, Singaraja, Bali, Indonesia.

E-mail: sariyasa64@yahoo.com

ASYMPTOTIC NORMALITY AND MODEL SELECTION IN NEURAL NETWORKS BY USING INFERENCE OF R^2 INCREMENTAL AND WALD TEST FOR TIME SERIES FORECASTING

SUHARTONO, SUBANAR, SURYO GURITNO

Abstract. The aim of this paper is to discuss an asymptotic normality of parameters estimator in neural networks (NN) and propose a new procedure for model selection in NN for time series forecasting. Firstly, we focus on the derivation of the asymptotic properties and asymptotic normality of NN parameters estimator. Then, we develop the model selection strategies based on statistical concepts particularly statistics test based on the Wald test. This test is combined with the inference of R^2 incremental. In this paper, we employ this new procedure in two main approaches for model selection in neural networks, i.e. bottom-up or forward by using the inference of R^2 incremental and top-down or backward by implementing Wald test. Bottom-up approach starts with an empty model, whereas top-down approach begins with a large neural networks model. We use simulation data as a case study. The results show that a combination between statistical inference of R^2 incremental and Wald test is an effective procedure for model selection in neural networks for time series forecasting.

Key words and Phrases: asymptotic normality, Wald test, R^2 incremental, model selection

1. INTRODUCTION

In recent years, an impressive array of publications has appeared claiming considerable successes of neural networks (NN) in data analysis and engineering applications. NN model is a prominent example of such a flexible functional form. The use of the NN model in applied work is generally motivated by a mathematical result stating that under mild regularity conditions, a relatively simple NN model is capable for approximating any Borel-measurable function to any given degree of accuracy (see e.g. [5, 6]).

In the application of NN, it contains limited number of parameters (weights). How to find the best NN model, i.e. how to find an accurate combination between number of input variables and unit nodes in hidden layer, is a central topic on the some NN literatures that discussed on many articles and books (see e.g. [2, 4, 11]). In general, there are two procedures usually used to find the best NN model (the optimal architecture), those are “general-to-specific” or “top-down” and “specific-to-general” or “bottom-up” procedures. “Top-down” procedure is started from complex model and then applies an algorithm to reduce number of parameters by using some stopping criteria, whereas “bottom-up” procedure works from a simple model. The first procedure in some literatures is also known as “pruning” (see [10]), or “backward” method in statistical modeling. The second procedure is also known as “constructive

learning” and one of the most popular is “cascade correlation” (see e.g. [3, 9]), and it can be seen as “forward” method in statistical modeling.

The aim of this paper is to discuss and propose a new procedure by combining the inference of R^2 incremental and Wald test. The inference of R^2 incremental is implemented on forward scheme to select the optimal number of hidden nodes, whereas Wald test is employed on backward scheme to determine the optimal input units. We emphasize on the used of NN for time series forecasting.

2. FEEDFORWARD NEURAL NETWORKS

Feedforward Neural Networks (FFNN) is the most popular NN models for time series forecasting applications. Figure 1 shows a typical three-layer FFNN used for forecasting purposes. The input nodes are the previous lagged observations, while the output provides the forecast for the future values. Hidden nodes with appropriate nonlinear transfer functions are used to process the information received by the input nodes.

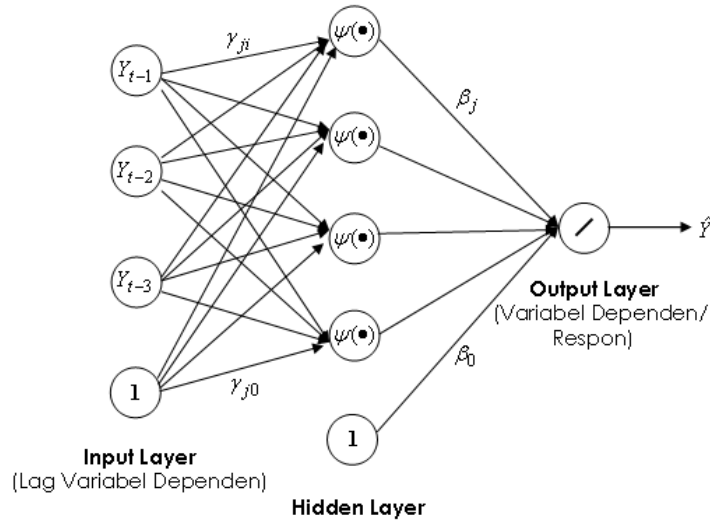


Figure 1. Architecture of neural network model with single hidden layer

The model of FFNN in Figure 1 can be written as

$$y_t = \beta_0 + \sum_{j=1}^q \beta_j \psi \left(\sum_{i=1}^p \gamma_{ij} y_{t-i} + \gamma_{oj} \right) + \varepsilon_t, \quad (1)$$

where p is the number of input nodes, q is the number of hidden nodes, ψ is a sigmoid transfer function such as the logistic, $\{\beta_j, j = 0, 1, \dots, q\}$ is a vector of weights from the hidden to output nodes and $\{\gamma_{ij}, i = 0, 1, \dots, p; j = 1, 2, \dots, q\}$ are weights from the input to hidden nodes. Note that equation (1) indicates a linear transfer function is employed in the output node.

Functionally, the FFNN expressed in equation (1) is equivalent to a nonlinear AR model. This simple structure of the network model has been shown to be capable of approximating arbitrary function (see e.g. [5, 6]). However, few practical guidelines exist for building a FFNN for a time series, particularly the specification of FFNN architecture in terms of the number of

input and hidden nodes is not an easy task.

Kaashoek and Van Dijk [7] introduced a “pruning” procedure by implementing three kinds of methods to find the best FFNN model; those are incremental contribution (R^2 incremental), principal component analysis, and graphical analysis. Whereas, Swanson and White [15, 16] applied a criterion of model selection, SIC, on “bottom-up” procedure to increase number of unit nodes in hidden layer and input variables until finding the best FFNN model. Recently, Suhartono et al. [14] proposed a new forward procedure based on the statistical inference of R^2 incremental contribution.

2.1. Backpropagation Algorithm

Backpropagation algorithm is an algorithm that usually used to estimate the FFNN weights (parameters). Ripley [11] stated that the existence of the function approximation was not useful if there was not known the way to find this function. This condition affected many researches about NN for many years. The main idea of the approximation by using NN is started by Rumelhart-McClelland learning for fitting parameters by employing least squares method. The training of the NN involves adjusting the weights of the network such that the output generated by the network for the given input $x_{(k)}$ is as “close” to $\hat{y}_{(k)} = f(x; w)$ as possible. Formally, this can be formulated as the optimization problem by finding weights, $w = (\gamma_{ij}, \beta_j)$, to minimize

$$E(w) = \sum_{k=1}^n \|y_{(k)} - f(x_{(k)}; w)\|^2 \quad (2)$$

as done in nonlinear regression (see e.g. [1, 12]).

Gradient descent is known as one of the oldest optimization methods. This method is based on a linear approximation of the error function given by

$$E(w + \Delta w) \approx E(w) + \Delta w^T E'(w). \quad (3)$$

The weights update is

$$\Delta w = -\eta E'(w), \quad \eta > 0, \quad (4)$$

where η is learning rate. Suhartono et al. [13] derived a corollary about backpropagation algorithm to find the optimal weights of FFNN for time series forecasting as illustrated in Figure 1.

2.2. Asymptotic Properties of FFNN Estimator

The large-sample properties of learning backpropagation in single hidden layer feedforward networks have been studied further by White [17, 18]. The aim of learning networks by using backpropagation is to find the solution w^* on the optimization problem $\arg \min_{w \in W} Q(w)$, i.e.

$$w^* = \arg \min_{w \in W} (Q(w) = E[(Y - f(X, w))^2 / 2]), \quad (5)$$

where w^* is index of an optimal networks.

With squared error penalty, learning must arrive at w^* , which solve

$$\begin{aligned} \min_{w \in W} (E[(Y - f(X, w))^2 / 2] = E[(Y - E(Y | X))^2 / 2] \\ + E([E(Y | X) - f(X, w)]^2 / 2)). \end{aligned} \quad (6)$$

Finding w^* is precisely the problem of finding the parameters of an optimal least squares approximation to $E(Y | X)$, the conditional expectation of Y given X .

Specifically, given target/input pairs (Y_t, X_t) with $t = 1, 2, \dots, n$, randomly drawn from the operating environment, then $\hat{w}_n$ is the nonlinear least squares estimator, i.e.

$$\hat{w}_n = \arg \min_{w \in W} Q_n(w) = n^{-1} \sum_{t=1}^n (Y_t - f(X_t, w))^2 / 2. \quad (7)$$

Nonlinear regression is an established method that has been completely analyzed in statistics and econometrics literatures.

White [18] provided a formal statement of condition sufficient to guarantee convergence of $\hat{w}_n$, as stated in the following theorem.

Theorem 2.1. (White [18]). *Let $(\Omega, \mathcal{F}, P)$ be a complete probability space on which is defined the sequence of independent identically distributed random variables $\{Z_t\} = (Z_t : \Omega \rightarrow \mathcal{R}^v, t = 1, 2, \dots)$, $v \in \mathbb{N} \equiv \{1, 2, \dots\}$. Let $l : \mathcal{R}^v \times W \rightarrow \mathcal{R}$ be a function such that for each w in W , a compact subset of $\mathcal{R}^s$, $s \in \mathbb{N}$, $l(\cdot, w)$ is measurable- B^v (where B^v is the Borel σ -field generated by the open sets of $\mathcal{R}^v$), and for each z in $\mathcal{R}^v$, $l(z, \cdot)$ is continuous on W . Suppose further that there exists $d : \mathcal{R}^v \rightarrow \mathcal{R}^+$ such that for all w in W , $|l(z, w)| \leq d(z)$ and $E(d(Z_t)) < \infty$ (i.e., l is dominated on W by an integrable function).*

Then for each $n = 1, 2, \dots$ there exists a solution $\hat{w}_n$ to the problem $\min_{w \in W} \hat{Q}_n(w) \equiv n^{-1} \sum_{t=1}^n l(Z_t, w)$ and $\hat{w}_n \rightarrow W^*$ a.s. - P , where $W^* \equiv \{w^* \in W : Q(w^*) \leq Q(w) \text{ for all } w \in W\}$, $Q(w) = E(l(Z_t, w))$.

2.2. Asymptotic Normality of FFNN Estimator

The appropriate formal concept for studying the limiting distribution of $\hat{w}_n$ is that of convergence in distribution. Asymptotic distribution of $\hat{w}_n$ depends on the nature of W^* . In general, W^* may consist of isolated points and/or isolated "flat". If convergence to a flat occurs, then the estimated weights $\hat{w}_n$ have a limiting distribution that can be analyzed using the theory of Phillips [8] for "partially identified" models. These distributions belong to the "limiting mixed Gaussian" (LMG) family introduced by Phillips. When w^* is locally unique, the model is said to be "locally identified" and estimated weights $\hat{w}_n$ converging to w^* have a limiting multivariate normal distribution.

The condition ensuring that $\hat{w}_n$ is the multivariate normal distribution have been studied further by White [18]. The following theorem is one of the results of White's works.

Theorem 2.2. (White [18]). *Let $(\Omega, \mathcal{F}, P)$, $\{Z_t\}$, W and l be as in Theorem 2.1, and suppose that $\hat{w}_n \rightarrow w^*$ a.s. - P where w^* is an isolated element of W^* interior to W .*

Suppose in addition that for each z in $\mathcal{R}^v$, $l(z, \cdot)$ is continuously differentiable of order 2 on $\int W$; that $E(\nabla l(Z_t, w^*)' \nabla l(Z_t, w^*)) < \infty$; that each element of $\nabla^2 l$ is dominated on W by an integrable function; and that $A^* \equiv E(\nabla^2 l(Z_t, w^*))$ and $B^* \equiv E(\nabla l(Z_t, w^*) \nabla l(Z_t, w^*)')$ are nonsingular $(s \times s)$ matrices, where ∇ and ∇^2 denote the $(s \times 1)$ gradient and $(s \times s)$ Hessian operators with respect to w .

Then $\sqrt{n}(\hat{w}_n - w^*) \xrightarrow{d} N(0, C^*)$, where $C^* = A^{*-1} B^* A^{*-1}$. If in addition each element of $\nabla l \nabla l'$ is dominated on W by an integrable function, then $\hat{C}_n \rightarrow C^*$ a.s. - P , where $\hat{C}_n = \hat{A}_n^{-1} \hat{B}_n \hat{A}_n^{-1}$, and

$$\hat{A}_n = \frac{\sum_{t=1}^n \nabla^2 l(Z_t, \hat{w}_n)}{n}, \quad \hat{B}_n = \frac{\sum_{t=1}^n \nabla l(Z_t, \hat{w}_n) \nabla l(Z_t, \hat{w}_n)'}{n}.$$

White [17] stated that taking one Nonlinear Least Squares (NLS) Newton-Raphson step from the backpropagation estimator asymptotically equivalent to NLS. Thus, tests of hypotheses bases on $\hat{w}_n$ can be conducted for selecting the optimal architecture of FFNN.

2.3. Hypothesis Testing by Using Wald Test

The Wald statistic allows the simplest analysis, although it may or may not be the easiest statistic to compute in a given situation. The motivation for the Wald statistic is that when the null hypothesis is correct $S\hat{w}_n$ should be close to $Sw^* = s$, so a value of $S\hat{w}_n - s$ far from zero is evidence against the null hypothesis.

The theorem about Wald statistic that be used for hypothesis testing of parameters in NN model is constructed as the following results.

Theorem 2.3. *Let the conditions of Theorem 2.2 hold, i.e.*

- (i) $C^{*-1/2}\sqrt{n}(\hat{w}_n - w^*) \xrightarrow{d} N(0, I)$, where $C^* \equiv A^{*-1}B^*A^{*-1}$, and C^{*-1} is $O(1)$,
- (ii) *there exists a matrix $\hat{B}_n$ positive semidefinite and symmetric such that $\hat{B}_n - B^* \xrightarrow{p} 0$. Then $\hat{C}_n - C^* \xrightarrow{p} 0$, where $\hat{C}_n = \hat{A}_n^{-1}\hat{B}_n\hat{A}_n^{-1}$, and*

$$\hat{A}_n = \frac{\sum_{t=1}^n \nabla^2 l(Z_t, \hat{w}_n)}{n}, \quad \hat{B}_n = \frac{\sum_{t=1}^n \nabla l(Z_t, \hat{w}_n) \nabla l(Z_t, \hat{w}_n)'}{n}.$$

And, let $\text{rank}(S) = q \leq k$. Then under $H_0 : Sw^* = s$,

- (i) $\Gamma_n^{-1/2}\sqrt{n}(S\hat{w}_n - s) \xrightarrow{d} N(0, I)$, where $\Gamma_n \equiv SC^*S' = SA^{*-1}B^*A^{*-1}S'$.
- (ii) *The Wald statistic*

$$W_n \equiv n(S\hat{w}_n - s)' \hat{\Gamma}_n^{-1} (S\hat{w}_n - s) \xrightarrow{d} \chi_q^2, \quad (8)$$

where $\hat{\Gamma}_n \equiv S\hat{C}_nS'$.

Proof: We use Corollary 4.24, Proposition 2.30 and Theorem 4.30 at [19] to prove Theorem 2.3 and the results are as follow.

- (i) Under H_0 , $S\hat{w}_n - s = S(\hat{w}_n - w^*)$, so

$$\Gamma_n^{-1/2}\sqrt{n}(S\hat{w}_n - s) = \Gamma_n^{-1/2}SC^{*1/2}C^{*-1/2}\sqrt{n}(\hat{w}_n - w^*).$$

It follows from Corollary 4.24 at [19] that $A_n = S$ and $b_n = \sqrt{n}(\hat{w}_n - w^*)$, so that $\Gamma_n^{-1/2}\sqrt{n}(S\hat{w}_n - s) \xrightarrow{d} N(0, I)$.

- (ii) Based on Theorem 2.2 we have that $\hat{C}_n - C^* \xrightarrow{a.s} 0$, so it imply that $\hat{C}_n - C^* \xrightarrow{p} 0$. By using Proposition 2.30 at [19], where $\hat{\Gamma}_n = g(\hat{C}_n)$ and $\Gamma_n = g(C^*)$, so that $\hat{\Gamma}_n - \Gamma_n \xrightarrow{p} 0$. Given the result in (i), i.e. $\Gamma_n^{-1/2}\sqrt{n}(S\hat{w}_n - s) \xrightarrow{d} N(0, I)$, so by implementing Theorem 4.30 at [19], this yields $W_n \equiv n(S\hat{w}_n - s)' \hat{\Gamma}_n^{-1} (S\hat{w}_n - s) \xrightarrow{d} \chi_q^2$. ■

Thus, a test about the relevance (significance) of input where the hypothesis can be stated as $H_0 : Sw^* = 0$ against $H_1 : Sw^* \neq 0$, can be evaluated by applying Theorem 2.3. As an example, Wald statistic to evaluate this hypothesis testing is

$$\hat{W}_n = n\hat{w}_n'S'(SC^*S')^{-1}S\hat{w}_n,$$

where C^* as defined in earlier section.

2.4. Statistically Inference of R^2 Incremental

Suhartono et al. [14] used statistical inference of R^2 incremental contribution on the forward procedure to determine the best architecture of FFNN. This approach involves three basic steps, which can be described in the following theorem.

Theorem 2.4. *Let the Reduced Model is defined as*

$$Y_t = f(X_t, \hat{\mathbf{w}}_n^{(R)}) + \varepsilon_t^{(R)}, \quad (9)$$

where l_R is the number of parameters to be estimated. And, let the Full Model that is more complex than Reduced Model is defined as

$$Y_t = f(X_t, \hat{\mathbf{w}}_n^{(F)}) + \varepsilon_t^{(F)}, \quad (10)$$

where l_F is the number of parameters in the Full Model, $l_F > l_R$. Then, under $H_0 : \mathbf{w}^{*+} = \mathbf{0}$ or testing for and additional parameters in the Full Model equal to zero, the F statistic can be constructed, i.e.

$$F = \frac{(SSE_{(R)} - SSE_{(F)}) / (df_{(R)} - df_{(F)})}{SSE_{(F)} / df_{(F)}}, \quad (11)$$

or

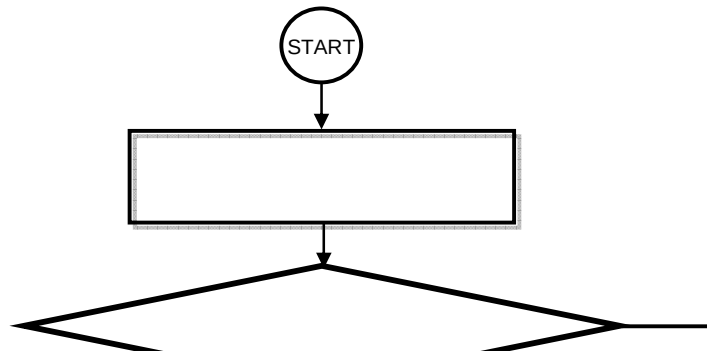
$$F = \frac{R_{\text{incremental}}^2 / (df_{(R)} - df_{(F)})}{(1 - R_{(F)}^2) / df_{(F)}}, \quad (12)$$

where $R_{\text{incremental}}^2 = R_{(F)}^2 - R_{(R)}^2$, $df_{(R)} = n - l_R$ is degree of freedom at Reduced Model, and $df_{(F)} = n - l_F$ is degree of freedom at Full Model.

The procedure proposed by Suhartono et al. [14] started with the forward scheme to find the optimal number of hidden nodes. Then, the forward scheme continued to find the optimal input units, starting with the input which has the largest R^2 .

3. ALGORITHM FOR FFNN MODEL BUILDING

Based on the Wald test and statistically inference of $R_{\text{incremental}}^2$, we proposed a new algorithm for FFNN model building that applied for time series forecasting. This algorithm is started with forward scheme for determining the optimal number of hidden nodes, and combined to the backward scheme for selecting the optimal input units. We illustrate this algorithm in Figure 2.



Apply nonlinearity test for detecting
nonlinear relationship in time series.

Does nonlinearity test
show that there is a nonlinear
relationship?

Specify FFNN model with many (relative) inputs
and 1 neuron unit in the hidden layer as
preliminary step for determining the optimal
number of hidden nodes.

Does the additional 1 node in
the hidden layer yield the significance
of $R^2_{\text{incremental}}$?

Specify FFNN model with FIXED number of
hidden nodes as result of the previous step,
use many (relative) lag inputs in the FFNN.

By using Wald test,
are there the not significance lag
inputs?

END

Figure 2. The proposed algorithm of FFNN model building for time series forecasting

4. RESEARCH METHODOLOGY

In this paper, this new procedure is applied on the simulated data. The simulation experiment is carried out to show how the proposed FFNN modeling procedures work. Finally, the result is compared to the procedures proposed by Kaashoek and Van Dijk [7] and Suhartono et al. [14].

Simulated data are generated as ESTAR (Exponential Smoothing Transition Autoregressive) model, i.e.

$$y_t = 6.5y_{t-1} \cdot \exp(-0.25y_{t-1}^2) + u_t, \quad (13)$$

where $u_t \sim \text{nid}(0, 0.5^2)$.

Time series and the lags plots of this simulated data can be seen in Figure 3. We can observe that data follow nonlinear autoregressive pattern at lag 1.

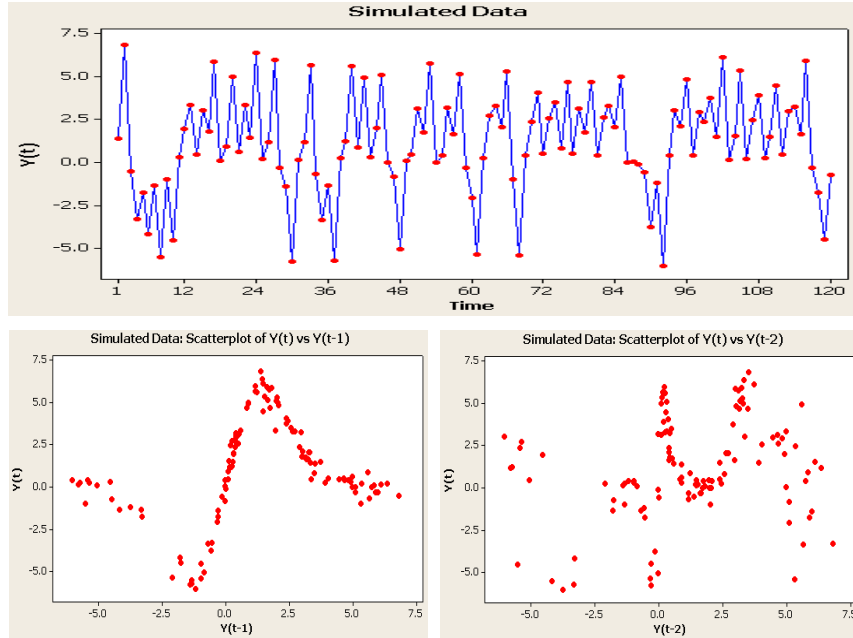


Figure 3. Time series and lags (y_{t-1} and y_{t-2}) plots of simulated data

5. EMPIRICAL RESULTS

In this section, the empirical results for the proposed procedure as illustrated in Figure 2 are presented and discussed. It contains three sub sections, i.e. the results of forward procedure by using $R^2_{\text{incremental}}$, the results of backward procedure by implementing Wald test, and the comparison result with other procedures.

5.1. Results of forward procedure for determining the optimal number of hidden nodes by using inference of $R^2_{\text{incremental}}$

First, we apply the proposed forward procedure starting with a FFNN with six variable inputs ($y_{t-1}, y_{t-2}, \dots, y_{t-6}$) and one constant input to find the optimal nodes in the hidden layer. It's done by implementing inference of $R^2_{\text{incremental}}$. The result of an optimization steps are reported in Table 1. Based on the results in Table 1, we can see that two unit hidden layer cells are the optimal result and further optimization runs are not needed. Then, we continue an optimization to find the optimal input cells.

Table 1. The results of the optimal hidden nodes determination in the forward scheme by implementing inference of $R^2_{\text{incremental}}$

Number of hidden nodes	SBC	R^2	$R^2_{\text{incremental}}$	F test	p-value

0	234.4843	0.161569	-	-	-
1	182.0737	0.547258	0.385689	28.5667	0.00000
2	-72.8918	0.975535	0.428277	7.7719	0.00000
3	-61.4821	0.981029	0.005494	0.0518	0.99993
4	-45.5007	0.984601	0.003572	0.0300	0.99999
5	-33.6011	0.987999	0.003398	0.0251	1.00000
6	2.70047	0.988065	0.000066	0.0004	1.00000

5.2. Results of backward procedure for selecting the optimal input unit by implementing Wald Test

The results of the significance test for FFNN estimator are presented in Table 2. It shows that only input unit 1, i.e. y_{t-1} , is the input cell of the network which has significance estimator, both to hidden node 1 and 2 (h1, h2). Hence, this backward procedure yields the optimal network is FFNN with one input cell (i.e. y_{t-1}) and two nodes in the hidden layer or FFNN(1,2).

Table 2. The results of the optimal input unit determination in the backward scheme by implementing Wald test

Weights	Coefficient	S.E.	Wald test	p-value
b->h1	-0.0122	0.0352	0.1203	0.728733
1->h1	0.9630	0.0556	300.0898	0.000000
2->h1	-0.0165	0.0108	2.3532	0.125021
3->h1	-0.0016	0.0068	0.0555	0.813763
4->h1	-0.0060	0.0068	0.7712	0.379829
5->h1	-0.0009	0.0071	0.0162	0.898732
6->h1	0.0020	0.0069	0.0846	0.771153
b->h2	-0.0005	0.0369	0.0002	0.989196
1->h2	1.3477	0.0746	326.0336	0.000000
2->h2	-0.0175	0.0116	2.2753	0.131440
3->h2	-0.0038	0.0081	0.2198	0.639206
4->h2	-0.0048	0.0080	0.3584	0.549406
5->h2	-0.0006	0.0080	0.0057	0.939963
6->h2	-0.0008	0.0078	0.0104	0.918691
b->o	0.3878	0.1474	6.9216	0.008515
h1->o	-77.4291	23.8600	10.5307	0.001174
h2->o	76.5030	23.9097	10.2381	0.001376

5.3. Comparison Results

There are two main evaluations for the comparison results between this proposed procedures and other procedures proposed Kaashoek and Van Dijk [7] and Suhartono et al. [14], i.e. the final result of architecture FFNN and the number of running steps. In general, the results of this simulation study show that the optimal FFNN architecture yielded by these procedures is the same, i.e. FFNN(1,2).

The comparison result on the number of running steps shows that a combination between inference of $R^2_{\text{incremental}}$ in forward scheme and Wald test in backward scheme yields less running steps than procedures proposed by

Kaashoek and Van Dijk [7] or Suhartono et al. [14]. The results in Table 1 and 2 show that the proposed procedure need 4 running steps, i.e. 3 running for determining the optimal hidden nodes and 1 running for input layer cells.

6. CONCLUSION

Based on the results at the previous sections, we can make two main conclusions, i.e.

- (i). The proposed procedure that contain a combination between inference of $R^2_{\text{incremental}}$ in forward scheme and Wald test in backward scheme work properly for determining the best FFNN architecture.
- (ii). The proposed procedure yields less running steps than procedures proposed by Kaashoek and Van Dijk [7] or Suhartono et al. [14].

In general, the results also show that the proposed procedure gives an advantage for FFNN modeling, i.e. the building process of FFNN model is not a black box. Additionally, we can do further research particularly on the application of this proposed procedure in the real time series data.

REFERENCES

1. D.M. Bates and D.G. Watts, *Nonlinear Regression Analysis and Its Applications*, Wiley, New York, 1988.
2. C.M. Bishop, *Neural Network for Pattern Recognition*, Oxford: Clarendon Press, 1995.
3. S.E. Fahlman and C. Lebiere, The Cascade-Correlation Learning Architecture, in Touretzky, D. S. (ed.), *Advances in Neural Information Processing Systems 2*, Los Altos, CA: Morgan Kaufmann Publishers, pp. 524-532, 1990.
4. H. Haykin, *Neural Networks: A Comprehensive Foundation*, Second edition, Prentice-Hall, Oxford, 1999.
5. K. Hornik, M. Stinchcombe and H. White, Multilayer feedforward networks are universal approximators, *Neural Networks*, **2**, 359–366, 1989.
6. K. Hornik, M. Stinchcombe and H. White, Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks, *Neural Networks*, **3**, 551–560, 1990.
7. J.F. Kaashoek and H.K. Van Dijk, Neural Network Pruning Applied to Real Exchange Rate Analysis, *Journal of Forecasting*, **21**, pp. 559–577, 2002.
8. P.C.B. Phillips, Partially identified econometric models. *Econometric Theory*, vol. **5**, pp. 181-240, 1989.
9. L. Prechelt, Investigation of the CasCor Family of Learning Algorithms, *Neural Networks*, **10**, 885-896, 1997.
10. R. Reed, Pruning algorithms - A survey, *IEEE Transactions on Neural Networks*, **4**, 740-747, 1993.
11. B.D. Ripley, *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge, 1996.
12. G.A.F. Seber and C.J. Wild, *Nonlinear Regression*, Wiley, New York, 1989.
13. Suhartono, S. Rezeki, Subanar and S. Guritno, Optimisation of Backpropagation Algorithm of Feedforward Neural Networks for Regression and Time Series Modeling. *Proceeding International Regional Conference on Mathematics, Statistics and It's Application (IRCMSA)*, Danau Toba, Medan, 2005.
14. Suhartono, Subanar and S. Guritno, Model Selection in Neural Networks by Using Inference of R^2 incremental, PCA, and SIC Criteria for Time Series Forecasting, *JOURNAL OF QUANTITATIVE METHODS: Journal Devoted to The Mathematical and Statistical Application in Various Fields*, Vol. **2**, No. 1, 41-57, 2006.

15. N.R. Swanson and H. White, A model-selection approach to assessing the information in the term structure using linear models and artificial neural networks, *Journal of Business and Economic Statistics*, **13**, 265–275, 1995.
16. N.R. Swanson and H. White, A model-selection approach to real-time macro-economic forecasting using linear models and artificial neural networks, *Review of Economic and Statistics*, **79**, 540–550, 1997.
17. H. White, Some asymptotic results for learning in single hidden layer feedforward networks, *Journal of the American Statistical Association*, Vol. **84**, No. 408, pp. 1003-1013, 1989.
18. H. White, Learning in neural networks: a statistical perspective, *Neural Computation*, **1**, 425-464, 1989.
19. H. White, *Asymptotic Theory for Econometricians*, Academic Press Inc., New York, 1999.

Suhartono: Statistics Department, Institut Teknologi Sepuluh Nopember, Indonesia
PhD Student, Mathematics Department, Gadjah Mada University
E-mail: suhartono@statistika.its.ac.id; har_arema@yahoo.com

Subanar: Mathematics Department, Gadjah Mada University, Indonesia
E-mail: subanar@yahoo.com;

Suryo Guritno: Mathematics Department, Gadjah Mada University, Indonesia
E-mail: suryoguritno@ugm.ac.id

OPTIMIZATION RADIAL BASIS FUNCTION NEURAL NETWORK

BRODJOL SUTIJO, SUBANAR AND SURYO GURITNO

Abstract. Radial Basis Function Neural Network (RBFNN) model is a model with flexible function structure and soft condition, so it can be used to approximate a function. To find best model of RBFNN is needed a combination between input variables and number of nodes in hidden layer. It implied to estimation mean and variance of input variables on each node. Variance component of Gaussian activation function of RBFNN can be considered as weight shift of mean. There are two problems of the variance component or weight. The large of weight or small variance implies useful information can not gain. In contrast if the weight is small or variance is large, then too much useless information will be gotten. In this study, we used K-mean method to estimate parameters in hidden layer. Confidence interval of variance is needed to solve the problem of weight. If variance of a variable on a node out of confidence interval, then link of the variable and the node is insignificant, so the link can be disconnected. To select best model of RBFNN is used Mean Square Error (MSE) and determination coefficient (R^2) criterion.

Key words : RBFNN, K-mean, Variance, MSE, R^2 .

1. INTRODUCTION

Neural network (NN) model is in a group of nonlinear model, which model structure is input pass unit hidden before processes to output. NN model has some design such as perceptron, feed forward neural network (FFNN), radial basis function neural network (RBFNN), etc. RBFNN design is a NN model transform input by using Gaussian activation function in hidden layer unit before linear process in output layer. Some references of NN can be found at Bishop (1995), Reply (1996) and Fine (1999). RBFNN articles are Broomhead and Lowe (1988), Moody and Darken (1989) and Poggio and Girosi (1990). NN model has flexible structure function, so it grows rapidly and has been applied in many areas. This is motivated by soft condition and can used to approximate a function (Cybenko, 1989), (Funahshi, 1989), (Hornik et.al, 1989, 1990), (White, 1990) and (Galland and White, 1992). NN model can be used to check nonlinear component on time series data (White, 1989), (Terasvirta, et.al., 1993) and (Sutijo and Subanar, 2004). Lapedes and Faber (1987) is the first researcher used NN model to forecast time series data.

Optimization RBFNN model had been done researchers; EM algorithm (Lazaro, 2003), bootstrap to select best model (Lendese, 2005), parameter regulation (Sutijo, Subanar and Guritno, 2005), Multiple comparison (Sutijo, Subanar and Guritno, 2006.a, 2006.b) and variance test to test links input variables to node in hidden layer unit (Sutijo, Subanar and Guritno, 2007). To find RBFNN model, we must be get a combination of input variables, number of

nodes or clusters in hidden layer unit. Mean and standard deviation input variables in each node in hidden layer unit are estimated separately. This is a main topic in some NN literature and articles, (Bhisop, 1995), (Reply, 1996), (Fine, 1999), Haykin, 1999), (Reed and Mark II, 1999) and (Lazaro, 2003).

2. RBFNN Modeling

RBFNN design is nonlinear mapping in hidden layer and linear mapping in output layer. RBFNN model map p ($p > 1$) dimension space input variables to one dimension space output variable.

$$F : \mathfrak{R}^p \rightarrow \mathfrak{R}^1 \quad (1)$$

Based on Multivariate theory: There is relationship N point observation $\{x_i \in \mathfrak{R}^p; i = 1, 2, \dots, N\}$ and n real numbers $\{d_i \in \mathfrak{R}; i = 1, 2, \dots, n\}$. A function $F : \mathfrak{R}^p \rightarrow \mathfrak{R}^1$ can be expressed by $F(x_i) = d_i$, $i = 1, 2, 3, \dots, N$.

The theory shows that interpolation by using function $F(x_i) = d_i$ has to pass all training data, so (1) can be expressed by :

$$F(x_i) = d_i, \quad i = 1, 2, 3, \dots, N. \quad (2)$$

RBFNN is designed to the interpolation and figure 1. is architecture of RBFNN.

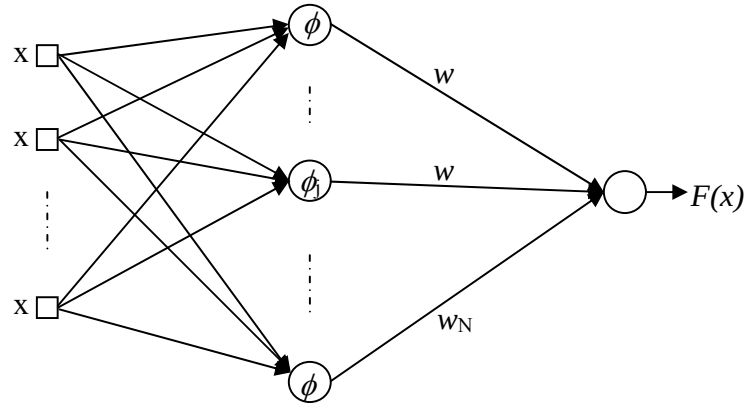


Figure 1. Architecture RBFNN

On RBFNN modeling, the function (2) must be satisfied. Based on the architecture RBFNN, input-output interpolation can be written:

$$F(x) = \sum_{i=1}^N w_i \phi(\|x - x_i\|) \quad (3)$$

Where, $\{\phi(\|x - x_i\|) | i = 1, 2, 3, \dots, N\}$ is nonlinear function set or Radial Basis Function (RBF) and $\|\cdot\|$ is norm Euclidean distance. Gaussian function is a function of RBF that most is used in modeling. There are some of RBF, such as :

1. Function *Thin Plate Spline* $\phi(z) = (z - \mu)^2 \log(z - \mu)$
2. Function Multi quadratic $\phi(z) = [(z - \mu)^2 + \sigma^2]^{1/2}$
3. Function Inverse Multi quadratic $\phi(z) = [(z - \mu)^2 + \sigma^2]^{-1/2}$
4. Function Gaussian $\phi(z) = \exp[-(z - \mu)^2 / \sigma^2]$.

Let, there are N point observations $\{x_i \in \mathcal{R}^p; i = 1, 2, \dots, N\}$ that are center of RBF, so (3) can be written :

$$\begin{bmatrix} \phi_{11} & \phi_{12} & \cdots & \phi_{1N} \\ \phi_{21} & \phi_{22} & \cdots & \phi_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{N1} & \phi_{N2} & \cdots & \phi_{NN} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_N \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_N \end{bmatrix} \quad (4)$$

Where, $\phi_{ii} = \phi(\|x_i - x_i\|)$ $i, j = 1, 2, 3, \dots, N$ and if equation (4) is written in matrix notation is :

$$\Phi \mathbf{w} = \mathbf{d} \quad (5)$$

Where, $\mathbf{d} = [d_1 \ d_2 \ \cdots \ d_N]$ is target vector, $\mathbf{w} = [w_1 \ w_2 \ \cdots \ w_N]$ is a weight vector and $\Phi = \phi_{ii}$ is a positive definite matrix of interpolation matrix.

Approximation function by using RBFNN is similar with interpolation high dimension space to low dimension space. Galerkin method is standard technique to solve the interpolation (Poggio and Girosi, 1990). The technique takes number of node in hidden layer unit less than total observation (N).

$$F^*(x) = \sum_{i=1}^M w_i \phi_i(x) \quad (6)$$

Where, $\{\phi_i(x) | i = 1, 2, \dots, M\}$ is new basis function new set, that linear independent is assumed. Generally, the basis function new set less than number of observation ($M \leq N$) and w_i is weight i node to output.

Model estimate or approximate function by using RBFNN is done two steps. First, unsupervised learning, it is a method to find mean and standard deviation input variables on each node in hidden layer unit. There are some methods of unsupervised learning; exact method, random method, orthogonal least square, K-mean method, etc. Second, weights of node are estimated by using pseudo inverse.

The main problem in RBFNN modeling is getting mean and standard deviation input variables each node. Random method needs large learning to get best performance. K-mean method is one of self-organizing map (SOM), that grows in NN modeling. The method partite data into subgroups or clusters. Data or items in a cluster have similar characteristic and items between clusters have dissimilar characteristic.

Let, RBFNN has M nodes in hidden layer unit, it has M mean vectors or center vectors. By using K-mean method, mean and variance vector are got by iteration. Mean vector at n iteration is $\mu(n)$. Algorithm of K-mean method is:

1. Initialization : Select M point randomly as initial value of mean, $\mu_i(0)$, $i = 1, 2, \dots, M$
2. Similarity : Find $k(x)$ is index of best goodness of fit for mean cluster by minimize Euclidean distance :

$$k(x) = \arg \min_k \|x(n) - \mu_k(n)\|, \quad k = 1, 2, \dots, M$$
3. Update : Calculate new mean cluster

$$\mu(n+1) = \begin{cases} \mu_k(n) + \eta[x(n) - \mu_k(n)], & k = k(x) \\ \mu_k(n) & , \text{yang lain} \end{cases}$$
4. Return (2 to 4) : go to step 2, until $\mu_k(n)$ stable.

3. Variance Test

Gaussian activation function of RBFNN model is :

$$\phi_j(x) = \exp \left(\frac{-\|x - \mu_j\|^2}{2\sigma_j^2} \right) \quad (9)$$

Variance component the model, $1/\sigma^2$, is a weight shift of mean $\|x - \mu_i\|$. There are two problems with the variance component or weight. The large weight or small variance, the Gaussian function is narrow, so do not get useful information. Conversely small weight or large variance, the function get much information include unusable. We need lower and upper limit for variance to solve the problem.

Lower and upper limit of variance are $(1 - \alpha)$ confidence interval limit of variance ($0 < \alpha < 1$). Bases on Chi-square distribution:

$$\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma^2} \sim \chi^2_{(n-1)} \quad (10)$$

The value of σ^2 is unknown, so we have to do simulation to approximate σ^2 . By using n times simulation and ordered, lower limit is $n\alpha/2$ and upper limit is $n(1 - \alpha/2)$ of the order.

Let, $(n-1)Sd^2 = \sum_{i=1}^n (x_i - \bar{x})^2$ on outside confidence interval, it is clue that link input variable to a node candidate to disconnect, because the variance do not get information or more unusable information are got.

4. METHODOLOGY

Time series data especially Indonesia consumer index (CPI), January 1999 – October 2006, is used to applied the theory approximation function by using RBFNN. The data is published by Statistic Center Agency (Badan Pusat Statistik = BPS). Steps analysis data are :

1. Modeling data by using time series method (ARIMA).
2. Data preprocessing, to eliminate nonrandom component, in this case is detrended.
3. K-mean Cluster, to find mean and variance input variables on each node and construct confidence interval of variance, to get link candidate to disconnect.
4. Construct RBFNN model, measure goodness of fit model bases on mean square error (MSE) and Determination coefficient (R^2).

5. DATA ANALYSIS

Indonesian CPI data on based year 2002, January 1999 to October 2006, show non stationer on mean and has linear trend (figure 2.a). This condition, time series ARIMA method cannot be used before data stationer. By using transformation difference, pattern of data become stationer and randomly (figure 2.b). Time series model for the data is ARIMA(0,1,0).

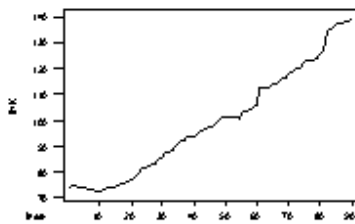


Figure 2.a. Plot CPI



figure 2.b. Plot Detrended CPI

To construct RBFNN model need preprocessing data to eliminate fix component. In this case preprocessing detrended is used, because data has linear trend. Pattern of detrended data is nonlinear. Sutijo and Subanar (2004) show NN model better to approximate nonlinear model than linear model. We hope RBFNN model, one of nonlinear models, give performance better than linear model (ARIMA). Model of detrended data is $y_t = F(y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4})$, because autocorrelation function (ACF) of detrended data show lag 1 to lag 4 significant. There are four clusters or nodes result of clustering data by using K-mean cluster method. Simulation is run to find confidence interval $(1-\alpha)100\%$ of σ^2 . Table 1. Show confidence interval of σ^2 of 1000 times running. Based on table 1, Link y_{t-1} to node 4 or link (1-4) candidates to disconnect, because variance on outside 95% confidence interval. Another links candidate to disconnect is link (2-3), (3-3), (4-1) and (4-3).

Table 1. Confidence Interval of σ^2

Var	Cls	mean	SD	CI 95% SD		Var	Cls	mean	SD	CI 95% SD	
				LL	UL					LL	UL
Y_{t-1} (1)	1	6.702	0.647	0.092	1.727	Y_{t-3} (3)	1	8.795	0.664	0.232	1.729
	2	5.282	1.335	0.310	1.690		2	6.314	0.645	0.469	1.929
	3	3.760	1.120	0.745	1.667		3	2.330	2.850	0.873	1.886
	4	-1.58	1.120	1.390	1.860		4	-1.54	1.418	0.268	1.576
Y_{t-2} (2)	1	7.763	0.854	0.875	1.762	Y_{t-4} (4)	1	9.236	0.020	0.049	1.137
	2	5.761	0.950	0.329	1.671		2	6.835	0.716	0.310	1.895
	3	3.140	2.050	0.788	1.728		3	1.550	2.690	0.992	1.935
	4	-1.57	1.388	1.370	1.609		4	-1.50	1.471	1.272	1.613

RBFFNN design for CPI detrended data is shown on figure 3. Links input variables to node by dash lines are links candidates to disconnect. There are five links candidates to disconnect, so there are 32 models may to construct based on combination that links.

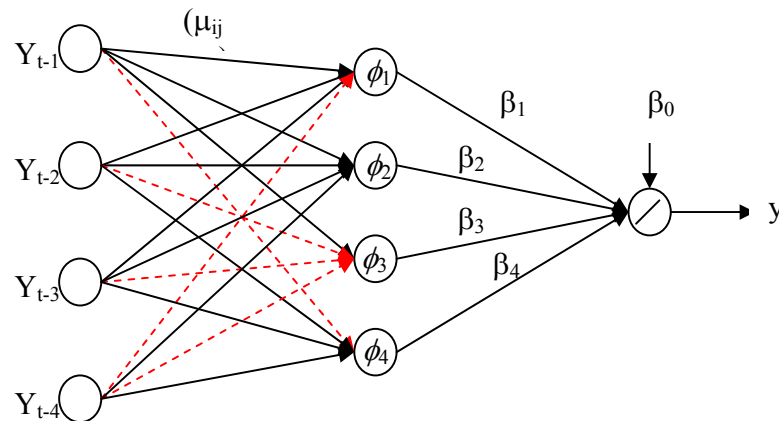


Figure 3. Architecture RBFFNN for Indonesian CPI

Performance models of RBFFNN are based on MSE and R^2 . Models are selected of all model may to construct, complete and uncompleted (one or more link are disconnected) models. Best model of each combination is shown on table 2.

Table 2. Performance RBFFNN model

Model	MSE	R^2 (%)
Full Model	1.304	70.1
(4,1)	1.106	74.6
(3,3) (4,3)	1.064	75.6
(2,3) (3,3) (4,3)	0.790	81.8
(2,3) (3,3) (4,3)	0.790	81.8
(2,3) (3,3) (4,1) (4,3)	0.736	83.1
(1,4) (2,3) (3,3) (4,1) (4,3)	0.749	82.8
Time Series AR(1)	1.289	

Based on table 2, complete model is model without disconnect link. Model (4,1) is best model of disconnect one link (link variable 4 to node 1). Model (3,3)(4,3) is best model of combination two links are disconnected. The best model of Indonesian CPI detrended data is model (2,3)(3,3)(1,4). The performance model is MSE 0.736 and R^2 83.1%. Time series data plot actual and prediction RBFNN is shown on figure 4. Figure 4.a is plot of RBFNN complete model and figure 4.b is plot of RBFNN best model.

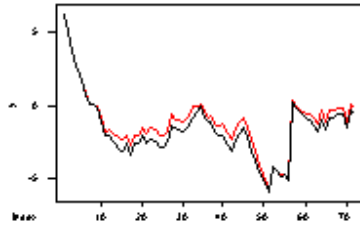
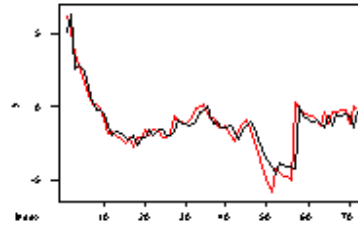


Figure 4.a RBFNN complete model



4.b. RBFNN best model

6. CONCLUSION

Data preprocessing is needed in RBFNN modeling to eliminate fix component and random or nonlinear component remain may RBFNN is applied use full. To increase performance RBFNN model is done by selecting use full link input variables to node in hidden layer unit. Variance component in gaussian activation function is weight of shift of mean, that need confidence interval of variance, small variance can not get information and large variance get more information, include unusable information. Variance input variable in a node on outside confidence interval show link candidate to disconnect.

Indonesian CPI data has linear trend. By using difference transformation in time series method, make stationer data. Time series model of Indonesian CPI is random or ARIMA(0,1,0). Preprocessing data uses detrended technique show that lag 1 to lag 4 of ACF the data are significant, so the model is $y_t = F(y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4})$. Based on clustering uses K-mean method and simulation had done links (1-4), (2-3), (3-3), (4-1), and (4-3) are candidate disconnect links.

Performance RBFNN model is measured by MSE and R^2 . The best RBFNN model of detrended Indonesian CPI data is RBFNN model with links(2-3), (3-3), (4-1) and (4-3) are disconnect from design RBFNN model. The performance is MSE 0.736 and R^2 83.1%.

REFERENCES

1. Bishop, C. M., (1995). *Neural Network for Pattern Recognition*. Oxford: Clarendon Press.
2. Broomhead, D.S. and Lowe, D. (1988). Multivariable functional interpolation and adaptive network. *Complex Systems*, **2**, 321–355.
3. Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, **2**, 304–314.
4. Fine, T. L. (1999). *Feedforward Neural Network Methodology*. Springer, New York.
5. Funahashi, K. (1989). On the approximate realization of continuous mappings by neural networks. *Neural Networks*, **2**, 183–192.
6. Haykin, H. (1999). *Neural Networks: A Comprehensive Foundation*, second edition, Prentice-Hall, Oxford.

Brodjol Sutijo: Statistics Department, Sepuluh Nopember Institute of Technology, Indonesia.

E-mail: sutijo_b@yahoo.com

Subanar: Mathematics Departments, Gadjah Mada University, Indonesia.

E-mail: subanar@yahoo.com

Suryo Guritno: Mathematics Departments, Gadjah Mada University, Indonesia.

E-mail: Guritno0@mailcity.ac.id

ASYMPTOTIC PROPERTIES OF INTEGRATED RENEWAL PROCESSES

SUYONO, J.A.M. VAN DER WEIDE

Abstract. In this paper we derived asymptotic properties of an integrated renewal process. An asymptotic form of the expected value of the process is derived by means of a Tauberian theorem. When the Laplace transform of the expected value is a rational function, we have a finer result of the asymptotic form of the expected value. In addition, under the same condition we may derive an asymptotic variance of the process.

Key words and Phrases: Renewal process, Tauberian theorem, Laplace transform.

1. INTRODUCTION

An integrated renewal process $(Y(t), t \geq 0)$ is a stochastic process defined by taking the integral of the underlying renewal process over the time interval $[0, t]$. It can be considered as a generalization of a shot noise process discussed by Gubner [1]. A special case of this process, where the corresponding renewal process is replaced by a homogeneous Poisson process, can also be found in Ross [3].

The marginal distribution of the integrated renewal process has been discussed by Suyono and van der Weide [4]. In this paper we discuss asymptotic properties of the process. In Section 2 we present the mathematical definition of an integrated renewal process. Finally Section 3 is devoted for the discussion of asymptotic forms of the expected value and the asymptotic variance of the process.

2. INTEGRATED RENEWAL PROCESSES

Let $(X_n, n \geq 1)$ be an independent and identically distributed sequence of non-negative random variables having a common distribution function F . Let

2000 Mathematics Subject Classification: 60G55, 60K05

$(N(t), t \geq 0)$ be the renewal process with inter-arrival times $(X_n, n \geq 1)$, that is

$$N(t) = \sup\{n \geq 0 : S_n \leq t\}$$

where $S_n = \sum_{i=1}^n X_i, n \geq 1$ and $S_0 = 0$.

Define for $t \geq 0$

$$Y(t) = \int_0^t N(s) ds.$$

We call the stochastic process $(Y(t), t \geq 0)$ an *integrated renewal process*. Note that we can express $Y(t)$ as

$$Y(t) = \sum_{i=1}^{N(t)} (t - S_i) = tN(t) - Z(t), \quad (1)$$

where

$$Z(t) = \sum_{i=1}^{N(t)} S_i. \quad (2)$$

Denote by F^* the Laplace-Stieltjes transform of the distribution function F , that is for $\beta \geq 0$,

$$F^*(\beta) = \int_0^\infty e^{-\beta t} dF(t).$$

The marginal distributions of the processes $(Z(t))$ and $(Y(t))$ are expressed in the following theorems, see [4].

Theorem 2.1. *For $Z(t)$ defined in (2)*

(a) *For $\alpha > 0$ and $\beta > 0$*

$$\int_0^\infty \mathbf{E}(e^{-\alpha Z(t)}) e^{-\beta t} dt = \frac{1}{\beta} [1 - F^*(\beta)] \sum_{n=0}^\infty \prod_{i=1}^n F^*(\alpha[n+1-i] + \beta)$$

with the usual convention that the empty product equals 1.

(b) *If $\mathbf{E}[X_1 e^{-\beta X_1}] < \infty$ for some $\beta > 0$, then*

$$\int_0^\infty \mathbf{E}[Z(t)] e^{-\beta t} dt = \frac{\int_0^\infty x e^{-\beta x} dF(x)}{\beta[1 - F^*(\beta)]^2}. \quad (3)$$

(c) *If $\mathbf{E}[X_1^2 e^{-\beta X_1}] < \infty$ for some $\beta > 0$, then*

$$\begin{aligned} \int_0^\infty \mathbf{E}[Z^2(t)] e^{-\beta t} dt &= \frac{[1 + F^*(\beta)] \int_0^\infty x^2 e^{-\beta x} dF(x)}{\beta[1 - F^*(\beta)]^3} \\ &\quad + \frac{2[2 + F^*(\beta)] [\int_0^\infty x e^{-\beta x} dF(x)]^2}{\beta[1 - F^*(\beta)]^4}. \end{aligned} \quad (4)$$

Theorem 2.2. For $Y(t)$ defined in (1)

(a) For $\alpha > 0$ and $\beta > 0$

$$\int_0^\infty \mathbf{E}(e^{-\alpha Y(t)})e^{-\beta t} dt = \sum_{n=0}^\infty \frac{1 - F^*(\alpha n + \beta)}{\alpha n + \beta} \prod_{i=1}^n F^*(\alpha[i-1] + \beta),$$

with the usual convention that the empty product equals 1.

(b) For $\beta > 0$

$$\int_0^\infty \mathbf{E}[Y(t)]e^{-\beta t} dt = \frac{F^*(\beta)}{\beta^2[1 - F^*(\beta)]},$$

(c) If $\mathbf{E}[X_1 e^{-\beta X_1}] < \infty$ for some $\beta > 0$, then

$$\int_0^\infty \mathbf{E}[Y^2(t)]e^{-\beta t} dt = \frac{2F^*(\beta)[1 - F^*(\beta)^2 + \beta \int_0^\infty t e^{-\beta t} dF(t)]}{\beta^3[1 - F^*(\beta)]^3}.$$

In the next section we will derive asymptotic properties of the mean and variance of the processes $(Z(t), t \geq 0)$ and $(Y(t), t \geq 0)$.

3. ASYMPTOTIC PROPERTIES OF THE MEAN AND THE VARIANCE

To derive asymptotic properties of the means of $(Z(t), t \geq 0)$ and $(Y(t), t \geq 0)$ we will use a Tauberian theorem, see Widder [5], stated as follows:

Theorem 3.1. If $\alpha(t)$ is non-decreasing and such that the integral

$$f(s) = \int_0^\infty e^{-st} d\alpha(t)$$

converges for $s > 0$ and if for some non-negative number γ and some constant C

$$f(s) \sim \frac{C}{s^\gamma} \quad \text{as } s \rightarrow 0,$$

then

$$\alpha(t) \sim \frac{Ct^\gamma}{\Gamma(\gamma + 1)} \quad \text{as } t \rightarrow \infty.$$

where Γ denotes the gamma function.

Now we will first consider asymptotic properties of the mean of $Z(t)$.

Theorem 3.2. *If $\mu_1 = \mathbf{E}[X_1] < \infty$ then*

$$\mathbf{E}[Z(t)] \sim \frac{t^2}{2\mu_1} \quad \text{as } t \rightarrow \infty.$$

Proof. Firstly observe that

$$\int_0^\infty e^{-\beta t} d\mathbf{E}[Z(t)] = \lim_{t \rightarrow \infty} \mathbf{E}[Z(t)]e^{-\beta t} + \beta \int_0^\infty \mathbf{E}[Z(t)]e^{-\beta t} dt.$$

Since $0 \leq Z(t) \leq tN(t)$ and $\mathbf{E}[N(t)] = \frac{t}{\mu_1} + o(t)$, see Cox (1962) for example, it follows that

$$\begin{aligned} 0 \leq \lim_{t \rightarrow \infty} \mathbf{E}[Z(t)]e^{-\beta t} &\leq \lim_{t \rightarrow \infty} t\mathbf{E}[N(t)]e^{-\beta t} \\ &= \lim_{t \rightarrow \infty} t \left[\frac{t}{\mu_1} + o(1) \right] e^{-\beta t} \\ &= 0. \end{aligned}$$

This implies

$$\begin{aligned} \int_0^\infty e^{-\beta t} d\mathbf{E}[Z(t)] &= \beta \int_0^\infty \mathbf{E}[Z(t)]e^{-\beta t} dt \\ &= \frac{\int_0^\infty x e^{-\beta x} dF(x)}{[1 - F^*(\beta)]^2}. \end{aligned}$$

where we have used Theorem 2.1 part (b) for the last equation. By dominated convergence it is easy to see that

$$\int_0^\infty x e^{-\beta x} dF(x) = \mu_1 + o(1)$$

and

$$F^*(\beta) = 1 - \mu_1\beta + o(\beta)$$

as $\beta \rightarrow 0$. Hence

$$\int_0^\infty e^{-\beta t} d\mathbf{E}[Z(t)] \sim \frac{1}{\mu_1\beta^2} \quad \text{as } \beta \rightarrow 0.$$

Obviously $\mathbf{E}[Z(t)]$ is non-decreasing. So we can apply Theorem 3.2 (Tauberian theorem) with $\gamma = 2$ to complete the proof.

Next we will derive a stronger version for the asymptotic form of $\mathbf{E}[Z(t)]$. We will assume that the inter-arrival times $(X_n, n \geq 1)$ are continuous random

variables. We also assume that the Laplace transform of $\mathbf{E}[Z(t)]$ given in Theorem 2.1 part (b) is a rational function, i.e.,

$$\frac{\int_0^\infty x e^{-\beta x} dF(x)}{\beta[1 - F^*(\beta)]^2} \quad (5)$$

is a rational function of β . This situation holds for example when X_1 has a gamma distribution with integer *shape* parameter.

Since we assume the Laplace transform of $\mathbf{E}[Z(t)]$ is a rational function, we can split (5) into partial fractions. To do this, firstly observe that $F^*(0) = 1$ and $F^{*'}(0) = -\mu_1 < 0$. So we conclude that the equation

$$1 - F^*(\beta) = 0 \quad (6)$$

has a simple root at $\beta = 0$. Hence the partial fraction expansion of (5) contains terms proportional to $1/\beta^2$ and $1/\beta$. For now, we will consider β as a complex variable and denote its real part by $\Re(\beta)$. Assuming $\mu_i = \mathbf{E}(X_1^i)$, $i = 1, 2, 3$ are finite we can express the Laplace transform of $\mathbf{E}[Z(t)]$ as

$$\int_0^\infty \mathbf{E}[Z(t)] e^{-\beta t} dt = \frac{1}{\mu_1 \beta^3} + \frac{1}{\mu_1^2} \left(\frac{\mu_3}{6} + \frac{\mu_2^2}{4\mu_1} \right) \frac{1}{\beta} + r(\beta). \quad (7)$$

where $r(\beta)$ is a rational function of β with non-zero poles at $\beta_1, \beta_2, \dots$. It follows from (6) that for every j , $\Re(\beta_j) < 0$. Also, since F is continuous, then there can be no purely imaginary roots of (6). There are some other remarks about the roots β_j . If β_j is a simple real root then it gives a term proportional to $1/(\beta - \beta_j)$ in the partial fraction of $r(\beta)$, inverting into $e^{\beta_j t}$. If β_j is a multiple root then it leads to a term proportional to $t^r e^{\beta_j t}$. Since $\Re(\beta_j) < 0$ these terms tend to zero exponentially fast as $t \rightarrow \infty$. Note also that the roots β_j must occur in conjugate pairs. Otherwise $\mathbf{E}[Z(t)]$ contains a term which is a complex number. So from (7) we obtain

$$\mathbf{E}[Z(t)] = \frac{1}{2\mu_1} t^2 + \frac{1}{\mu_1^2} \left(\frac{\mu_3}{6} + \frac{\mu_2^2}{4\mu_1} \right) + o(1) \quad \text{as } t \rightarrow \infty. \quad (8)$$

In (8) the term $o(1)$ tends to 0 as $t \rightarrow \infty$ exponentially fast.

Similar arguments as before can be used to obtain the asymptotic variance of $Z(t)$. Assume that $\mu_4 = \mathbf{E}[X_1^4] < \infty$. If the Laplace transform of $\mathbf{E}[Z^2(t)]$ in (4) is a rational function of β , then we can prove that

$$\int_0^\infty \mathbf{E}[Z^2(t)] e^{-\beta t} dt = \frac{6}{\mu_1^2 \beta^5} + \frac{2(\mu_2 - \mu_1^2)}{\mu_1^3 \beta^4} + \frac{C}{\beta} + r(\beta)$$

where C is a constant depending on $\mu_i, i = 1, 2, 3, 4$ and $r(\beta)$ is a rational function of β having non-zero poles. Inverting this transform we obtain

$$\mathbf{E}[Z^2(t)] = \frac{1}{4\mu_1^2} t^4 + \frac{\mu_2 - \mu_1^2}{3\mu_1^3} t^3 + C + o(1) \quad \text{as } t \rightarrow \infty. \quad (9)$$

It follows, from (8) and (9), that

$$\frac{\text{Var}[Z(t)]}{t^3} \rightarrow \frac{\mu_2 - \mu_1^2}{3\mu_1^3} \quad \text{as } t \rightarrow \infty.$$

For the process $(Y(t))$ we have the following asymptotic properties. The proof is quite similar to that of $(Z(t))$, and is therefore omitted.

Theorem 3.3. *Let $N(t)$ be a renewal process with inter-arrival times (X_n) . Let $\mu_i = \mathbf{E}[X_1^i]$. Let $Y(t) = \int_0^t N(s)ds$. Then*

(a) *If $\mu_1 < \infty$ then*

$$\mathbf{E}[Y(t)] \sim \frac{t^2}{2\mu_1} \quad \text{as } t \rightarrow \infty.$$

(b) *If μ_i , $i = 1, 2, 3$ are finite and if the Laplace transform of $\mathbf{E}[Y(t)]$ stated in Theorem 2.2 part (b) is a rational function of β , then*

$$\mathbf{E}[Y(t)] = \frac{1}{2\mu_1}t^2 + \left(\frac{\mu_2}{2\mu_1^2} - 1\right)t + \left(\frac{\mu_2^2}{4\mu_1^3} - \frac{\mu_3}{6\mu_1^2}\right) + o(1) \quad \text{as } t \rightarrow \infty.$$

(c) *If μ_i , $i = 1, \dots, 5$ are finite and if the Laplace transform of $\mathbf{E}[Y^2(t)]$ stated in Theorem 2.2 part (c) is a rational function of β , then*

$$\begin{aligned} \mathbf{E}[Y^2(t)] = & \frac{1}{4\mu_1^2}t^4 + \left(\frac{5\mu_2 - 2\mu_1^2}{6\mu_1^3}\right)t^3 + \frac{1}{\mu_1^2}\left(1 + \frac{3\mu_2^2}{2\mu_1^2} - \frac{21\mu_2}{2} - \frac{2\mu_3}{3\mu_1}\right)t^2 \\ & + \frac{1}{\mu_1^2}\left(\frac{2\mu_3}{3} - \frac{\mu_2^2}{\mu_1} + \frac{\mu_4}{4\mu_1} - \frac{3\mu_2\mu_3}{2\mu_1^2} + \frac{3\mu_2^3}{2\mu_1^5}\right)t \\ & + C + o(1) \quad \text{as } t \rightarrow \infty, \end{aligned}$$

where C is a constant depending on μ_i , $i = 1, 2, 3, 4, 5$.

(d) *Under the assumptions as in (c),*

$$\frac{\text{Var}[Y(t)]}{t^3} \rightarrow \frac{\mu_2 - \mu_1^2}{3\mu_1^3} \quad \text{as } t \rightarrow \infty.$$

REFERENCES

1. J.A. GUBNER, Computation of shot noise probability distributions and densities, *SIAM Journal on Scientific Computing*, **17** no. 3, 750–761, 1996.

2. D.R. COX, *Renewal Theory*, Spottiswoode Ballantyne, London, 1962.
3. S.M. ROSS, *Stochastic Processes*, John Wiley, New York, 1996.
4. SUYONO AND J.A.M. VAN DER WEIDE, Integrated Renewal Processes, *Accepted for publication at MIHMI*, 2007.
5. D.V. WIDDER, *The Laplace transform*, Princeton University Press, Princeton, 1946.

SUYONO: Department of Mathematics, Universitas Negeri Jakarta, Jakarta, Indonesia.
E-mail: synjkt@yahoo.com.

J.A.M. VAN DER WEIDE : Delft University of Technology, The Netherlands.

MODEL, CALCULATIONS, AND APPLICATIONS OF SPATIAL SCAN STATISTICS

YEKTI WIDYANINGSIH

Abstract. *A common problem in spatial scan statistic is whether a set of points are randomly distributed or if they show signs of clusters or clustering. A spatial scan statistic is a statistical tool to detect location of clusters of interest. This paper shows a model of scan statistic, an algorithm to calculate the value of the test statistic, and application of an algorithm. The data is about the number of mortality of Aids in Indonesia, 2006. Computation of the data was calculated with SaTScan software.*

Keywords: *spatial scan statistic, geography, cluster, likelihood ratio test.*

1. INTRODUCTION

The scan statistics is a statistical method with many potential applications, design to detect a local excess of events and to test if such an excess can reasonably have occurred by chance. The detection can be based on one, two, or three dimension-space. Purely spatial scan statistics concern in two-dimensional space. This paper presents a framework for two-dimensional scan statistics fit. Three basic properties of the scan statistic are the geometry of the area being scanned, the probability distribution generating events under the null hypothesis, and the shapes and sizes of the scanning window. Depending on the application, different models will be chosen, and depending on the model, the test statistics may be evaluated either through explicit mathematical derivations and approximations or through Monte Carlo sampling. In the latter case, random data sets are generated under the null hypothesis, and the scan statistics is calculated in each case, comparing the values from the real and random data sets to obtain a hypothesis test. [1].

2. MODEL

As mentioned above, the three basic properties of the scan statistic are the geometry of the area being scanned, the probability distribution generating events under the null hypothesis, and the shapes and sizes of the scanning window. Kulldorff (1999) defined a general model for the multidimensional scan statistic.

Spatial Scan Statistic Model

Let G be the area in which events may occur, a subset of Euclidean space where different dimensions may represent either physical space or time. For example, G could be particular geographical area during a time period. On G define a measure μ , representing a known underlying intensity that generates events under the null hypothesis. For a homogeneous Poisson process on an area G , we

have $\mu(x) = \lambda$ for all $x \in G$ and some constant λ . The measure could also be discrete, so that it is only positive on a finite number of population points, where $\mu(B)$ is the combine measure of the population points located in area $B \subset G$. We require that $\mu(B) > 0$ for all areas B . Let X denote a spatial point process where $X(B)$ is the random number of events in the set $B \subset G$. Two different probability models are considered, based on Bernoulli counts and the Poisson process, respectively.[2]

For the Bernoulli model, we consider only discrete measures μ such that $\mu(B)$ is an integer for all subsets $B \subset G$. Each unit of measure corresponds to an “entity” or “individual” who could be in either one of two states, for example with or without some disease, or being of a certain species or not. Individuals in one of the states are defined as events, and the locations of those individuals constitute the point process. In the model, there is exactly one zone $A \subset G$ such that each individual within that zone has probability p of being a point, while the probability for individuals outside the zone is q . The probability for any one individual is independent of all the others. Under the null hypothesis, the number of events in any given area is binomially distributed, so that $X(B) \sim \text{Binomial}(\mu(B), p)$ for some value p and for all sets B . The alternative hypothesis H_a : $X(B) \sim \text{Binomial}(\mu(B), p)$ for all sets $B \subset A$ and $X(B) \sim \text{Binomial}(\mu(B), q)$ for all sets $B \subset A^c$, where $p > q$.[3]

For the Poisson model, events are generated by a homogenous or non homogenous Poisson process. Under the null hypothesis, $X(B) \sim \text{Poisson}(p\mu(B))$, for some value p and for all sets $B \subset G$. The alternative hypothesis H_a : $X(B) \sim \text{Poisson}(p\mu(B \cap A), q\mu(B \cap A^c)) \forall B$. The measure μ may either be defined continuously so that events may occur any where, or discretely so that events may occur only at pre specified locations, or as a combination of the two. The discrete case is useful when we are dealing with individual counts or with aggregate data.

The window scan statistic for purely spatial is thought of as an area of fixed size and shape, which then moves across the study area. As it moves, it defines a collection W of zones $W \subset G$. To be more general, we allow for windows of variable size and shape, by defining the window as a collection w of zones $W \subset G$ of any size and shape.

The definition of the scan statistic is the maximum likelihood ratio over all possible zones

$$S_W = \frac{\max_{w \in W} L(W)}{L_0} = \max_{w \in W} \frac{L(W)}{L_0} \quad (1)$$

$L(W)$ is the likelihood function for zone W , expressing how likely the observed data are given a differential rate of events within and outside the zone, and L_0 is the likelihood function under the null hypothesis.

Let $X(G \setminus W) = X(G) - X(W)$ and $\mu(G \setminus W) = \mu(G) - \mu(W)$.

For the Bernoulli model,

$$\frac{L(W)}{L_0} = \frac{\left(\frac{X(W)}{\mu(W)}\right)^{X(W)} \left(1 - \frac{X(W)}{\mu(W)}\right)^{\mu(W) - X(W)} \left(\frac{X(G \setminus W)}{\mu(G \setminus W)}\right)^{X(G \setminus W)} \left(1 - \frac{X(G \setminus W)}{\mu(G \setminus W)}\right)^{\mu(G \setminus W) - X(G \setminus W)}}{\left(\frac{X(G)}{\mu(G)}\right)^{X(G)} \left(1 - \frac{X(G)}{\mu(G)}\right)^{\mu(G) - X(G)}} \quad (2)$$

if $X(W)/\mu(W) > X(G \setminus W)/\mu(G \setminus W)$, and $\frac{L(W)}{L_0} = 1$ otherwise.

For the Poisson model,

$$\begin{aligned} L(W) &= \frac{e^{-p\mu(W)-q\mu(G \setminus W)} [p\mu(W) + q\mu(G \setminus W)]^{X(G)}}{X(G)!} \\ &\quad \times \prod_{r_i \in W} \frac{p\mu(r_i)}{p\mu(W) + q\mu(G \setminus W)} \prod_{r_i \notin W} \frac{q\mu(r_i)}{p\mu(W) + q\mu(G \setminus W)} \\ &= \frac{e^{-p\mu(W)-q\mu(G \setminus W)}}{X(G)!} p^{X(W)} q^{X(G \setminus W)} \prod_{r_i} \mu(r_i) \end{aligned} \quad (3)$$

As before, the likelihood ratio is defined as in equation (1). We have

$$L_0 = \max_p \frac{e^{-p\mu(G)} p^{X(G)}}{X(G)!} \prod_{r_i} \mu(r_i) = \frac{e^{-X(G)} \left(\frac{X(G)}{\mu(G)} \right)^{X(G)}}{X(G)!} \prod_{r_i} \mu(r_i)$$

Equation 3 takes its maximum when $p = X(W)/\mu(W)$ and $q = X(G \setminus W) / \mu(G \setminus W)$ so

$$L(W) = \begin{cases} \frac{e^{-pX(G)} \left(\frac{X(W)}{\mu(W)} \right)^{X(W)} \left(\frac{X(G \setminus W)}{\mu(G \setminus W)} \right)^{X(G \setminus W)}}{X(G)!} \prod_{r_i} \mu(r_i) & \text{if } \frac{X(W)}{\mu(W)} > \frac{X(G \setminus W)}{\mu(G \setminus W)} \\ \frac{e^{-X(G)} \left(\frac{X(G)}{\mu(G)} \right)^{X(G)}}{X(G)!} \prod_{r_i} \mu(r_i) & \text{otherwise} \end{cases}$$

The test statistic of the likelihood ratio test can now be written as

$$\frac{L(W)}{L_0} = \frac{\left(\frac{X(W)}{\mu(W)} \right)^{X(W)} \left(\frac{X(G \setminus W)}{\mu(G \setminus W)} \right)^{X(G \setminus W)}}{\left(\frac{X(G)}{\mu(G)} \right)^{X(G)}} \quad (4)$$

if $X(W)/\mu(W) > X(G \setminus W)/\mu(G \setminus W)$, and $\frac{L(W)}{L_0} = 1$ otherwise. The expression

$X(W) / \mu(W) > X(G \setminus W) / \mu(G \setminus W)$ simply states that there are more than the expected number of events within the window as compared to outside the window.[2]

3. CALCULATIONS

Monte Carlo-Based Hypothesis Testing

The mathematics for obtaining the distribution of the scan statistic is quite complex, and exact derivations have proved elusive for all but the simplest scenarios. There are some very interesting and impressive probabilistic approximations though. Details of these developments can be found in other references [2]. When probabilistic approximations are not available, Monte Carlo-based hypothesis can be applied to any special case of the general model presented above. Monte Carlo-based hypothesis testing was proposed by Dwass (1956), who pointed out that the probability of falsely rejecting the null hypothesis is exactly according to the significance level, in spite of the simulation involved. The Four-step procedure of Monte Carlo hypothesis testing for a scan statistic:

1. Calculate the value of the test statistic for the real data.
2. Create a large number of random data sets generated under the null hypothesis.
3. Calculate the value of the test statistic for each of the random replications.
4. Sort the values of the test statistic, from the real and random data sets, and note the rank of the one calculated from the real data set. If it is ranked in the highest α percent, then reject the null hypothesis at α percent significance level.

The key in terms of minimizing computing time is Step 3, as it can be complex in nature, and most of all, because it must be repeated once for each random replication of the data set. Some algorithms are built for circular window process. The following is an algorithm for discrete non-homogeneous Bernoulli or Poisson process. [1]

Scanning algorithm (circular window, variable size, circle centroids on grid).

1. *Pick a grid point. Calculate the distance to the different population points and sort those in increasing order. Memorize the sorted population points in an array.*
2. *Repeat Step 1 for each grid point.*
3. *Pick a grid point*
4. *Create a circle centered at the grid point and continuously increase the radius. For each population point entering the circle, update the number of events n and the measure $\mu(W)$ inside the circular area W .*
5. *Repeat Steps 3 and 4 for each grid point. Report the largest likelihood based on all $(n, \mu(W))$ -pairs as the scan statistic, where the likelihood is calculated according to equation (2) or (3).*
6. *Repeat Step 3 to 5 for each Monte Carlo replication.*[1]

This algorithm can work for three-dimensional spherical windows by simply defining the population points in three-dimensional space. The complexity remains the same but for a complete coverage, the number of grid points G may have to be larger.[1]

4. APPLICATION

There is a long history of geographical surveillance of disease by publishing atlases. If there are areas with exceptionally high rates, they may give us clues to the etiology of the disease, it may indicate areas where health care needs improvement, or it may indicate areas to be targeted for preventive measures. In those atlases that are not purely descriptive, analysis is often done by dividing the study region into non overlapping districts, making of separate test each district to see if it has an excess incidence or mortality. With a spatial scan statistic, we do the surveillance adjusting for multiplicity of possible cluster locations.

Example: AIDS mortality in Indonesia

HIV AIDS disease is continuing to increase in number, and spreading through the regions. Recently, according to the Jakarta Post news paper, the number of AIDS cases was found in Jogjakarta and more number in Papua. Through the scan statistics calculation, we analyzed the cluster regions of the AIDS mortality cases in year 2002 and 2006[7]. This application compared the two results. Based on the number of HIV AIDS cases, assumed that the mortality number as Bernoulli distribution with possibilities of die or not die. The equation (2) is used as the model of statistical test. The data was calculated with SaTScan software.[5] The results are as followed.

5. RESULTS

Table 1 and 2 shows the result of SatScan software for the spatial scan statistics. Figure 1 and 2 are the atlases of the results.

Table 1. Cluster Areas of Aids Mortality
The Application of Spatial Scan Statistic Calculation:
Indonesia, 2002

Hotspot Cluster	Number of Mortality Cases	Number of Expected	Relative Risk	Log Likelihood Ratio	p-value
<i>Most Likely Cluster:</i> MALUKU, SULTENG, PAPUA, SULUT, SULSEL, NTT, KALTIM	155	131.32	1.305	5.175177	0.020
<i>Secondary cluster:</i> DIY, JATENG	12	7.11	1.711	2.625381	0.401
<i>Secondary cluster:</i> SUMSEL, JAMBI, RIAU	28	20.20	1.417	2.454825	0.425

In 2002, the regions of the most likely cluster of aids mortality were Maluku, centre of Sulawesi, Papua, North of Sulawesi, South Sulawesi, Nusa Tenggara Timur, and East Kalimantan. This cluster is statistically significant with relative risk = 1.305, LLR = 5.175177, and p-value= 0.020.

Table 2. Cluster Areas of Aids Mortality
The Application of Spatial Scan Statistic Calculation:
Indonesia, 2006

Hotspot Cluster	Number of Mortality Cases	Number of Expected	Relative Risk	Log Likelihood Ratio	p-value
<i>Most Likely Cluster:</i> NTT, NTB, SULSEL, SULTRA, BALI, SULBAR, SULTENG, KALSEL, JATIM, GRTALO, DIY, KALTENG, MALUKU, SULUT, KALTIM, JATENG, MALUT	599	437.86	1.578	50.587215	0.001
<i>Secondary cluster:</i> SUMBAR, RIAU, KEPRI	150	77.03	2.042	41.156584	0.001
<i>Secondary cluster:</i> LAMPUNG	31	22.68	1.374	1.884332	0.767

In 2006, the regions of the most likely cluster of aids mortality were Nusa Tenggara Timur, Nusa Tenggara Barat, South Sulawesi, South East Sulawesi, Bali, West Sulawesi, Center of Sulawesi, South Kalimantan, East Java, Gorontalo, Yogyakarta, Center of Kalimantan, Maluku, North Sulawesi, East Kalimantan, Center of Java, and North Maluku. This cluster is statistically significant with relative risk = 1.578, LLR = 50.587, and p-value= 0.001.

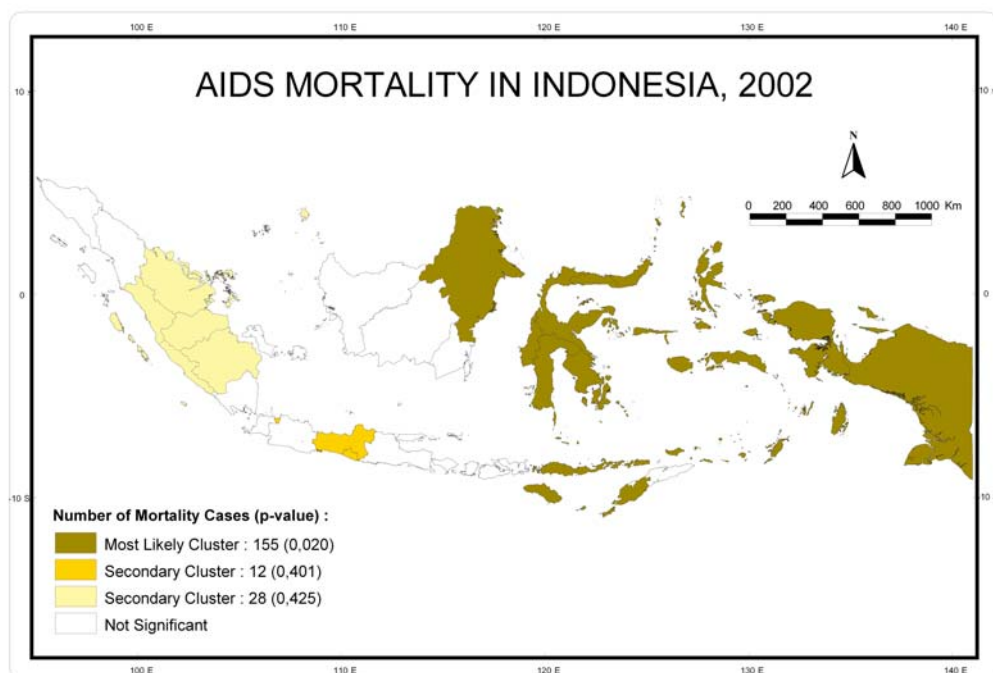


Figure 1. Aids Mortality in Indonesia, 2002

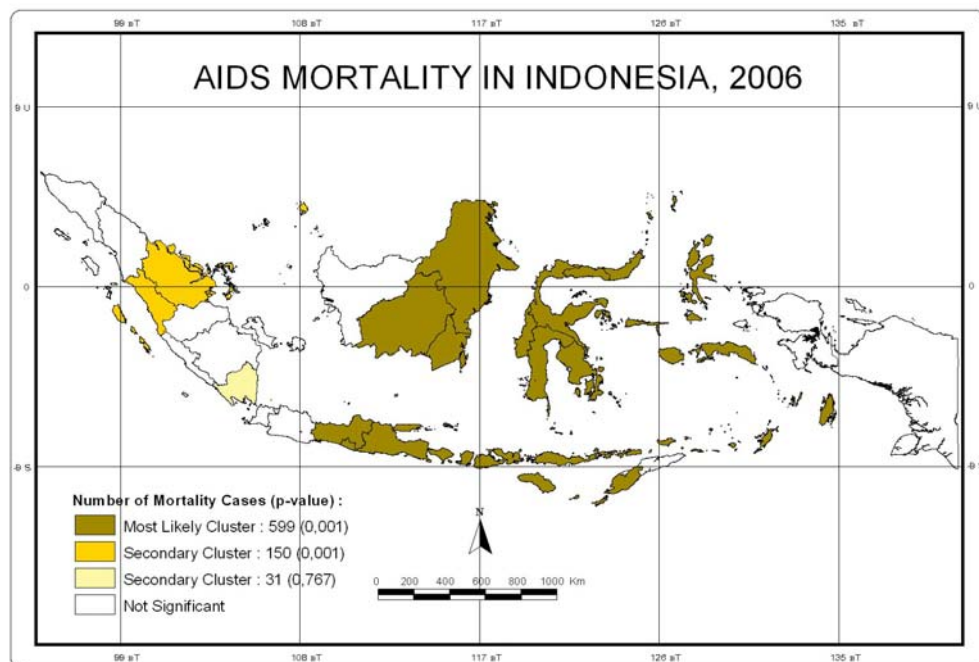


Figure 2. Aids Mortality in Indonesia, 2006

6. DISCUSSION AND CONCLUSION

According to the result of spatial scan statistic calculation for Aids mortality in Indonesia, there was a moving hotspot areas of the mortality number from year 2002 to 2006. In 2002, aids mortality hotspot is around the east part of Indonesia. From 2002 to 2006, AIDS mortality hotspot was move from the east to the central part of Indonesia, include Central Java and Yogyakarta. The news about rising AIDS cases in Yogyakarta in 2007 is relevant with the analysis results. The second highest was in the Central part of Sumatra, and the third (not significant) was in the south part of Sumatra. In this paper, the using of spatial scan statistic calculation was apply for two period of times, that is the year 2002 and 2006, each with the purely spatial. The analysis was applied for each year data. It will be better if the application was calculated for spatial and time simultaneously, with addition periods of time. The method is space-time scan statistics.

By Model, calculations, and application of spatial scan statistics method, the results of the study are about the information of hotspot (cluster) regions. As long as the purely spatial scan statistic was used for the data, the result is about the locations of the hotspot information. Spatial scan statistics compare many regions simultaneously and find the Most Likely Cluster by scanning all parts of the study region.

REFERENCES

1. Kulldorff M, 1999. Spatial Scan Statistics: Models, Calculations, and Application. National Cancer Institute, Bethesda, MD
2. Joseph G. and Balakrishnan N. 1999. Scan Statistics and Applications, Birkhauser.
3. Kulldorff M. 1997. A Spatial Scan Statistic. *Communication in Statistics: Theory and Methods*, 26:1481—1496.
4. Dwass M. 1956. Modified Randomization Test for Nonparametric Hypothesis, Northwestern University.
5. Kulldorff M. 2006. *SatScan User Guide version 6.1*.
6. NN, HIV/AIDS Cases on the Rise in Yogya. *The Jakarta Post*, June 5, 2007.
7. Departemen Kesehatan Republik Indonesia. 2006. Ditjen PP & PL. Jakarta

Yekti Widyaningsih: Department of Mathematics, University of Indonesia, Depok, Indonesia.

E-mail: yekti@ui.edu