Lecture Notes in Electrical Engineering 275

Gi-Chul Yang
Sio-Iong Ao
Xu Huang
Oscar Castillo
Editors

Transactions on Engineering Technologies

International MultiConference of Engineers and Computer Scientists 2013

Springer

Lecture Notes in Electrical Engineering

Volume 275

For further volumes:
http://www.springer.com/series/7818

Gi-Chul Yang · Sio-Iong Ao
Xu Huang · Oscar Castillo
Editors

Transactions on Engineering Technologies

International MultiConference of Engineers and Computer Scientists 2013

Springer

Editors
Gi-Chul Yang
Department of Multimedia Engineering
College of Engineering
Mokpo National University
Chonnam
Korea, Republic of Korea

Xu Huang
Faculty of Information Sciences and
Engineering
University of Canberra
Canberra, ACT
Australia

Sio-Iong Ao
International Association of Engineers
Hong Kong SAR

Oscar Castillo
Computer Science in the Graduate Division
Tijuana Institute of Technology
Tijuana, BC
Mexico

ISSN 1876-1100 ISSN 1876-1119 (electronic)
ISBN 978-94-007-7683-8 ISBN 978-94-007-7684-5 (eBook)
DOI 10.1007/978-94-007-7684-5
Springer Dordrecht Heidelberg New York London

Library of Congress Control Number: 2013953195

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

A large international conference on Advances in Engineering Technologies and Physical Science was held in Hong Kong, 13–15 March, 2013, under the "International MultiConference of Engineers and Computer Scientists 2013 (IMECS 2013)". The IMECS 2013 is organized by the International Association of Engineers (IAENG). IAENG is a non-profit international association for the engineers and the computer scientists, which was founded originally in 1968 and has been undergoing rapid expansions in recent few years. The IMECS congress serves as good platforms for the engineering community to meet with each other and to exchange ideas. The congress has also struck a balance between theoretical and application development. The conference committees have been formed with over 300 committee members who are mainly research center heads, faculty deans, department heads, professors, and research scientists from over 30 countries with the full committee list available at our congress web site (http://www.iaeng.org/IMECS2013/committee.html). The congress is truly international meeting with a high level of participation from many countries. The response that we have received for the congress is excellent. There have been more than 600 manuscript submissions for the IMECS 2013. All submitted papers have gone through the peer review process and the overall acceptance rate is 50.97 %.

This volume contains 30 revised and extended research articles written by prominent researchers participating in the conference. Topics covered include engineering physics, engineering mathematics, scientific computing, control theory, automation, artificial intelligence, electrical engineering, and industrial applications. The book offers the state-of-art of tremendous advances in engineering technologies and physical science and applications, and also serves as an excellent reference work for researchers and graduate students working with/on engineering technologies and physical science and applications.

Gi-Chul Yang
Sio-Iong Ao
Xu Huang
Oscar Castillo

Contents

Statistics of Critical Avalanches in Vertical Nanopillar Arrays

Zbigniew Domański, Tomasz Derda and Norbert Sczygiol

Abstract Nanopillar arrays are encountered in numerous areas of nanotechnology such as bio-medical and chemical sensing, nanoscale electronics, photovoltaics or thermoelectrics. Especially arrays of nanopillars subjected to uniaxial microcompression reveal the potential applicability of nanopillars as components for the fabrication of electro-mechanical sense devices. Thus, it is worth to analyze the failure progress in such systems of pillars. Under the growing load pillars destruction forms an avalanche and when the load exceeds a certain critical value the avalanche becomes self-sustained until the system is completely destroyed. In this work we have explored the distributions of such catastrophic avalanches appearing in overloaded systems. Specifically, we analyze the relations between the size of an avalanche being the numbers of instantaneously crushed pillars and the size of the corresponding array of nanopillars using different load transfer protocols.

Keywords Avalanche · Array of pillars · Fracture · Load transfer · Probability distribution function · Scaling

Z. Domański (✉)
Institute of Mathematics, Czestochowa University of Technology, Dabrowskiego 69, PL-42201 Czestochowa, Poland
e-mail: zbigniew.domanski@im.pcz.pl

T. Derda
Faculty of Mechanical Engineering and Computer Science, Czestochowa University of Technology, Dabrowskiego 69, PL-42201 Czestochowa, Poland
e-mail: tomasz.derda@im.pcz.pl

N. Sczygiol
Institute of Computer and Information Sciences, Czestochowa University of Technology, Dabrowskiego 69, PL-42201 Czestochowa, Poland
e-mail: norbert.sczygiol@icis.pcz.pl

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 275, DOI: 10.1007/978-94-007-7684-5_1,

1 Introduction

During the last decade, rapid progress has been achieved in creation and development of new sub-micron scale devices. Among them bundles of nanopillars have attracted much attention, especially these assembled in an ordered fashion on flat substrates under the form of vertical pillar arrays (VPA) [1]. Such arrangement is applied in systems of micromechanical sensors. Our work is motivated by uniaxial tensile and compressive experiments on nano- and microscale metallic pillars that confirm substantial strength increase via the size reduction of the sample [2]. Especially the studies on arrays of free-standing nanopillars subjected to uniaxial microcompression reveal the potential applicability of nanopillars as components for the fabrication of micro- and nano-electromechanical systems, micro-actuators or optoelectronic devices [3, 4].

In a technological application, if a VPA is subjected to an external load, it begins to fracture immediately when the internal stress intensity equals or exceeds the critical value of weakest pillars and the failure develops in a form of avalanches of simultaneously fractured elements.

Avalanches are phenomena on different length scales encountered in an ample set of complex systems. Examples involve magnetic avalanches progressed through tiny crystals, mass movements of geological materials forming rock, sand or snow avalanches as well as fires destroying huge forests. Their presence is not limited to natural science or technology. Avalanches are reported, e.g. in the world of economy where the stock market crashes evolve as an explosive instability. Such instabilities are commonly present in sand and snow avalanches, earthquakes, nuclear chain reactions as well as in damage evolutions of mechanical systems [5, 6]. They appear when a small increase in the external load excludes an element from the working community in such a way that this exclusion alters the internal load pattern sufficiently to trigger the rupture of the other elements and, in consequence, provoking a wave of destruction. For this reason it is worth analyzing the evolution of mechanical destruction within an array of nano-sized pillars.

We simulate a failure by accumulation of pillars crushed under the influence of an axial load. The stepwise increasing load causes the progressive damage of the system in an avalanche-like manner. When the load on a pillar exceeds the threshold the pillar crashes and its load has to be redistributed among the other pillars and carried by them. In this context, an important issue concerns the so-called load sharing rules because the behaviour of a system depends on them. Typically, the avalanche statistical characteristics are immersed in the distribution $D(\Delta)$ of burst size Δ being the number of events triggered by the single failure. We focus our analysis on different load sharing protocols with respect to the range of the load transfer. More specifically, we employ the short-range (local load), long-range (global load) and variable-range protocols. Within their framework we compute $D(\Delta)$ for an ample set of system sizes and report the results of statistical analysis of the system destruction.

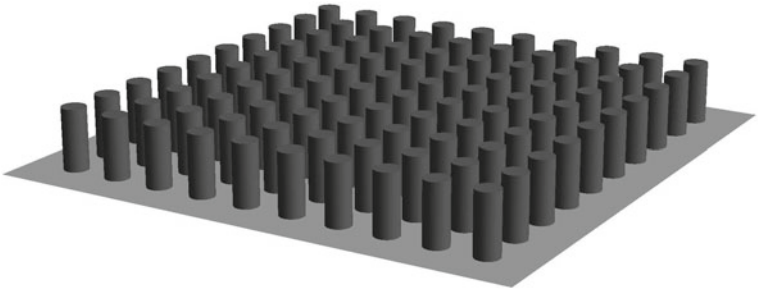


Fig. 1 Schematic view of an array of $N = 10 \times 10$ nano-sized pillars

2 Mathematical Model

Ruptured parts of mechanisms are encountered in virtually all kind of devices. They cause machine malfunction and are dangerous for users. Thus, the knowledge of the fracture evolution as well as its effective description represent an important issue in materials science and technology. Our example system is an array of nanopillars [1, 4]. A schematic view of such array is presented in Fig. 1.

2.1 Load Transfer Rules

The pillars are treated as fibres in the framework of a Fibre Bundle Model (FBM) [6]-[16]. FBM is a load transfer model. In a static FBM, a set of N pillars is located in the nodes of the supporting lattice. Due to various defects during fabrication the pillar-strength-thresholds are quenched random variables. Here, for our numerical analysis purpose they are governed by independent probability uniformly distributed on the interval $[0, 1]$. The set of elements is subjected to an external load that is increased quasi-statically. After a pillar breakdown, its load is transferred to the other intact elements and, as a consequence, the probability of subsequent failure increases. Among several load transfer rules there are two extreme schemes: global load sharing (GLS) the load is equally redistributed to all the remaining elements and local load sharing (LLS) the load is transferred only to the neighboring elements [6]. The GLS model being a mean-field approximation with long-range interactions among the elements can be solved analytically. In the case of the LLS rule the distribution of load is not homogeneous and regions of stress accumulation appear throughout the system. This gives severe problems for analytical treatment and one has to answer the questions by means of numerical simulations. Load redistribution in free-standing pillars should be placed somewhere in between the LLS and the GLS rules. For this reason we employ an approach based on Voronoi polygons which merges the GLS and LLS rules. The extra load is equally redistributed among the elements lying inside the Voronoi regions generated by a group of elements destroyed within an

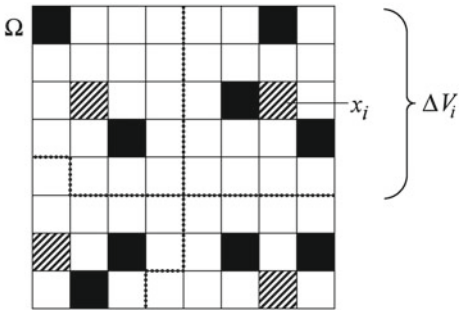


Fig. 2 The Voronoi polygons for a set of square-shaped pillars: *white squares*-intact pillars, *black squares*-previously destroyed pillars and *shaded squares*-just damaged pillars

interval of time taken to be the time step. We call this load transfer rule Voronoi load sharing (VLS) [17].

Voronoi polygons are one of the most fundamental and useful constructs defined by irregular lattices [18]. For set $\mathbf{X} = \{x_1, x_2, \ldots, x_N\}$ of N distinct points in $\Omega \subset \mathbf{R}^2$, the Voronoi tessellation is the partition of Ω into N polygons denoted by ΔV_i. Each ΔV_i is defined as the set of points which are closer to x_i than to any other points in $\mathbf{X}$. In Fig. 2, an example of Voronoi polygons is shown in the case of square array of pillars under the load.

All of the Voronoi regions are convex polygons. Each polygon is defined by the lines that bisect the lines between the central point and its neighboring points. The bisecting lines and the connection lines are perpendicular to each other. Using this rule for every point in the area yields this area completely covered by the adjacent polygons representing pillars. Numbers of intact elements inside of Voronoi regions vary randomly and this is the source of supplemental stochasticity in the model. It is worth mentioning that an approach based on the Voronoi tessellation was used for the failure analysis of quasi-brittle materials and fiber-reinforced brittle-matrix composites [19].

2.2 Loading of the System

At the beginning of the damage process all the pillars are intact. Then the system is loaded in a quasi-static way by a longitudinal external force F gently growing from its initial value $F = 0$. More precisely, the system is uniformly loaded until the weakest intact pillar fails and then the increase of load stops. After this failure the load dropped by the damaged pillar has to be redistributed to the intact pillars according to a given transfer rule. The increased stress on the intact pillars may give rise to other failures, after which the load transfer from the destroyed elements may cause subsequent failures. If the load transfer does not trigger further failures there is a stable configuration and external load F has to be increased with a small amount, just to provoke damage of the subsequent weakest intact pillar. By that

means a single failure induced by the load increment can cause an entire avalanche of failures. The above described procedure is repeated until the system completely fails. In the quasi-static approach the force F is the control parameter of the model.

3 Statistics of Catastrophic Avalanches

During the loading process, cascades of simultaneous crashes of several pillars appear. These ruptures resemble avalanches occurring in snow or sands movement. Hence, we consider the avalanche of size Δ being the number of damaged pillars under an equal external load and the distribution $D(\Delta)$ of the magnitude of such crushed-pillars avalanches is the main characteristics in our work.

The problem we consider is the distribution of Δ. It is known that under the GLS rule $D(\Delta)$ can be expressed in a power law form

$$D(\Delta) \sim \Delta^{-\tau}, \ \tau = 5/2. \tag{1}$$

The mean filed exponent $\tau = 5/2$ is presumably independent of disorder distribution with the only exception for distributions which allow unbreakable elements [20] in the system. This $5/2$ power law is valid when all avalanches are considered, i.e. it is a global exponent. If we trace *sub*-critical avalanches, i.e. the avalanches close to the critical breakdown of the system a crossover from $5/2$ to $3/2$ emerges in Eq. (1) [9].

3.1 Catastrophic Avalanches Within the GLS and the VLS Rules

We start our analysis with avalanches which propagate themselves through the system when the load is transferred according to the GLS rule. The case of the LLS rule we discuss in the following subsection.

We pay special attention to the critical avalanche which sustains fracture development among all the pillars. This final, from the system's integrity point of view, avalanche has its own dynamics related to the stress redistribution. It can be viewed as a cascade of sub-avalanches, also called inclusive avalanches [21]. Here, such an avalanche is the number of crashed pillars per step of internal stress redistribution.

The $\tau = 3/2$ scaling related to the distribution of *sub*-critical avalanches also emerges if we analyze the inclusive avalanches inside a critical avalanche and this is true when the load is transferred according to the GLS rule. It is seen in Fig. 3, which shows the distribution of the size Δ_{inc} of inclusive avalanche developed in the array of pillars with the GLS rule. The results are averaged over 10^5 independent samples.

We have performed simulations on arrays of different sizes, i.e. with the number of pillars N ranged from 5×5 up to 250×250. For $N \leq 100 \times 100$ we have carried out at least 10^4 simulations for each array's size, whereas for the other sizes

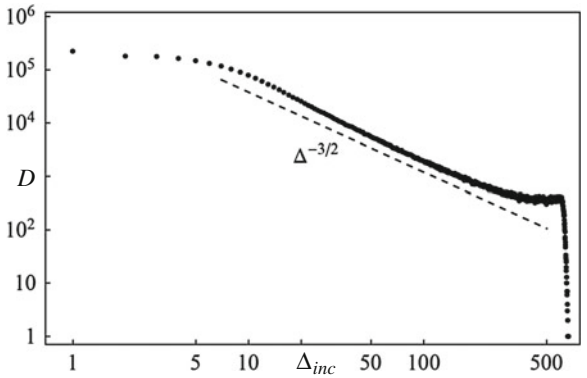


Fig. 3 The inclusive avalanche size distribution for 50 × 50 pillars loaded quasi statically. The GLS rule was applied. The *dashed line* represents the power law obtained analytically [9]

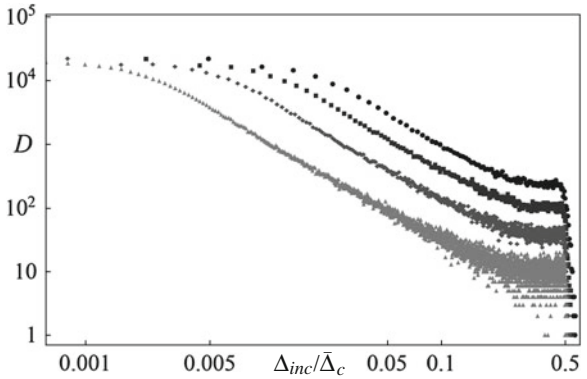


Fig. 4 Scaled inclusive avalanche size distribution for the system with the GLS rule. Different system sizes are compared; from the *top* to *bottom*: the array with 20 × 20, 30 × 30, 50 × 50 and 100 × 100 pillars. The scaling parameter is the mean size $\overline{\Delta}_c$ of the appropriate catastrophic avalanche

the number of simulations was reduced below 10^3 simulations per system's size, mainly due to the long execution time of the algorithm. Based on the results of these simulations we have computed corresponding statistics characterizing avalanches along with the values of such empirical estimators as, e.g. the mean values and the standard deviations. The outcome distributions are presented in Figs. 4 and 5, for the GLS and the VLS rules, respectively. In both figures these distributions involve the sizes Δ_{inc} of inclusive avalanches scaled by the mean size $\overline{\Delta}_c$ of the appropriate catastrophic avalanche. Figure 4 clearly displays the $\tau = 3/2$ scaling (1) in the middle range of $\Delta_{inc}/\overline{\Delta}_c$ values as it was already seen in the Fig. 3. However, in the case of the VLS rule we observe a strong departure of the distribution from the power law form [22].

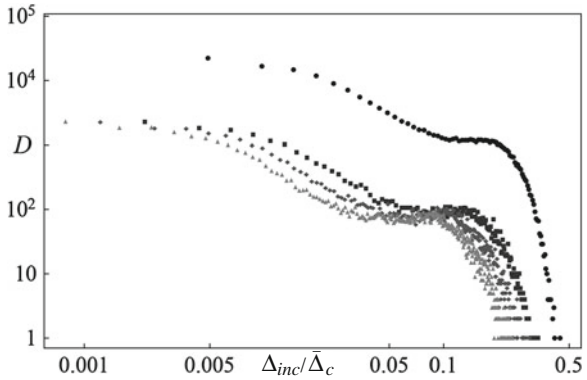


Fig. 5 Scaled inclusive avalanche size distribution for the system with the VLS rule. Different system sizes are compared; from the *top* to *bottom*: the array with 20×20, 30×30, 40×40 and 50×50 pillars. The scaling parameter $\overline{\Delta}_c$ is the mean size of the appropriate catastrophic avalanche

3.2 Catastrophic Avalanches Within the LLS Rule

Limited-range load transfer rules, including the LLS rule, seem to be applicable to an array with pillars fabricated on a nonrigid substrate. It is because of the stress concentrations appearance around crushed pillars if a substrate has certain compliance. Under the LLS rule $D(\Delta)$ can also be expressed in a power law form (1) but with an exponent $\tau = 9/2$ [12, 16]. This value of τ is substantially greater than $\tau = 5/2$ in the case of the GLS rule. To look closer at how the LLS rule influences avalanches development we have performed a series of simulations related to systems with different numbers of pillars. We have varied the array's size for about one decade (from 5×5 to 200×200) and have averaged over at least 10^4 runs. We have also simulated compressions of arrays with number of pillars $(200 \times 200) < N \leq (400 \times 400)$, but with significantly smaller number of runs (typically less than 10^3). Then we have fitted the resulting empirical probability density functions $p_N(\Delta_c)$. Figure 6 displays one of such pdf.

Based on these pdfs we have computed the mean values $\overline{\Delta}_c$ and the standard deviations σ for $N \leq 200 \times 200$. It is interesting to see that these quantities can be approximated by scaling relations. It turns out that $\overline{\Delta}_c$ can be cast into the following equations:

$$\overline{\Delta}_c/N = 2.5876\,(\ln N)^{-2} - 1.75\,(\ln N)^{-1} + 0.9463 \tag{2}$$

or using $N = L \times L$ to replace N by L, Eq. (2) can be written as

$$L^{-2}\overline{\Delta}_c = 0.6469\,(\ln L)^{-2} - 0.875\,(\ln L)^{-1} + 0.9463 \tag{3}$$

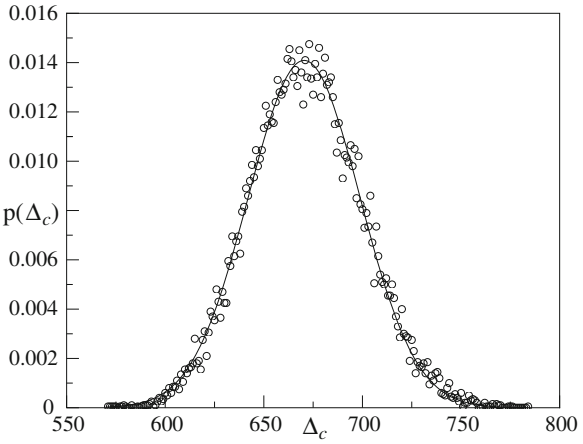


Fig. 6 The probability density function of the length Δ_c of a critical avalanche in an array with 30×30 nanopillars obtained from 2×10^4 samples. Here the LLS rule was applied and the solid line represents normally distributed Δ_c with the mean and the standard deviation computed from these samples

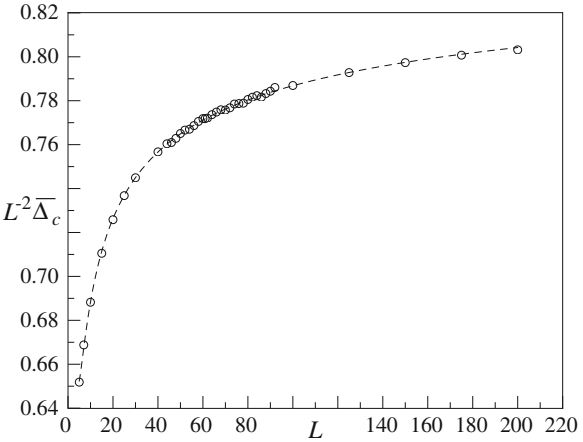


Fig. 7 The mean size $\overline{\Delta}_c$ of the critical avalanche scaled by the number of nanopillars $N = L \times L$ versus the linear array's size L. The results are obtained from at least 5×10^3 samples for each value of N with the LLS rule. The *dashed line* is drawn according to the Eq. (3)

The relation (3) is presented in Fig. 7.

Similarly we have fitted values of σ by a function $\tilde{\sigma}$ which follows the relation:

$$\widetilde{\sigma} = 0.3702N^{1/2} + 0.045N^{7/8} \tag{4}$$

or equivalently

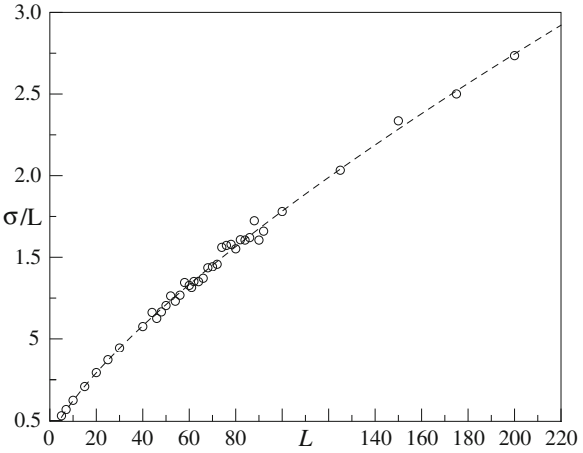


Fig. 8 Scaled standard deviation σ/L of the size Δ_c as a function of L. The *dashed line* is drawn according to the Eq. (5)

$$L^{-1}\widetilde{\sigma} = 0.3702 + 0.045L^{3/4} \tag{5}$$

The computed standard deviations σ and the scaling (5) are shown in Fig. 8 and the relative error $(\widetilde{\sigma}/\sigma - 1)$ of this approximation lies within the interval $(-0.04, 0.02)$.

For arrays with numbers of pillars $N \gg 1$, Eqs. (2) and (4) allow us to cast the coefficient of variation $c_v = \widetilde{\sigma}/\overline{\Delta}_c$ into the scaling $c_v \sim N^{-1/8}$. Since c_v decreases with growing N then the variability in $p_N(\Delta_c) \to 0$.

4 Summary

In this paper, we have analyzed the statistics of avalanches during the failure process in longitudinally loaded arrays of nano-sized pillars with statistically distributed thresholds for breakdown of an individual pillar. We have paid special attention to the inclusive avalanches, i.e. to the avalanches which belong to the final avalanche among all the intact pillars in the system. The sizes Δ_{inc} of these avalanches were computed in series of simulations conducted under different load transfer protocols: the GLS, the LLS and the VLS protocol formulated around the Voronoi segmentation concept.

When the GLS protocol is applied, the resulting distribution of Δ_{inc} displays the $\tau = 3/2$ scaling signature (1). The VLS scenario yields the distribution $D(\Delta_{inc})$ without a trace of $\Delta^{-\tau}$ scaling.

Within the LLS scenario we have numerically linked the mean critical avalanche size with the number of pillars: $\overline{\Delta}_c \sim N(1 - a/\ln N + b/\ln^2 N)$, as well as the corresponding standard deviation: $\sigma \sim N^{7/8}(1 + cN^{-1/2})$. We see that apart from

the leading term N in $\overline{\Delta}_c$ there are some logarithmic corrections which slow down the growth of $\overline{\Delta}_c$ with N. Similarly, the leading term $N^{7/8} < N$ means that the standard deviation σ increases slower than the system size. Hence, the coefficient of variation scales with N as $c_v \sim N^{-1/8}$ and this scaling indicates that the variability in $p_N(\Delta_c)$ vanishes slowly when N grows.

References

1. Chekurov N, Grigoras K, Peltonen A, Franssila S, Tittonen I (2009) The fabrication silicon nanostructures by local gallium implantation and cryogenic deep reactive ion etching. Nanotechnology 20:65307
2. Jang D, Greer JR (2010) Transition from a strong-yet-brittle to a stronger-and-ductile state by size reduction of metallic glasses. Nat Matter 9:215–219
3. Greer JR, Jang D, Kim J-Y, Burek J (2009) Emergence of New Mechanical Functionality in Material via Size Reduction. Adv Funct Mater 19:2880–2886
4. Sievila P, Chekurov N, Tittonen I (2010) The fabrication of silicon nanostructures by focused-ion-beam implantation and TMAH wet etching. Nanotechnology 21(14):145301
5. Chakrabarti B, Benguigui LG (1997) Statistical physics of fracture and breakdown in disordered systems. Claredon Press, Oxford
6. Alava MJ, Nukala PKVV, Zapperi S (2006) Statistical models of fracture. Adv Phys 55:349–476
7. Delaplace A, Roux S, Pijaudier-Cabot G (2001) Avalanche statistics of interface crack propagation in fibre bundle model: characterization of cohesive crack. J Eng Mech 127:646–652
8. Mahesh S, Leigh Phoenix S, Beyerlein IJ (2002) Strength distributions and size effects for 2D and 3D composites with Weibull fibers in an elastic matrix. Int J Fract 115:41–85
9. Pradhan S, Hansen A, Hemmer PC (2005) Crossover behavior in burst avalanches: signature of imminent failure. Phys Rev Lett 95(12):125501
10. Pradhan S, Hansen A, Chakrabarti BK (2010) Failure processes in elastic fiber bundles. Rev Mod Phys 82:499–555
11. Hemmer PC, Hansen A (1992) The distribution of simultaneous fiber failures in fiber bundles. J Appl Mech 59:909–914
12. Kloster M, Hansen A, Hemmer PC (1997) Burst avalanches in solvable models of fibrous materials. Phys Rev E 56:2615–2625
13. Hidalgo RC, Kun F, Herrmann HJ (2001) Bursts in fiber bundle model with continuous damage. Phys Rev E 64(6):066122
14. Pradhan S (2011) Can we predict the failure point of a loaded composite material? Comp Phys Commun 182:1984–1988
15. Hidalgo RC, Moreno Y, Kun F, Herrmann HJ (2002) Fracture model with variable range of interaction. Phys Rev E 65(4):046184
16. Raischel F, Kun F, Herrmann HJ (2006) Local load sharing fiber bundless with a lower cutoff of strength disorder. Phys Rev E 74(3):035104 (R)
17. Domański Z (2011) Geometry-induced transport properties of two dimensional networks. In: Schmidt M (ed) Advances in computer science and engineering. InTech, Rijeka, pp 337–352
18. Ocabe A, Hoots B, Sugihara K, Chiu NS, Kendeall DG (2008) Spatial tessellations: concepts and applications of Voronoi diagrams. John Wiley & Sons, Hoboken, NJ
19. Bolander JE, Sukumar N (2005) Irregular lattice model for quasistatic crack propagation. Phys Rev B 71(9):094106
20. Hemmer PC, Pradhan S (2007) Failure avalanches in fiber bundles for discrete load increase. Phys Rev E 75(4):046101

21. Pradhan S, Chakrabarti BK (2006) Search for precursors in some models of catastrophic failures. In: Bhattacharyya P, Chakrabarti BK (eds.) Modelling critical and catastrophic phenomena in geoscience. A statistical approach Springer, Heidelberg, pp. 459–477
22. Domański Z, Derda T, Sczygiol N (2013) Critical avalanches in fiber bundle models of arrays of nanopillars, lecture notes in engineering and computerScience. In: Proceedings of the international multiConference of engineers and computer scientists, Hong Kong, 13–15 Mar 2013, pp 765–768

Quantum Chemical Study of Point Defects in Tin Dioxide

Richard Rivera, Freddy Marcillo, Alexander Chamba, Patricio Puchaicela and Arvids Stashans

Abstract First-principles calculations based on the density functional theory (DFT) within the generalized gradient approximation (GGA), and the introduction of intra-atomic interaction term for strongly correlated *d*-electrons (DFT+*U*), have been utilized to study defective SnO_2 crystals. Introduction of some impurities, such as fluorine, gallium, aluminium and chromium affect the structural, electronic properties and magnetic properties of tin dioxide. F-doping produces alterations in the structure, with Sn atoms moving away from the impurity and O atoms moving closer to it; and, the system presents *n*-type electrical conductivity. Ga impurity incorporation distorts its surrounding, with the atoms moving closer to the impurity whereas the electrical properties of crystal remain unchanged. Results for Al impurity doping are almost the same as those for the Ga-doping. Cr presence produces the atoms in the neighbourhood of the point defect to move towards it, the band gap width has been slightly reduced and we observe the occurrence of a local magnetic moment.

R. Rivera (✉) · A. Stashans
The Grupo de Fisicoquímica de Materiales, Universidad Técnica Particular de Loja, Loja, Ecuador
e-mail: rariveraxx@utpl.edu.ec

A. Stashans
e-mail: arvids@utpl.edu.ec

F. Marcillo · A. Chamba
The Grupo de Fisicoquímica de Materiales, and Escuela de Ingeniería Química, Universidad Técnica Particular de Loja, Loja, Ecuador
e-mail: fpmarcillo@utpl.edu.ec

A. Chamba
e-mail: wachamba@utpl.edu.ec

P. Puchaicela
The Grupo de Fisicoquímica de Materiales, and Departamento de Geología, Minas e Ingeniería Civil, Universidad Técnica Particular de Loja, Loja, Ecuador
e-mail: ppuchaicela@utpl.edu.ec

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 275, DOI: 10.1007/978-94-007-7684-5_2,

Keywords DFT+U · Electrical conductivity · Electronic properties · Impurity doping · Microstructure · SnO_2.

1 Introduction

Tin dioxide (SnO_2) is a semiconductor oxide that crystallizes with the rutile structure. It can be described through a tetragonal crystalline lattice that belongs to the space group D_{4h}^{14} ($P4_2/mnm$), in which tin atoms are in the centre of an almost regular oxygen octahedron [1, 2]. The primitive unit cell of this material is composed by two formula units (6 atoms), and the lattice parameters are $a = 4.74$ Å and $c = 3.188$ Å [1].

The stoichiometric form of SnO_2 acts as an insulator, but it usually shows a no stoichiometric form, which contains a high presence of oxygen deficiencies (tin interstitials and oxygen vacancies). These defects are responsible of its behaviour as an n-type semiconductor with a direct band-gap width of 3.6 eV [2]. The formation energy of these intrinsic point defects is reduced [3], being the reason of their high presence in this crystal.

Tin dioxide belongs to a certain type of materials that present high electrical conductivity and also a high optical transparency in the visible range of the electromagnetic spectra. These oxides are called transparent conducting oxides (TCOs). In order to understand the importance of these substances it is necessary to remember that most of electrical conductors, such as metals, are opaque; and most of the transparent solids are insulators. Hence, systems that present both characteristics are rare and have importance from both scientific and industrial point of view. SnO_2 is the prototype TCO [4], because of its band-gap, width which offers up to 97 % of optical transparency in the visible range [3]. SnO_2 films are inexpensive, chemically stable in acidic and basic solutions, thermally stable in oxidizing environments at high temperature, and also mechanically strong. Nowadays, it is unclear the reason of the coexistence of the optical transparency alongside with the electrical conductivity, but it has been postulated that it is due the presence of oxygen vacancies [5].

The previously mentioned characteristics make TCOs suitable for applications that require the presence of electrical contacts that allow the pass of visible light [6, 7], solar energy panels, low-emission glasses, heat mirrors [8–10]. In addition, tin dioxide has some other interesting applications such as its use in sub-wavelength waveguides of ultraviolet and visible light [11], gas sensing applications [12], etc.

The nature of the properties of the SnO_2 materials depend on different kind of defects and impurities that are present in the structure of this crystal. These defects could affect its structural, electronic properties, optical and/or magnetic properties. That means a strong necessity to understand the nature of the alterations being produced by the point defects in order to succeed in successful application of the tin dioxide. A large number of theoretical and experimental studies have been carried out for the pristine structure [12–16], and some other ones for the structure containing impurities [17–20]. The present work has the purpose to understand better what is

happening at the fundamental quantum level in this crystal if some impurities such as fluorine, gallium, aluminium and chromium are incorporated in the otherwise pure material.

2 Methodology

Density functional theory (DFT) is an approach that uses the electron density rather than Schrödinger wave function to describe a many-electron system. This method is efficient and is easily accessible; therefore, it is the choice of many research groups. Among the different computational codes that use the DFT method, Vienna *ab initio* Simulation Package (VASP) [21–24] is one of the most known and has been utilised throughout this investigation. The interactions between the core electrons and the valence electrons is implemented through the projector augmented wave (PAW) method, as it was proposed by Blöch [25] and later adapted by Kresse and Joubert [26]. Perdew-Burke-Ernzerhof (PBE) developed [27] GGA functionals are used to describe the exchange-correlation interactions.

A cut-off kinetic energy of 500 eV is used by converging the total energy to less than 1 meV/atom. Γ-centred Monkhorst-Pack (MP) [28] grid with a 0.035 $Å^{-1}$ separation is applied, which corresponds to a k-point mesh of $6 \times 6 \times 8$ for the 6-atom primitive unit cell of the tin dioxide. The previously mentioned parameters are obtained through the atomic relaxation until all the forces are <0.008 eV/Å and the equilibrium state of the system has been achieved.

DFT theory describes inappropriately strong Coulomb repulsion between the d electrons localised on metal ions. In order to take into account these issues, an unrestricted Hartree-Fock (UHF) approximation U-term has been included for the Sn-4d electrons. That results in the so-called DFT+U method [29–32]. Some values for the U parameter have been tested, and finally $U = 4.0$ eV was obtained as a proper magnitude for our system. The computed band-gap width was found to be equal to 1.93 eV. The corresponding experimental value is about 3.6 eV^2, but larger U values have negative impact on the equilibrium structural parameters, so we refused to enlarge the magnitude of this parameter. The computed lattice parameters have been found to be equal to $a = 4.73$ Å and $c = 3.16$ Å, in close agreement with the available experimental data [1].

Finally, in order to study the effects of fluorine, gallium, aluminium and chromium impurities in the SnO_2 crystals, the 6-atom primitive unit cell was expanded sixteen times ($2 \times 2 \times 4$ extension), which resulted in 96-atom supercell, with the corresponding k-point mesh equal to $3 \times 3 \times 2$.

3 Results and Discussion

3.1 F-doped SnO_2

One of the O atoms situated in the central part of the supercell was replaced by an F atom. As a result, atoms in the neighbourhood of the defective region have a tendency to displace themselves in order to find new equilibrium positions. The Sn atoms move outwards the impurity doping by approximately 0.13 Å for Sn(1), and 0.16 Å for the Sn(2) and Sn(3), meanwhile the O atoms have a tendency to displace themselves towards the impurity by 0.04 Å in case of O(5), O(6), O(7) and O(8) atoms. The value of the displacement is not the same for all O atoms since O(10), O(11), O(12) and O(13) atoms shifts only by 0.01 Å. Finally, the O(9) atom does not experience any distortion from its original site. The movements of the atoms are shown in Table 1 as well as Fig. 1.

In an attempt to explain the reason of these motions, Bader charge analysis [33] has been carried out. In this method, a robust algorithm [34–36] is employed to calculate the charge on a particular atom within the crystalline lattice. These computations show that the charge on the incorporated F atom is $-0.76\,e$. The replaced O atom had a charge of $-1.21\,e$ in the pure crystal. This means that the tin atoms move outwards the impurity due to the reduction in the Coulomb electrostatic attraction because of the defect incorporation. Similarly, the O atoms reduce their initial distance until the F atom because of the reduction in the Coulomb repulsion. It is worth to mention that O(9) atom, which has not changed its initial distance with the defect, is the closest O atom to impurity, and has chemical bonds with Sn(3) and Sn(4) atoms. Thus, O(9) atom is trying to preserve its bond length with these tin atoms, and that explains why the distance between the impurity and the O(9) atom remains unchanged. The fact, that Coulomb electrostatic interaction is responsible for the atomic distortion

Table 1 Atomic displacements and charges for F-doped SnO_2 crystal

Atom	Q1(e)	Q2(e)	ΔR (Å)
F (1)	–	−0.76	–
Sn (2)	2.41	2.43	0.13
Sn (3)	2.42	2.43	0.16
Sn (4)	2.42	2.43	0.16
O (5)	−1.20	−1.21	−0.04
O (6)	−1.21	−1.21	−0.04
O (7)	−1.21	−1.21	−0.04
O (8)	−1.20	−1.21	−0.04
O (9)	−1.21	−1.21	–
O (10)	−1.21	−1.22	−0.01
O (11)	−1.21	−1.22	−0.01
O (12)	−1.20	−1.22	−0.01
O (13)	−1.21	−1.22	−0.01

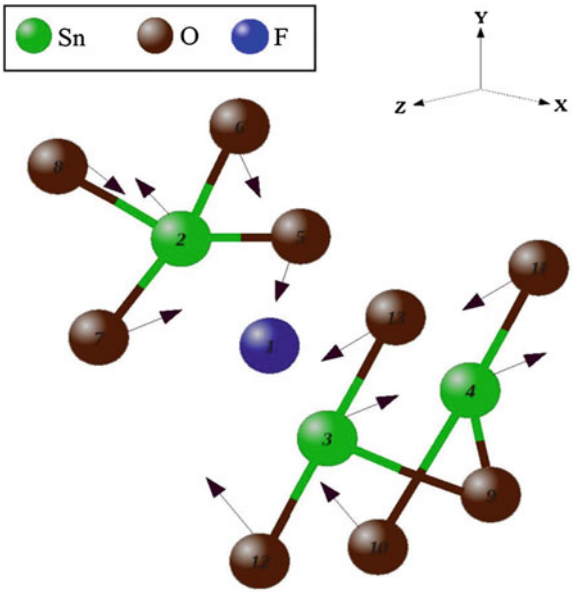


Fig. 1 Atomic displacements in the SnO_2 crystalline lattice in the neighbourhood of the F-dopant. The Sn atoms move towards the impurity while the O atoms move outwards the defect

was also found in F-doped TiO_2, [37] which has almost equal structure to the SnO_2 crystal.

Density of States (DOS) pattern for F-doped tin dioxide is shown in Fig. 2. It is possible to observe that the Fermi level has been displaced from the top of the occupied states, and now it is situated at the bottom of the conduction band (CB), which means that the introduction of the fluorine impurity, which brings to the system one extra valence electron, produces a metallic state in the CB and leads to the *n*-type electrical conductivity. That is in accordance to a number of available experimental observations [38–40].

3.2 Ga-doped SnO_2

Gallium doping was done by replacing one of the lattice central host Sn atoms by the impurity. The atoms surrounding the impurity move themselves in order to find new equilibrium positions, as it is possible to see in Fig. 3. All the atoms in the neighbourhood of the impurity reduce their initial distance with respect to the

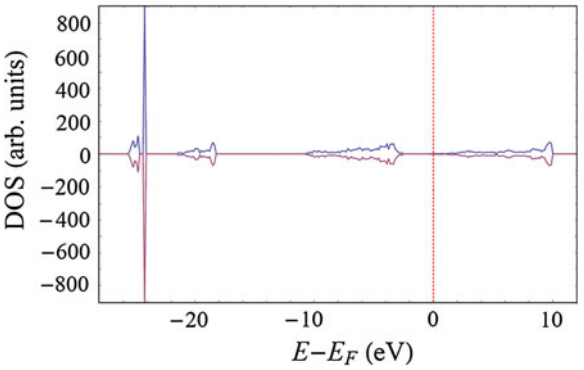


Fig. 2 Total DOS of the F-doped SnO_2 crystal. *The dotted line* marks the Fermi level (E_F)

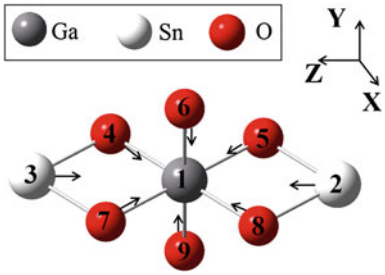


Fig. 3 Atomic displacements in the SnO_2 crystalline lattice in the neighbourhood of the Ga dopant. Both Sn and O atoms have a tendency to move towards the impurity (Atomic shifts in case of the Al- and Cr-doping are essentially the same)

Table 2 Atomic displacements and charges for Ga-doped SnO_2 crystal

Atom	Q1(e)	Q2(e)	ΔR (Å)
Ga (1)	–	1.77	–
Sn (2)	2.42	2.43	−0.04
Sn (3)	2.42	2.43	−0.04
O (4)	−1.21	−1.17	−0.02
O (5)	−1.21	−1.17	−0.02
O (6)	−1.21	−1.16	−0.02
O (7)	−1.21	−1.17	−0.02
O (8)	−1.21	−1.17	−0.02
O (9)	−1.20	−1.16	−0.02

defect. Sn(2) and Sn(3) atoms move by 0.04 Å, and O atoms are displacing towards the impurity by 0.02 Å. Distances can be seen in Table 2.

The Bader charge analysis shows that the impurity has a charge of 1.77 *e* instead of 2.41 *e* for the replaced Sn atom. This means that the Coulomb repulsion force to the nearest tin atoms is smaller than that in pure crystal state, and as a consequence Sn atoms have a tendency to move closer to the dopant. O atoms move slightly closer too, most likely due to their intention to keep their bond lengths with their nearest Sn atoms, as it can be seen in Fig. 3.

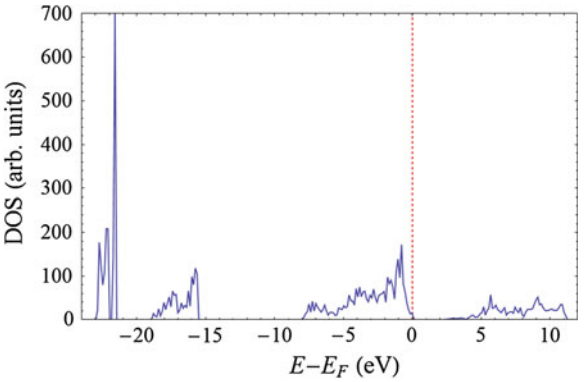


Fig. 4 Total DOS of the Ga-doped SnO_2 crystal. *The dotted line* marks the Fermi level (E_F)

Table 3 Atomic displacements and charges for Al-doped SnO_2 crystal

Atom	Q1(e)	Q2(e)	ΔR (Å)
Al (1)	–	2.48	–
Sn (2)	2.42	2.46	−0.07
Sn (3)	2.42	2.46	−0.07
O (4)	−1.21	−1.29	−0.09
O (5)	−1.21	−1.29	−0.09
O (6)	1.21	−1.26	−0.07
O (7)	−1.21	−1.29	−0.09
O (8)	−1.21	−1.29	−0.09
O (9)	−1.21	−1.26	−0.07

Figure 4 shows the DOS pattern for the given system. Ga atom has some small contributions in the upper valence band (VB) and the CB as well. However, there is no presence of any local energy level within the band-gap region, and no changes for the band-gap width have been detected. That means the presence of the Ga impurity in SnO_2 has no notable influence upon the electrical conductivity in this material.

3.3 Al-doped SnO_2

An Al atom substituted one of the central host tin atoms in the lattice. Once the procedure of doping has been occurred, the atoms in the neighbourhood of defect start to displace themselves in order to find new equilibrium positions. The obtained relevant atomic displacements are shown in Table 3 (the movements are similar to those in the Ga-doped case, so Fig. 3 can be used to visualize the movements). It is possible to see that the nearest Sn atoms move towards Impurity doping by 0.07 Å, and most of the oxygens have a similar tendency, moving 0.09 Å, except for the O(6) and O(9) atoms, which move by 0.07 Å.

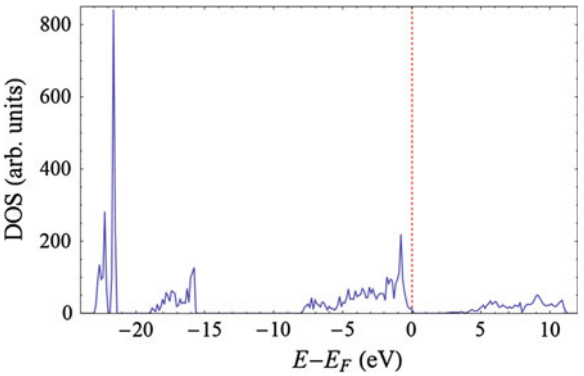


Fig. 5 Total DOS of the Al-doped SnO_2 crystal. *The dotted line* marks the Fermi level (E_F)

Bader charge analysis shows that the charge on the dopant atom is about 2.48 *e*, which is slightly larger than the charge of the replaced Sn atom (2.41 *e*) in undoped material. Despite small change of this atomic charge, we observe notable charge alterations on the defect-closest atoms. The tin atoms become more positive whereas the O atoms are more negative. This phenomenon means that the nature of the chemical bonding has become more ionic, which explains why the O atoms move towards the impurity. The Sn atoms try to preserve their bond lengths with the nearest O atoms, moving along with them, even if it reduces their distance regarding the dopant.

The DOS (Fig. 5) is practically unchanged due the presence of the Al atom. Impurity atom has some small contributions in the upper VB and the CB, but no presence of any local band-gap level has been observed. The band-gap width increases slightly up to 1.78 eV, which means there are no major changes on electrical conductivity.

3.4 Cr-doped SnO_2

Cr-doping was done similarly to the previous cases: one of the central Sn atoms was replaced by the impurity. In order to describe appropriately the *d*electrons of the chromium impurity, parameter $U = 4.0$ eV has been utilized [41, 42]. The introduction of impurity affects its surroundings with local microstructure being altered. Movements of both, Sn and O atoms are towards the impurity, and can be seen in Table 4 (atoms have moved in a similar way as in the case of Ga-doping, so Fig. 3 can be used to visualize the movements), with the closest tin atoms moving by approximately 0.03 Å towards the impurity, and most of the O atoms displacing by 0.06 Å, except for O(6) and O(9) atoms, which have moved 0.09 Å.

Bader charge analysis shows that the charge on the Cr impurity is 2.01 *e*, which is smaller than the 2.41 *e* on the replaced Sn atom. That could explain the movements

Table 4 Atomic displacements and charges for Cr-doped SnO_2 crystal

Atom	Q1(e)	Q2(e)	ΔR (Å)
Cr (1)	–	2.01	–
Sn (2)	2.42	2.47	−0.03
Sn (3)	2.42	2.47	−0.03
O (4)	−1.21	−1.16	−0.06
O (5)	−1.21	−1.16	−0.06
O (6)	−1.21	−1.13	−0.09
O (7)	−1.21	−1.16	−0.06
O (8)	−1.21	−1.16	−0.06
O (9)	−1.21	−1.13	−0.09

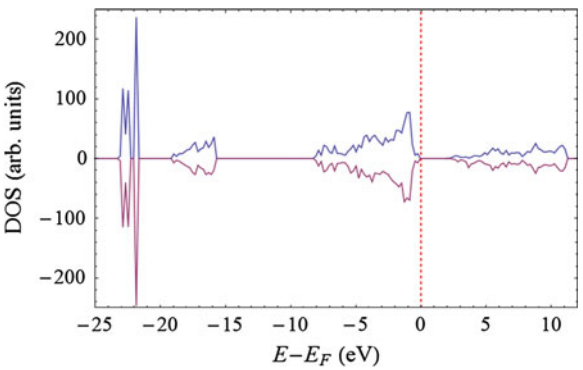


Fig. 6 Total DOS of the Cr-doped SnO_2 crystal. *The dotted line* marks the Fermi level (E_F)

of the nearest Sn atoms: the Coulomb repulsion has decreased. On the other hand, it is found that O atoms also move towards the impurity. In this case, it is necessary to state that the atomic radius of the Cr atom, r_{Cr} = 0.615 Å [43] is smaller than the corresponding radius of the replaced Sn atom r_{Sn} = 0.69 Å [43]. That is why due to the impurity incorporation, the bond length is artificially larger than the value it should be according to the atomic radii, and the O atoms are rearranging themselves according to this fact, trying to keep an appropriate bond length with the defect, which leads to the defect-inward atomic shifts. The same effect, regarding the atomic radii, has been observed in some other researches [44–46].

DOS has been also calculated and is depicted in Fig. 6. There are no major changes upon the band structure compared to the corresponding DOS picture of the pure crystal. Nevertheless, some contributions of the Cr 3*d* atomic orbitals (AOs) in the CB for the β spin can be noticed around 4 eV above the Fermi level. The band-gap width increases slightly up to 1.85 eV.

The presence of the Cr impurity doping leads to the occurrence of local magnetic moment in the lattice. The total magnetic moment of supercell has been found to be around +2.0 μ_B, with impurity contribution being approximately +2.27 μ_B. Small

negative contributions of defect-nearest O atoms towards the magnetism have been also observed.

4 Conclusions

A quantum-mechanical study of tin dioxide (SnO_2) in the presence of some impurities through the DFT+U approach has been carried out. The obtained structural parameters for pure crystal are in accordance with the available experimental results.

The main conclusions due to the doping of impurities can be drawn from the analysis of the obtained results. In case of the F-doping, it is found that the O atoms move towards the F atom whereas the Sn atoms move outwards the defect, due to changes in the electrostatic interaction within the defective region. Fluorine substitution leads to the increase in the n-type electrical conductivity.

The presence of a Ga impurity doping produces atomic shifts towards the defect. Neither local energy levels within the band-gap region nor changes in the band-gap width have been observed.

Al-doping changes slightly the nature of chemical bonding in its neighbourhood. It becomes more ionic due to increase of both positive and negative atomic charges on Sn and O atoms, respectively. The DOS analysis shows slight increase of the band-gap width for the doped material.

The incorporation of a Cr atom produces defect-inward atomic shifts. Cr atom itself has local influence upon the band structure properties of the material with some contributions towards the CB of material for the β spin. Small increase in the band-gap width is also observed. Finally, Cr dopant produces local magnetic moment in the crystalline lattice, being equal to +2.0 μ_B.

References

1. Yamanaka T, Kurashima R, Mimaki J (2000) X-ray diffraction study of bond character of rutile type SiO_2, GeO_2 and SnO_2. Z Kristallogr 215(7):424–428
2. Godinho KG, Walsh A, Watson GW (2009) Energetic and electronic structure analysis of intrinsic defects in SnO_2. J Phys Chem 113:439–448
3. Kiliç C, Zunger A (2002) Origins of coexistence of conductivity and transparency in SnO_2. Phys Rev Lett 88(9):095501
4. Lewis BG, Paine DC (2000) Applications and processing of transparent conducting oxides. MRS Bulletin 25(8):22–27
5. Maestre D, Ramirez-Castellanos J, Hidalgo P, Cremades A, Gonzalez-Calbet JM, Piqueras J (2007) Study of defects in sintered SnO_2 by high resolution transmission electron microscopy and cathodoluminescence. Eur J Inorg Chem 11:1544–1548
6. Wagner JF (2003) Transparent Electronics. Science 300:1245–1246
7. Presley RE, Munsee CL, Park CH, Hong D, Wager JF (2004) Tin oxide transparent thin-film transistors. J Phys D 37(20):2810–2813

8. Zhang B, Tian Y, Zhang J, Cai W (2010) The FTIR studies on the structural and electrical properties of SnO_2:F films as a function of hydrofluoric acid concentration. Optoelectron Adv Mat 4(8):1158–1162
9. Moholkar AV, Pawar SM, Rajpure KY, Bhosale CH, Kim JH (2009) Effect of fluorine doping on highly transparent conductive spray deposited nanocrystalline tin oxide thin films. Appl Surf Sci 255(23):9358–9364
10. Kuantama E, Han DW, Sung YM, Song JE, Han CH (2009) Structure and thermal properties of transparent conductive nanoporous F:SnO_2 films. Thin Solid Films 517(14):4211–4214
11. Sirbuly DJ, Law M, Yan H, Yang P (2005) Semiconductor nanowires for subwavelength photonics integration. J Phys Chem B 109,:15190–15213
12. Chaisitsak S (2011) Nanocrystalline SnO_2:F thin films for liquid petroleum gas sensors. Sensors 11(7):7127–7140
13. Mäki-Jaskari MA, Rantala TT (2001) Band structure and optical parameters of the SnO_2(110) surface. Phys Rev B 64(7):075407–075413
14. Robertson J, Xiong K, Clark SJ (2006) Band gaps and defect levels in functional oxides. Thin Solid Films 496(1):1–7
15. Errico LA (2007) Ab initio FP-LAPW study of the semiconductors SnO and SnO_2. Physica B 389(1):140–144
16. Alterkop B, Parkansky N, Goldsmith S, Boxman RL (2003) Effect of air annealing on optoelectrical properties of amorphous tin oxide films. J Phys D 36(5):552–558
17. Joseph J, Mathew V, Abraham KE (2007) Studies on Cu, Fe, and Mn doped SnO_2 semiconducting transparent films prepared by a vapour deposition technique. Chin J Phys 45(1):84–97
18. Liu XM, Wu SL, Chu PK, Zheng J, Li SL (2006) Characteristics of nano Ti-doped SnO_2 powders prepared by sol-gel method. Mater Sci Eng A 426:274–277
19. Kawamura F, Kamei M, Yasui I (1999) Effect of impurity cations on the growth and habits of SnO_2 crystals in the $SnO_2 - Cu_2O$ flux system. J Am Ceram Soc 82(3):774–776
20. Rivera R, Marcillo F, Chamba W, Puchaicela P, Stashans A (2013) SnO_2 physical and chemical properties due to the impurity doping. Lecture notes in engineering and computer science: proceedings of the international multiConference of engineers and computer scientists. Hong Kong, pp 814–818, 13–15 Mar 2013
21. Kresse G, Hafner J (1993) Ab initio molecular dynamics for liquid metals. Phys Rev B 47(1):558–561
22. Kresse G, Hafner J (1994) Ab initio molecular-dynamics simulation of the liquid-metal-amorphous-semiconductor transition in germanium. Phys Rev B 49(20):14251–14269
23. Kresse G (1996) Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis. Comput Mater Sci 6(1):15–50
24. Kresse G, Furthmüller J (1996) Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. Phys Rev B 54(16):11169–11186
25. Blöchl PE (1994) Projector augmented-wave method. Phys Rev B 50(24):17953–17979
26. Kresse G, Joubert J (1999) From ultrasoft pseudopotentials to the projector augmented-wave method. Phys Rev B 59(3):1758–1775
27. Perdew JP, Ernzerhof M, Burke K (1996) Generalized gradient approximation made simple. Phys Rev Lett 77(18):3865–3868
28. Monkhorst HJ, Pack JD (1976) Special points for Brillouin-zone integrations. Phys Rev B 13(12):5188–5192
29. Anisimov VI, Zaanen J, Andersen OK (1991) Band theory and Mott insulators: Hubbard U instead of Stoner I. Phys Rev B 44(3):943–954
30. Solovyev IV, Dederichs PH, Anisimov VI (1994) Corrected atomic limit in the local-density approximation and the electronic structure of d impurities in Rb. Phys Rev B 50(23):16861–16871
31. Liechtenstein AI, Anisimov VI, Zaanen J (1995) Density-functional theory and strong interactions: orbital ordering in Mott-Hubbard insulators. Phys Rev B 52(8):R5467–R5470

32. Dudarev SL, Botton GA, Savrasov SY, Humphreys CJ, Sutton AP (1998) Electron-energy-loss spectra and the structural stability of nickel oxide: an LSDA+U study. Phys Rev B 57(3):1505–1509
33. Bader RFW (1990) Atoms in molecules: a quantum theory, the international series of monographs on chemistry 22. Oxford University Press, Oxford
34. Henkelman G, Arnaldsson A, Jónsson H (2006) A fast and robust algorithm for Bader decomposition of charge density. Comput Mater Sci 36(3):354–360
35. Sanville E, Kenny SD, Smith R, Henkelman G (2007) Improved grid-based algorithm for Bader charge allocation. J Comput Chem 28(5):899–908
36. Tang W, Sanville E, Henkelman G (2009) A grid-based Bader analysis algorithm without lattice bias. J Phys Condens Matter 21(8):084204
37. Stashans A, Lunell S, Grimes RW (1996) Theoretical study of perfect and defective TiO_2 crystals. J Phys Chem Solids 57(9):1293–1301
38. Zhang B, Tian Y, Zhang JX, Cai W (2011) Structural, optical, electrical properties and FTIR studies of fluorine doped SnO_2 films deposited by sprays pyrolysis. J Mater Sci 46(6):1884–1889
39. Wu S, Yuan S, Shi L, Zhao Y, Fang J (2010) Preparation, characterization and electrical properties of fluorine-doped tin dioxide nanocrystals. J Colloid Interface Sci 346(1):12–16
40. Elangovan E, Ramamurthi K (2005) A study on low cost-high conducting fluorine and antimony-doped tin oxide thin films. Appl Surf Sci 249(1–4):183–196
41. Maldonado F, Rivera R, Stashans A (2012) Structure, electronic and magnetic properties of Ca-doped chromium oxide studied by the DFT method. Physica B 407(8):1262–1267
42. Maldonado F, Novillo C, Stashans A (2012) Ab initio calculation of chromium oxide containing Ti dopant. Chem Phys 393(1):148–152
43. Shannon RD (1976) Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides. Acta Crystallogr A 32:751–767
44. Rivera R, Pinto HP, Stashans A, Piedra L (2012) Density functional theory study of Al-doped hematite. Phys Scr 85(1):015602
45. Patiño E, Stashans A (2001) Structural and electronic effects in $BaTiO_3$ due to the Nb doping. Ferroelectrics 256(1):189–200
46. Stashans A, Eras L, Chamba G (2010) Modelling of Al impurity in perovskite and ilmenite structures of $MgSiO_3$. Phys Chem Minerals 37(4):191–199

The Investigation of the Adsorption of Thiophene on NiMoS Surface: A Density Functional Theory Study

Wahyu Aji Eko Prabowo, Mohammad Kemal Agusta, Nugraha, Subagjo, Ahmad Husin Lubis and Hermawan Kresno Dipojono

Abstract In the last decades, research on hydrotreating processes has regularly been played an essential role in producing clean transportation fuels based international fuel standards. Hydrodesulfurization is one of the most important process in petroleum refinery industry because the limits of sulfur concentration in fuels are currently below 10 ppm. Molybdenum sulfides promoted by nickel or cobalt have been widely used as hydrotreating catalysts for the removal of sulfur from oil fractions. To uncover the physical phenomena responsible for the adsorption of thiophene on the NiMoS active edge sites, the electronic structure of the recommended material is investigated by using density functional theory. The minimum energy of thiophene on the vertical configuration is in bridge S-Mo sites which is about −1.76 eV. In horizontal cofiguration however is in hollow site and is at about −1.70 eV.

Keywords Adsorption energy · Density functional theory · Hydrodesulfurization · NiMoS · Structure optimization · Thiophene

W. A. E. Prabowo (✉) · M. K. Agusta · Nugraha · Subagjo · A. H. Lubis · H. K. Dipojono (✉)
Department of Engineering Physics, Institut Teknologi Bandung, Bandung, Indonesia
e-mail: wahyu.a.e.prabowo@gmail.com

M. K. Agusta
e-mail: kemal@fti.itb.ac.id

Nugraha
e-mail: nugraha@tf.itb.ac.id

H. K. Dipojono
e-mail: dipojono@tf.itb.ac.id

Subagjo
Department of Chemical Engineering, Institut Teknologi Bandung, Bandung, Indonesia
e-mail: subagjo@che.itb.ac.id

Subagjo
Department of Electrical Engineering, Al Azhar Indonesia University, Jakarta, Indonesia
e-mail: ahmlubis@yahoo.com

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 275, DOI: 10.1007/978-94-007-7684-5_3,

1 Introduction

The removal of sulfur, nitrogen, oxygen and metals from crude oil by reductive treatments in so called hydrotreating processes has been very importance ever since oil began to be used as an energy source. In this chemical process, the crude oil is converted into transportation fuels, such as gasoline and diesel oil. SO_2 and NO_x are formed during the combustion of the hydrocarbon containing those hetero atoms. Crude oil and oil products have to be purified because most catalysts which are used for sulfur processing of oil products cannot tolerate sulfur and metals. A further reason for cleanup is to decrease air polluting emissions of sulfur and nitrogen oxides which contribute to acid rain. Furthermore, these types of hydrocarbon have a detrimental effect in further refining processes and in car exhausts.

Hydrotreating also converts olefins and aromatics into saturated hydrocarbon, which burns more cleanly (i.e. fully to CO_2 and H_2O). The annual sale of hydrotreating catalysts is 30 % of the total global catalyst market, which emphasizes the importance of hydrotreating. In particular, the hydrodesulfurization (HDS) is a key process to reduce sulfur contents in diesel and gasoline below 10 ppm [1, 2].

Catalysts based on molybdenum sulfide are widely used in oil industries for the hydrotreatment of petroleum derived feedstocks. In hydrodesulfurization processes , organosulfur molecules are removed by reaction with hydrogen to form H_2S and hydrocarbons [3, 4]. Desulfurization of feedstocks is important for two reasons. First, it prevents sulfur-containing molecules from reaching and deactivating catalysts [5]. Second, it improves the quality of gasoline-related products by reducing the amount of SO_x pollutants formed during the combustion of these fuels [6].

Thiophene is frequently used as a test molecule in HDS studies [9–14], because petroleum feedstocks contain a wide range of sulfur containing aromatic molecules like it. Nickel is known to have a high reactivity towards thiophene [7, 8]. It is able to break the C-S bonds at temperatures as low as 100 K. The industrial catalysts are complex systems that contain a mixture of molybdenum sulfide promoted with Co or Ni on a γ-alumina support [14]. Some comparing effects of several metals (V, Cr, Fe, Co, Ni, Cu, and Zn) on the HDS activity of molybdenum sulfide catalysts, it was found that nickel was the best promoter [15, 16].

The HDS of gasoline produced from the fluid catalytic cracking (FCC) unit requires a selective sulfur removal from thiophene derivatives while avoiding the hydrogenation of olefins (HydO) present in FCC gasoline. This represents a technical challenge to prevent the loss of octane number of gasoline. The HDS selectivity has been the subject of recent experimental works on model molecules for FCC gasoline or real feed [17–22]. Numerous experimentals [23–30], and theoretical works [31–37], have provided atomistic descriptions of the NiMoS active phases.

Although some experimental investigations on NiMoS active phase had been conducted in the recent years, however some problems remain unresolved. A reasonable starting point to begin with is uncovering the adsorption mechanism of thiophene on NiMoS surface. The density functional theory (DFT) [38, 39] based on ab initio computational method will be used for this purpose.

2 Hydrodesulfurization

Hydrodesulfurization (HDS) is a catalytic chemical process widely used to remove sulfur from oil processing (petroleum). The industrial technology used for desulfurization is hydrodesulfurization process. In the solid state model, the desulfurization of MoS_2 start by adsorption of the sulfur atom of the thiophene molecule on the sulfur vacancy. A four electron reduction process [40, 41] then leads to the information of H_2S and butadiene from thiophene:

$$C_4H_4S + \text{butadiene} + 2H^+ + 4e \rightarrow C_4H_6 + S^{2-}$$
$$2H_2 \rightarrow 4H^+ + 4e$$
$$S^{2-} + 2H^+ \rightarrow \text{butadiene} + H_2S$$

The four electrons are delivered by a redox couple like:

$$4Mo^{n+} \rightarrow 4Mo^{(n+1)+} + 4e$$

or

$$2Mo^{n+} \rightarrow 2Mo^{(n+2)+} + 4e$$

Hydrogen is used in a reductive addition reaction on sulfide surfaces, as on some oxides surface [**?**]. When Nickel atom promoted on MoS_2, the surface become metal oxide. The HDS reaction clearly has much in common with the Birch reduction of functional groups in organic molecules in a solution of Na or K in liquid NH_3, in which the reduction is not performed by hydrogen atoms, but by electrons and protons. Under the high H_2 pressure conditions during HDS, transition metal sulfides become sulfur deficient and good electron conductors. Protons are always available at the surface of transition metal sulfides in the form of SH groups (Sulfhydryl group). The analogy with Birch reduction explains the observations made by Pecoraro and Chianelli [42], Vissers et al. [43], and Ledoux et al. [44] that most transition metal sulfides are capable of catalyzing the HDS reaction.

Molybdenum disulfide (MoS_2) promoted by Ni are used to catalyze this reaction. Although pure transition metal surfaces such as Ni [45] show catalytic activity for hydrogenation processes, the difficulty of removing the adsorbed sulfur makes them unsuitable for industrial hydrodesulfurization. It is because the sulfur content saturated on the surface. Number of recent studies has been dedicated to the investigation of the desulfurization of thiophene over MoS_2, Co or Ni surfaces to gain detailed insights into fundamental aspects of the desulfurization [45]. Sulfur-containing molecules react with hydrogen in the presence of a catalyst:

$$C_4H_4S + x\,H_2 \rightarrow H_2S + \text{hydrocarbons}$$

where x is number of H_2

3 Computational Details

The calculations are implemented in the opEn Source Package for Research in Electronic Structure, Simulation, and Optimization (Quantum ESPRESSO) [46]. The ultra-soft pseudopotential method is employed to describe the interaction between ion cores and electrons. The electron exchange correlation is treated by a generalized gradient approximation (GGA) based on Perdew, Burke, and Ernzerhof (PBE) functional [47]. The planewave basis set with a cutoff energy of 340 eV is used for all calculations for the system of the surface.

We use a supercell (12.294 $\times$ 10.683 $\times$ 20.120) Å for NiMoS model with 48 atoms, namely 32 (S) atoms, 12 (Mo) atoms, and 4 (Ni) atoms, respectively. The Monkhorst-Pack method [48] is used to sample k-points by using 3 $\times$ 3 $\times$ 1 grid. Based on the model of MoS_2 the four Mo edge atoms in the surface are substituted by Ni promoter atoms.

Figure 1 shows the structure of the NiMoS surface with the vacuum space of 16.6 Å, which is located above NiMoS surface in the z direction to avoid interactions between surfaces. Figure 1 shows the initial condition for thiophene before being adsorbed in NiMoS surface.

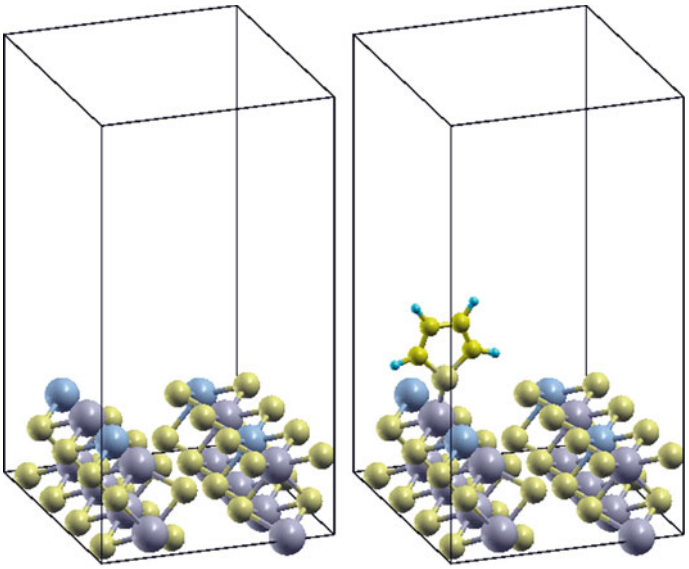


Fig. 1 The structural model for NiMoS surface (*left side*) and the thiophene adsorption on NiMoS surface (*right side*). (*gold* S atom, *grey* Mo atom, *blue* Ni atom, *yellow* C atom, *green* H atom)

4 Results and Discussion

4.1 Optimization Geometry

4.1.1 Optimization Geometry of Thiophene

Thiophene is a heterocyclic compound C_4H_4S that resemble benzene which is especially common in petroleum. Figure 2 shows a molecular model of thiophene. The thiophene content (sulfur) is removed through HDS process . Many kinds of thiophene derivative are formed in petroleum ranging from thiophene itself to more complicated ones like benzothiophenes and dibenzothiophenes. Thiophene itself and its alkyl derivatives are easier to be hydrogenalized. Whereas dibenzothiophene particularly its derivatives are considered to be the most challenging substrates. Benzothiophenes are midway between the thiophenes and dibenzo-thiophenes in their susceptibility to HDS. Therefore, as starting point we calculated the geometry optimization of thiophene to investigate HDS.

The final thiophene optimized-structure is displayed in Table 1. From the computational results, one may conclude that the values of the optimization geometry depend on the calculational method of the same scheme of calculation. If we compare the final result of the optimized structure of thiophene with the experimental and other computational calculation[a] by Itamar et al. [49], they will be in a good

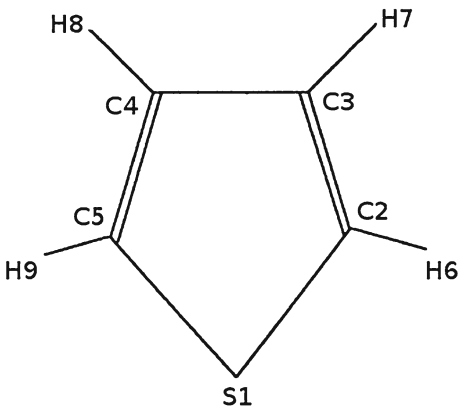


Fig. 2 Thiophene (C_4H_4S) molecular model and atom numbering scheme. (*C* carbon atom, *H* hydrogen atom, and *S* sulphur atom)

Table 1 Geometric parameters for the optimized isolated thiophene

Distance	Calculated[a] (Å)	Calculated[b] (Å)	Calculated[c] (Å)
S_1–C_2	1.71	1.73	1.71
S_1–C_5	1.71	1.73	1.71
C_2–C_3	1.37	1.37	1.37
C_3–C_4	1.42	1.43	1.42
C_4–C_5	1.37	1.37	1.37

[a] by Bouwens et al. [29], [b] by Itamar Borges Jr et al. [49]. [c] by Prabowo et al. [50]

Table 2 Structure comparation for the isolated thiophene

Angle (°)	Calculated[a]	Calculated[b]
$\angle C_2–S_1–C_5$	92.4	92.3
$\angle S_1–C_2–C_3$	111.6	111.3
$\angle C_2–C_3–C_4$	112.2	112.4
$\angle S_1–C_2–H_6$	119.9	120.1
$\angle C_4–C_3–H_8$	124.4	124.3
$\angle C_3–C_2–H_6$	128.5	128.4
$\angle C_2–C_3–H_7$	123.4	123.1

[a] by Kochikov et al. [51], [b] by Prabowo et al. [50]

agreement. Itamar et al. calculations were done using the B3LYP functional. They used the G03 default parameters and quadratic convergence (QC) methods to improve DFT calculation. In our calculations[b], we used Perdew, Burke, and Ernzerhof (PBE) functional for exchange correlation energy. In our calculations[b], the distance of ($S_1–C_2$ is 1.70 Å), ($C_2–C_3$ is 1.37 Å), ($C_3–C_4$ is 1.42 Å) have a good agreement with the experimental result.

Table 2 shows the comparation of the angle in thiophene structure. Our result[b] is in a good agreement with that of Kochikov et al.[a] [51].

4.1.2 Optimization Geometry of NiMoS and Thiophene

The final NiMoS optimized structure is displayed in Fig. 3. It is slightly different from the model obtained by Itamar et al. The model in this investigation consists two-side slabs where the upmost layer of each slabs contains two Ni atoms and two Mo atoms. Itamar et al. used slightly different model with only one side slab and four Ni atoms in the uppermost layer.

Using DFT with periodic boundary conditions, calculation for optimization of geometry is relevant with that of Itamar et al. [49] and with the experimental data and calculation of Raybaud et al. [52]. Periodic boundary condition approach indicates that there is only a slightly distortion in local structure around the Ni atoms, not in the all of surface area. In spite of the distortion, distances between metal atoms are in good agreement with theoretical approach [49, 52].

The Ni-Mo distances from extended X-ray adsorption fine structure experimental calculation and our DFT based calculation are in a good agreement with experimental data [29]. The distances of $Ni_1–Mo_3$, $Ni_2–Mo_5$, $Ni_2–Mo_4$, $Ni_2–Mo_6$, $Ni_1–S_1$, $Ni_2–S_3$, $Ni_2–S_4$ are also in a good agreement with experimental results [29]. The Ni-S distances are a bit larger than experimental values, and a slight distortion is also present. Overall, our calculation for NiMoS is relevant with Raybaud et al. The comparative studies of structure optimization are shown in Table 3, containing four references: [a]from Itamar et al. [49], [b]work by us, experimental works from Bouwens et al. [29], and [c]from Raybaud et al. [52].

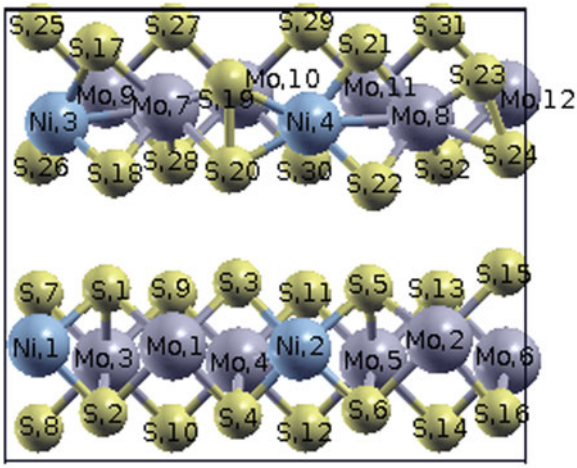


Fig. 3 Scheme of atomic number in z direction of NiMoS surface

Table 3 Structure comparation for NiMoS + thiophene

Distance	Calculated[a] (Å)	Calculated[b] (Å)	Calculated[c] (Å)	Calculated[d] (Å)
Ni_1–Mo_3	2.66	2.96	2.85	2.75
Ni_2–Mo_5	2.86	2.96	2.85	2.75
Ni_2–Mo_4	2.76	2.79	2.85	2.75
Ni_2–Mo_6	5.46	5.45	–	–
Ni_1–S_1	2.30	2.19	–	2.17
Ni_2–S_3	2.25	2.19	–	2.17
Ni_2–S_4	2.25	2.14	–	2.17

[a] by Itamar et al. [49], [b] by Prabowo et al. [50], [c] by Bouwens et al. [29], [d] by Raybaud et al. [52]

Table 4 shows the optimization of our calculation for NiMoS structure. There are two-side slabs in y direction. Both of them have appropriate distances.

4.2 Adsorption Energies of Thiophene

Adsorption of thiophene is a substantial starting step of the hydrodesulfurization process. It is impossible to activate the thiophene molecule suitable for further reaction, if the interaction of thiophene with a catalyst is too weak [53].

Adsorption energies are computed by optimized gas-phase thiophene, optimized NiMoS surface and optimized NiMoS/thiophene according to:

$$E_{ads} = E_{NiMoS/Thiophene} - (E_{NiMoS} + E_{Thiophene}) \quad (1)$$

where E_{ads} is adsorption energy , $E_{NiMoS/Thiophene}$ is total energy of system, E_{NiMoS} is total energy of NiMoS surface, and $E_{Thiophene}$ is total energy of thiophene.

Table 4 Optimized structure for NiMoS surface

Distance	Calculated (Å)	Distance	Calculated (Å)
Ni_1–Mo_1	3.18	Ni_2–Mo_2	3.31
Ni_1–S_2	2.21	Ni_1–Ni_2	6.16
Ni_1–S_1	2.19	Mo_1–Mo_2	6.28
Mo_1–S_2	2.32	Ni_1–Mo_3	2.96
Mo_1–S_1	2.32	Mo_1–Mo_3	3.23
Mo_1–S_4	2.30	Mo_1–Mo_4	3.22
Mo_1–S_3	2.30	Ni_2–Mo_4	2.79
Ni_2–S_4	2.14	Ni_2–Mo_5	2.96
Ni_2–S_3	2.19	Mo_2–Mo_5	3.20
Ni_2–S_6	2.13	Mo_2–Mo_6	3.23
Ni_2–S_5	2.27	Ni_2–Mo_3	5.17
Mo_2–S_6	2.43	Ni_2–Mo_6	5.45
Mo_2–S_5	2.29	Ni_1–Mo_4	5.36
Mo_2–S_{15}	2.33	Mo_1–Mo_5	5.50
Mo_2–S_{16}	2.23	Mo_2–Mo_4	5.48
Mo_1–Ni_2	3.02	–	–
Ni_3–Mo_7	2.60	Mo_8–S_{23}	2.42
Mo_7–Ni_4	3.60	Ni_3–Mo_9	3.22
Ni_4–Mo_8	2.73	Mo_7–Mo_9	3.21
Ni_3–S_8	2.11	Mo_7–Mo_{10}	3.37
Ni_3–S_7	2.30	$Mo_7 - Mo_8$	6.02
Mo_7–S_8	2.28	Ni_3–Ni_4	6.07
Mo_7–S_{17}	2.44	Ni_2–Mo_{10}	4.39
Mo_7–S_{20}	2.55	Ni_4–Mo_{11}	4.33
Mo_7–S_{19}	2.48	Mo_8–Mo_{12}	3.84
Ni_4–S_{20}	2.18	Ni_3–Mo_{10}	5.25
Ni_4–S_{19}	2.26	Mo_7–Mo_{10}	5.68
Ni_4–S_{22}	2.14	Ni_4–Mo_9	6.31
Ni_4–S_{21}	2.18	Ni_4–Mo_{12}	6.18
Mo_8–S_{22}	2.23	Mo_8–Mo_{10}	5.37
Mo_8–S_{21}	2.34	Mo_8–Mo_{11}	3.44
Mo_8–S_{24}	2.48	–	–

4.2.1 Vertical Configuration

4.2.2 Horizontal Configuration

4.3 Hydrodesulfurization Process

After we know the active sites of the adsorption of thiophene on the NiMoS surface, we try to investigate the disulfurization process from the reaction:

$$C_4H_4S + xH_2 \rightarrow H_2S + \text{hydrocarbons}$$

Table 5 Adsorption energy for vertical configuration of thiophene

Position	ΔE_a (eV)	Distance (Å)
Top S	−1.50	1.80
Top Mo	−1.59	2.51
Top Ni	−0.56	2.48
Bridge Ni–Mo	−0.66	2.05
Bridge S–Mo	−1.76	2.21
Bridge S–Ni	−1.66	2.20
Bridge S–S	−0.29	2.24
Hollow 1	−1.70	3.77
Hollow 2	−1.59	1.77

ΔE_a is Adsorption Energy

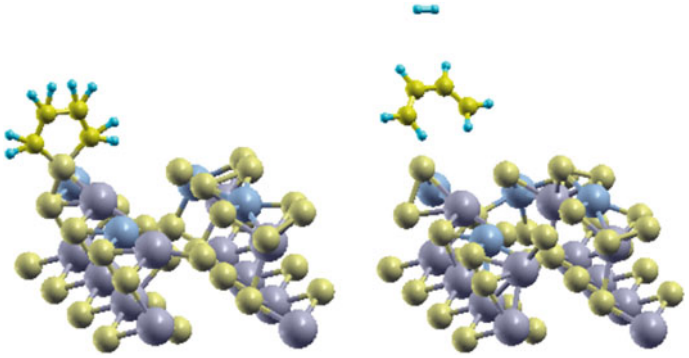


Fig. 4 Adsorption energy of thiophene $+4H_2$ on NiMoS surface (vertical configuration). *left side* is the initial condition of the system, and the *right one* is the final condition of the system

We put some of H_2 on the near of thiophene. As we had done before, we investigate the effect of H_2 promoted on the thiophene. There are two different result of this investigation. On the vertical configuration (Fig. 4), the sulfur content remove from thiophene whereas on the horizontal configuration (Fig. 5), the sulfur content cannot remove.

5 Discussion

The most stable site for adsorption on the surface is bridge S-Mo. The adsorption energy is −1.76 eV or it is about −40.58 kcal/mol. The distance of S atom in thiophene and the surface is 2.21 Å. The adsorption energy of thiophene in the top site of Ni atom is −12.94 kcal/mol or it is about −0.56 eV (Table 5). Using similar orientation of the adsorbed thiophene (Fig. 6), this result is in agreement with a periodic DFT calculation [53], that is −11.5 kcal/mol or it is about −0.49 eV. The adsorption

Table 6 Adsorption energy for horizontal configuration of thiophene

Position	ΔE_a (eV)	Distance (Å)
Top S	−1.52	2.86
Top Mo	−1.40	1.50
Top Ni	−1.49	2.39
Bridge S–Ni	−1.40	1.49
Bridge S–Mo	−1.30	0.83
Bridge Ni–Mo	−0.80	3.39
Bridge S–S	−1.65	2.74
Hollow 1	−1.70	2.56
Hollow 2	−1.52	1.11

ΔE_a is Adsorption Energy

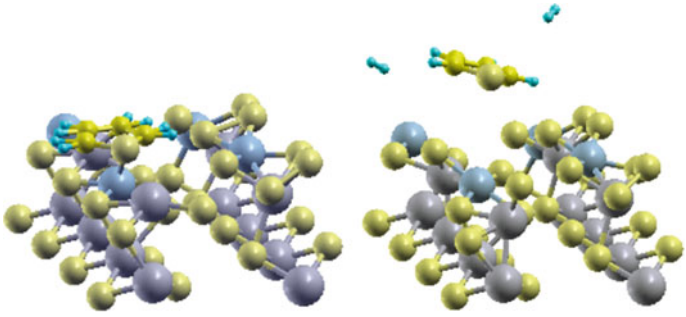


Fig. 5 Adsorption energy of thiophene $+4H_2$ on NiMoS surface (horizontal configuration). *left side* is the initial condition of the system, and the *right one* is the final condition of the system

energy of DFT calculations [53] used a $Ni_1Mo_{15}S_{32}$ cluster with two Mo atoms on the corners and one Ni atom is about −7.8 kcal/mol or −0.34 eV. The distance of S atom in thiophene and Ni atom on the surface is 2.48 Å. It is in a good agreement with Orita et al. The adsorption energy of thiophene in the top site of Mo atom is −36.88 kcal/mol or it is about −1.5 eV. This result is a bit larger than that of Orita et al, that is −25.37 kcal/mol or it is about −1.1 eV. This is due to the distortion of Mo atom with four S atom (Mo_3 with S_1, S_2, S_3, S_4), causing Mo atom bounded. The distance of S atom from thiophene and Ni atom from the surface is 2.51 Å, a bit larger than that calculated by Orita et al., that is 2.35 Å.

The next case we want to deal with is the interaction of NiMoS and thiophene in horizontal configuration (Fig. 7). The most stable site for horizontal configuration is hollow (Table 6). The energy is about −39.34 kcal/mol or −1.7 eV. This result is a bit different with that of Sun et al. [56]. Their calculation shows that the energy adsorption for the thiophene with horizontal configuration is −48.2 kcal/mol or it is about −2.09 eV. Our system consists of two-side slabs, whereas that of Sun et al. [56] used one cluster for their calculations. This is opening up many possibilities to get an adsorption energy from horizontal configuration. Although the above results are different, it is still consistent to the sulfur-metal bond strength energy of 1–2 eV

and is almost identical to that in adsorption energy of thiophene. This relationship is explained by the basic idea that a value of sulfur-metal bond energy optimizes the appropriate interaction between sulfur organic compounds and coordinately unsaturated active sites [25].

6 Conclusion

The NiMoS model is constructed based on the model of MoS_2. Geometry optimization of NiMoS surface, thiophene, and NiMoS/thiophene are carried out. The substitution of the Mo atom with Ni atom changes the structural and electronic properties of the surface.

The adsoprtion energy of thiophene $+4H_2$ molecules in the vertical configuration is -3.75 eV. Whereas, the energy for horizontal configuration is -5.78 eV. The next step of our research will be investigating the saturated of the surface by sulfur content.

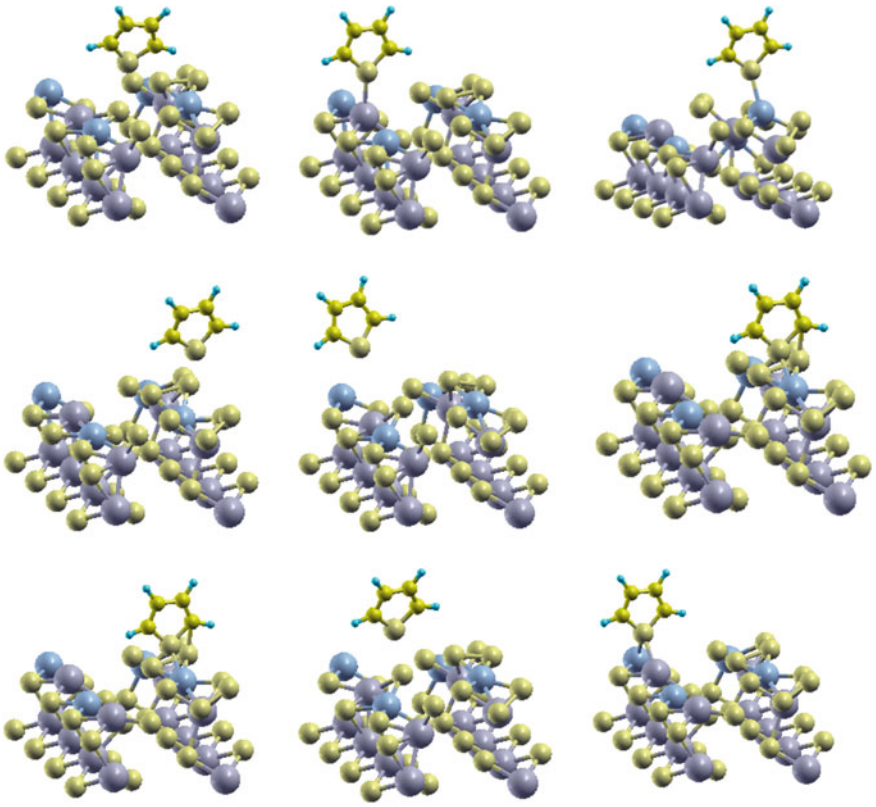


Fig. 6 Adsorption energy of thiophene on NiMoS surface. From *left to right* and *top to bottom* top S, top Mo, top Ni, bridge Ni–Mo, bridge S–Mo, bridge S–Ni, bridge S–S, hollow-1, hollow-2

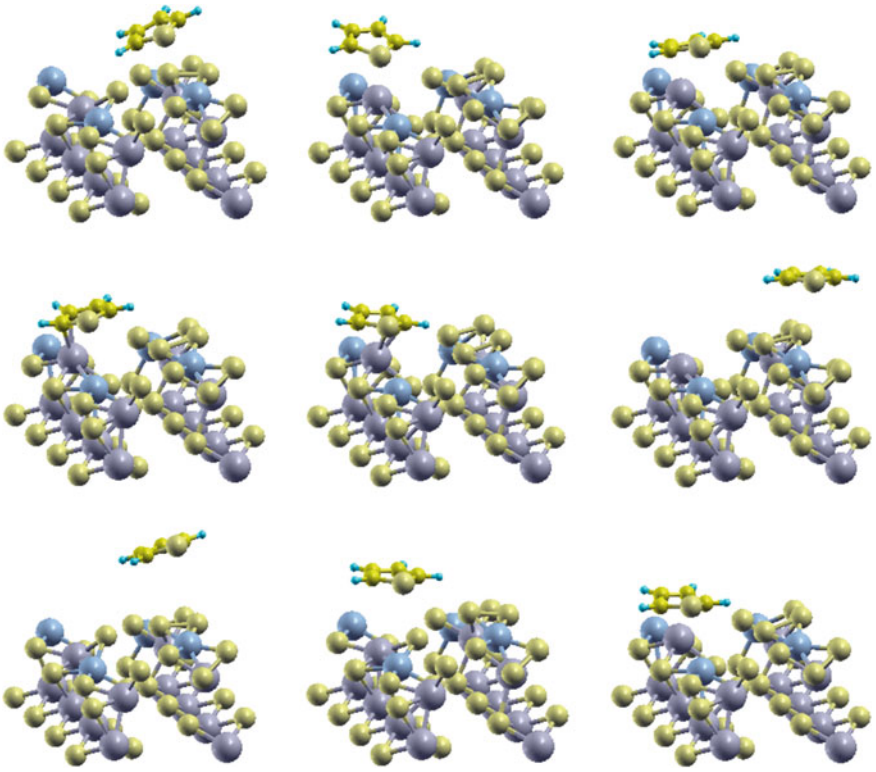


Fig. 7 Adsorption energy of thiophene on NiMoS surface. From *left to right* and *top to bottom* top S, top Mo, top Ni, bridge S–Ni, bridge S–Mo, bridge Ni–Mo, bridge S–S, hollow-1, hollow-2

It is because sulfur adsorbed on the surface. Some of theoretical works have provided atomistic descriptions and the mechanism of HDS process [45, 54, 55, 57–60], but it is not a specific investigation of the HDS reaction of thiophene adsorption on NiMoS surface.

Acknowledgments The authors would like to express their gratitude for the generous financial support from the Directorate General of Higher Education of the Republic of Indonesia through the International Join Research and Publication Project Grant.

References

1. Prins R (1997) In: Ertl G, Knzinger H, Weitkamp J (eds) Handbook of heterogeneous catalysis, vol 4. Wiley-VHC Verlagsgesellschaft, Weinheim, p 1908
2. Topse H, Clausen BS, Massoth FE (1996) In: Anderson JR, Boudart M (eds) Hydrotreating catalysis-science and technology, vol 11. Springer, Berlin/Heidelberg

3. Thomas JM, Thomas WJ (1997) Principles and practice of heterogeneous catalysis. VCH, New York
4. Delmon B (1995) Selectivity in HDS, HDN, HDO and Hydrocracking contribution of remote control and other new concepts, Bull Soc Chim Belg 104:173–187
5. Bartholomew CH, Agrawal PK, Katzer JR (1982) Sulfur poisoning of metals, Adv Catal 31:135–242
6. Stern AC, Boubel RW, Turner DB, Fox DL (1984) Fundamentals of air pollution, 2nd edn. Academic Press, Orlando, FL
7. Zaera F, Kollin EB, Gland JL (1987) Vibrational characterization of thiophene decomposition on the Mo(100) surface, Surf Sci 184:75–89
8. Chen JG (1996) Carbide and nitride overlayers on early transition metal surfaces: preparation, characterization and reactivities, Chem Rev 96:1477–1498
9. Chisholm MA (ed) (1997) Symposium on modeling the chemistry of hydrotreating processes. Polyhedron 3071–3246
10. Huntley DR, Mullins DR, Wingeier MP (1996) Desulfurization of thiophene compounds byMi(111): adsorption and reactions of thiophene, 3-methylthiophene adn 2,5 dimethylthiophene, J Phys Chem 100(50):19620–19627
11. Stohr J, Kollin EB, Fischer DA, Hastings JB, Zaera F, Sette F (1985) Surface extended X-Ray-absorption fine structure of low-z adsorbates studied with fluorescene detection, Phys Rev Lett 55:1468–1471
12. Rodriguez JA (1992) Bonding and decomposition of thiophene, sulfhydryl, thiomethoxy and phenyl thiolate on Mo surfaces, Surf Sci 278:326–338
13. Roberts JT, Friend CM (1987) The reactions of thiophene on Mo(110) and Mo(110)-p(2x2)-s, Surf Sci 186:201–218
14. Chianelli RR, Daage M, Ledoux MJ (1994) Fundamental studies of transition metal sulfide catalitic materials, Adv Catal 40:177–232
15. Harris S, Chianelli RR (1986) Catalysis by transition metal sulfides: a theoretical and experimental study of the relation between the synergic system and the binary transition metal sulfides, J Catal 98:17–31
16. Brito JL, Barbosa AL (1997) Effect of phase composition of the oxidic precursor on the HDSactivity of the sulfided molybdates of Fe(II), Co(II), and Ni(II), J Catal 171:467–475
17. Daudin A, Brunet S, Perot G, Raybaud P, Bouchy C (2007) Transformation of a model FCC-gasoline olefin over transition monometallic sulfide catalysts, J Catal 248:111–119
18. Lamic AF, Daudin A, Brunet S, Legens C, Bouchy C, Devers E (2008) Effect of H2S partial-pressure on the transformation of a model FCC gasoline over unsupported molybdenumsulfide-based catalysts, Appl Catal A 344:198–204
19. Mey D, Brunet S, Canaff C, Maug F, Bouchy C, Diehl F (2004) HDS of a model FCC gasolineover a sulfided CoMo/Al2O3 catayst: effect of the addition of potassium, J Catal 227:436–447
20. Brunet S, Mey D, Perot G, Bouchy C, Diehl F (2005) On the hydrodesulfurization of FCC gasoline: a review, Appl Catal A 278:143–172
21. Miller JT, Reagan WJ, Kaduk JA, Marshall CL, Kropf AJ (2000) Selective hydrodesulfurizationof FCC naphta with supported MoS2 catalyst: the role of cobalt, J Catal 193:123–131
22. Choi JS, Mauge F, Pichon C, Olivier-Fourcade J, Jumas JC, Petit-Clair C, Uzio D (2004) Alumina-supported cobalt-molybdenum sulfide modified by tin via surface organometallic chemistry: application to the simultaneous hydrodesulfurization of thiophene compounds and the hydrogenation of olefins, Appl Catal A 267:203–216
23. Daudin A, Lamic AF, Perot G, Brunet S, Raybaud P, Bouchy C (2008) Microkinetic interpretation of HDS/HYDO selectivity of the transformation of a model FCC gasoline over transition metal sulfides, Catal Today 130:221–230
24. Toulhoat H, Raybaud P, Kasztelan S, Kresse G, Hafner J (1999) Transition metals to the sulfur binding energies relathionship to catalytic activities in HDS: back to sabatier with first principle calculations, Catal Today 50:629–636

25. Chianelli RR, Berhault G, Raybaud P, Kasztelan S, Hafner J, Toulhoat H (2002) Periodic trends in hydrodesulfurization: in support of the sabatier principle, Appl Catal A 227:83–96
26. Toulhoat H, Raybaud P (2003) Kinetic interpretation of catalytic activity patterns based on theoretical chemical descriptors, J Catal 216:63–72
27. Lauritsen JV, Bollinger MV, Lgsgaard E, Jacobsen KW, Nrskov JK, Clausen BS, Topse H, Besenbacher F (2004) Atomic-scale insight into structure and morphology changes of MoS2 nanoclusters in hydrotreating catalysts, J Catal 221:510–522
28. Lauritsen LJV, Kibsgaard J, Olesen GH, Moses PG, Hinnemann B, Helveg S, Nrskov JK, Clausen BS, Topse H, Laegsgaard E, Besenbacher F (2007) Location and coordination of promoter atoms in Co- and Ni-promoted MoS2 based hydrotreating catalysts, J Catal 249:220–233
29. Bouwens SMAM, van Veen JAR, Koningsberger DC, de Beer VHJ, Prins R (1991) Extended X-Ray absorption fine structure determination of the structure of cobalt in carbonsupported Co and Co-Mo sulfide hydrodesulfurization catalysts, J Phys, Chem 95:123–134
30. Kasztelan S, Toulhoat H, Grimblot J, Bonnelle JP (1984) A geometrical model of the activephase hydrotreating catalysts, Appl Catal 13:127–159
31. Schweiger H, Raybaud P, Toulhoat H (2002) Promoter sensitive shapes of Co(Ni)MoS nanocatalysts in sulfo-reductive conditions, J Catal 212:33–38
32. Krebs E, Silvi B, Raybaud P (2008) Mixed sites and promoter segregation: a DFT study of the manifestation of le chaterlier's principle for the Co(Ni)MoS active phase in reaction condition, Catal Today 130:160–169
33. Schweiger H, Raybaud P, Kresse G, Toulhoat H (2002) Shape and edge sites modifications of MoS2 catalytic nanoparticles induced by working conditions: a theoretical study, J Catal 207:76–87
34. Raybaud P (2007) Understanding and predicting improved sulfide catalysts: insight from first principles modeling, Appl Catal A 322:76–91
35. Byskov LS, Nrskov JK, Clausen BS, Topse H (1999) DFT calculations of unpromoted and promoted MoS2-based hydrodesulfurization catalysts, J Catal 187:109–122
36. Travert A, Nakamura H, van Santen RA, Cristol S, Paul JF, Payen E (2002) Hydrogen activation on Mo-based sulfides catalyst, a periodic DFT study, J Am Chem Soc 124:7084–7095
37. Sun M, Nelson AE, Adjaye J (2004) On the incorporation of nickel and cobalt into MoS2-edgestructures, J Catal 226:32–40
38. Hohenberg P, Kohn W (1964) Inhomogeneous electron gas, Phys Rev 136:B864–B871
39. Kohn W, Sham LJ (1965) Self-consistent equations including exchange and correlation effects, Phys Rev 140:A1133–A1138
40. de Beer VHJ, Schuit GCA (1976) In: Delmon B, Jacobs PA, Poncelet G (eds) Preparation of catalysts. Elsevier, Amsterdam, p 1908
41. Gates BG, Katzer JR, Schuit GCA (1979) Chemistry of catalytic processes. Mc Graw-Hill, New York, p 423
42. Pecoraro TA, Chianelli RR (1981) Hydrodesulfurization catalysis by transition metal sulfides, J Catal 67:430–445
43. Vissers JPR, Groot CK, van Oers EM, de Beer VHJ, Prins R (1984) Carbon-supportedtransition metal sulfides, Bull Soc Chem Belg 93:813–822
44. Ledoux MJ, Michaux O, Agostini G, Panissod P (1986) The influence of sulfide structure on hydrodesulfurization activity of carbon-supported catalysts, J Catal 102:275–288
45. Mittendorfer F, Hafner J (2003) Initial steps in desulfurization of thiophene/Ni(100)-a DFT study, J Catal 214:234–241
46. Paolo G (2009) QUANTUM ESPRESSO: a modular and open source software project forquantum simulations of materials, J Phys: Condens Matter 21:39
47. Perdew JP, Burke K, Ernzerhof M (1996) Generalized gradient approximation made simple, Phys Rev Lett 77:3865–3868
48. Monkhorst HJ, Pack JD (1976) Special points for brillouin-zone integrations, Phys Rev B 13:5188

49. Jr Itamar Borges, Silva Alexander M (2012) Probing topological electronic effects in catalysis: thiophene adsorption on NiMoS and CoMoS clusters, J Braz Chem Soc 23
50. Prabowo WAE, Agusta MK, Nugraha S, Lubis AH, Dipojono HK (2013) Density functional theory study of the adsorption of thiophene on NiMoS surface. Lecture notes in engineering and computer science: proceedings of he international multiconference of engineers and computer scientists 2013, 13–15 March 2013, Hong Kong, pp 794–797
51. Kochikov et al (2001) The equilibrium structure of thiophene by the combined use of electron diffraction, vibrational spectroscopy and microwave spectroscopy guided by theoretical calculations, J Mol Struct 567–568:29–40
52. Raybaud P, Hafner J, Kresse G, Kasztelan S, Toulhoat H (2000) Structure, energetics, and electronic properties of the surface of a promoted MoS2 catalysts: an Ab initio local density functional study, J Catal 190:128–143
53. Orita H, Uchida K, Itoh N (2004) A volcano-type relationship between the adsorption energy of thiophene on promoted MoS2 cluster model catalysts and the experimental HDS activity: Ab initio density functional study, Appl Catal A 258:115–120
54. Cristol S, Paul JF, Schovsbo C, Veilly E, Payen E (2006) DFT study of thiophene adsorptionon molybdenum sulfide, J Catal 239:145–153
55. Silva AM, Borges I (2011) How to find an optimum cluster size through topological siteproperties: MoSx model clusters, J Comput Chem 32:2186–2194
56. Sun M, Nelson AE, Adjayeb J (2006) Adsorption thermodynamics of sulfurandnitrogen-containing molecules on NiMoS: a DFT study, Catal Lett 109:113–138
57. Morin C, Eichler A, Hirschl R, Sautet P (2003) DFT study of adsorption and dissociationof thiophene molecules on Ni(110), Surf Sci 540:474–490
58. Yaoa X-Q, Lia Y-W, Jiao H (2005) Mechanism of thiophene hydrodesulfurizationon a Mo3S9 model catalysts: a computational study, J Mol Struct THEOCHEM 726:81–92
59. Hinnemann B, Moses PG, Nrskov JK (2008) Recent density functional studiesof hydrodesulfurization catalysts: insight into structure and mechanism, J Phys: Condens Matter 20:064236
60. Kogan Victor M, Nikulshin Pavel A, Rozhdestvenskaya Nadezhda N (2012) Evolution and interlayer dynamics of active sites of Promoted transition metal sulfide catalysts under hydrodesulfurization conditions, Fuel 100:2–16

Investigation on Control Issues in Power Converters for Advanced Micro-Grid Operations

Tsao-Tsung Ma

Abstract This paper investigates a novel design concept, in which a hybrid power interface system (HPIS) is constructed to work smartly in various micro-grid (MG) operations. Some distributed generation (DG) systems, e.g. the wind turbine generator (WTG) and the photovoltaic (PV) systems conventionally generate real power based on natural conditions thus the average utilization rate of the entire asset is normally low. To eliminate this shortcoming, the proposed HPIS aims to use the DG inverter system optimally. To achieve a cost-effective design, the modular design concept and the related droop control algorithms are incorporated into the proposed HPIS to maximize its operating capability in terms of real power regulation, active power filter (APF) functions for current harmonics compensation and reactive power compensator for MG voltage support and power factor correction. The HPIS is designed to fully utilize the DG inverter capacity after performing various real power control functionalities required by the system operator. In this paper, the mathematical model of HPIS and its related controllers designed in two-axis reference frame are firstly addressed. Then, simulation studies on a simplified MG network are carried out in the Matlab/Simulink software environment. Typical results are presented with brief discussions to demonstrate the feasibility and performances of the proposed control schemes.

Keywords Active power filter (APF) · Distributed generator (DG) · Hybrid power interface system (HPIS) · Micro-grid (MG) · Photovoltaic (PV) · Wind turbine generator (WTG)

T.-T. Ma (✉)
Department of Electrical Engineering, National United University,
No.1, Lien-Da, Kung-Ching Li, Miao-Li city 36003, Taiwan R.O.C
e-mail: tonyma@nuu.edu.tw

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 275, DOI: 10.1007/978-94-007-7684-5_4,

1 Introduction

The rapid development of distributed generation (DG) and micro-grid (MG) systems is an inevitable trend; however, with the addition of these new generation units the conventional distribution network has to face many new challenges in terms of system operation, protection, optimization and stability problems. In recent years, renewable energy resources (RES) based distributed power generations (DG), micro-grids (MG) and state-of-the-art communication and control technologies have been recognized to play important roles in the achievement of some energy policies set by Taiwan's government. The goal of these energy policies include reductions in high-polluting power generations and global greenhouse gas emissions, improved diversity and security of energy supply and the exploitation of possible incentives for creating local value added opportunities for the related industrial sectors in Taiwan. Based on the related technical reports in the open literature, potential RES based power generations may include wind turbine generator (WTG) [1], photovoltaic (PV) [2], and fuel cell (FC) [3]. Of these power generating methods, the interests in PV energy is growing worldwide over last ten years. Although the PV generation system is still expensive, according to recent published reports, PV prices have dropped by 45 % over last two years and further drop is expected in the near future [4]. In fact, a number of different PV incentive programs have been introduced in Taiwan since 2008. With the same objectives, some of the developed countries are currently promoting residential and commercial uses of PV generation systems [5–7]. Based on the standards such as IEEE1547, IEEE 929 and UL1741, PV inverter systems should operate at unity power [8]; however, this regulation has some limitations in practical applications. It has been proved that with proper design of the controllers these inverter systems working as the power interfaces in WTG, PV and FC all have the capability to provide additional control functions in addition to the regulation of real power generated from the RES. Intrinsically, some DG systems, e.g. the wind turbine generator and the PV system generate real power based on practical conditions thus the average utilization rate of the entire asset is normally low. This has resulted in that the payback time period for the system owners becomes longer. To make the best use of the DG hardware systems, the concept of utilizing PV inverter as a reactive power controller during the night time for voltage control thereby increasing the connectivity of a nearby wind farm, is proposed in [9]. A number of potential operations of a 3-phase PV grid connected inverter are discussed in [10, 11]. Some similar examples regarding reactive power compensation and voltage support in MG are also illustrated in [12, 13]. However, in the above published papers, only individual control function has been included in the operation of DG inverters. This paper investigates a novel design concept concerning the feasibility of performing multiple control functions in the DG inverter system working in micro-grids. To achieve a cost-effective design, the modular design concept is considered in the proposed HPIS to maximize the operating capability of DG inverter. With the proposed design concept the DG inverter is able to optionally utilize the unused portion of rated capacity after performing the function of real power generation.

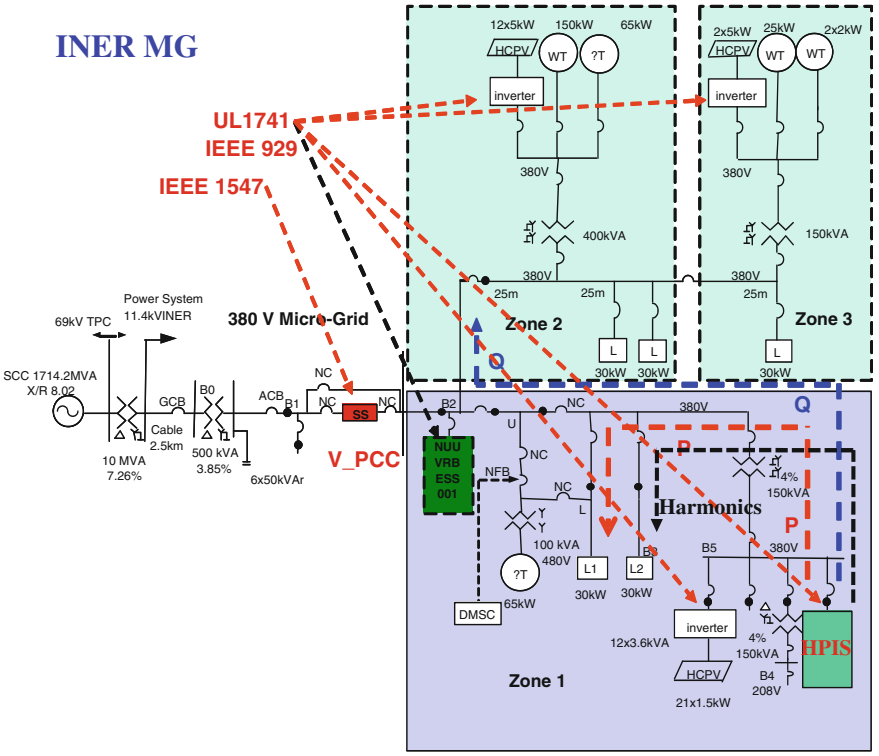


Fig. 1 System diagram of a typical micro-grid with various DG inverters (Taiwan)

2 The First Micro-Grid Test Bed in Taiwan

Distribution systems possessing DGs and controllable loads with the ability to operate in both grid-connected and standalone modes are an important class of the so-called MG power systems [14–16]. A typical MG system constructed by the Institute of Nuclear Energy Research (INER) in Taiwan is shown in Fig. 1 [17]. As can be seen in Fig. 1, the INER micro-grid with a designed maximum capacity of 475.5 kVA has three independent Zones. In Zone1, there is a 31.5 kVA PV, a 65 kVA microturbine, a total of 60 kVA controllable load bank, a 150 kVA wind turbine and a 100 kVA ABB PCS-100 BESS. Zone2 has a 60 kVA PV system, a 65 kVA microturbine and a total of 60 kVA controllable load bank. Zone3 is equipped with a 4 kW small wind turbine, a 25 kVA wind turbine, a 10 kVA PV, a 65 kVA microturbine and a 30 kVA controllable load bank. To facilitate possible tests considering the conditions of practical operating modes, Zone 2 and Zone 3 are connected in series, while Zone 1 and Zone 2 are in parallel.

It is well known that MG strives for optimized operation of the aggregated distribution systems by coordinating the various DGs, ESS and load resources not only when connected to the distribution system but also in a standalone mode. In either modes of operation, advanced local controls, energy management, power quality and protection technologies are required for robustness and reliability. In practice, the energy management optimization objective function can be tailored to the needs of each application. In a practical MG project, the overall objective is to optimize operating performance and cost in the normally grid-connected mode, while ensuring that the system is capable of meeting the performance requirements in standalone mode. To satisfy the needs of possible applications, some ESS units are inevitably required. In the INER MG project, two ESS are to be installed to achieve some advanced operations and power management functions.

3 Modular HPIS Principles and Control Schemes

3.1 HPIS Hardware Configuration

The operating principles and control concepts of the proposed HPIS are actually derived from the static synchronous compensator, a popular shunt-type FACTS device. In this paper, a basic HPIS hardware configuration which consists of a 3-phase switching converter using 6 IGBT switches and a three-phase power grid model, as shown in Fig. 2 is chosen to introduce the proposed HPIS operating principles and its control schemes. The IGBT converter in the HPIS is designed to be operated from a DC link voltage provided by a PV or any kind of RES. In normal operations, the active power can be controlled in either direction between the AC terminals of the converter and the grid to regulate the DC voltage and thus the real power generation of the DG. In this hardware topology, the converter can also generate or absorb reactive power independently at its AC output terminals to act as a reactive power generator or an APF.

In practical applications, two or more HPIS might be needed to work in parallel to meet the system operating requirement. Therefore, implementing effective control over P and Q is very important from the operational and control points of view. Figure 3 shows the system block diagram of a two HPIS modules working in parallel. Of the feasible control schemes, the droop control method has been widely used for controlling the parallel connected DG inverters. In this case, the inverters are controlled in such a way that the amplitude and frequency of the reference voltage signal will follow a droop as the load current increases and these droops are used to allow independent HPIS to share the load in proportion to their capacities.

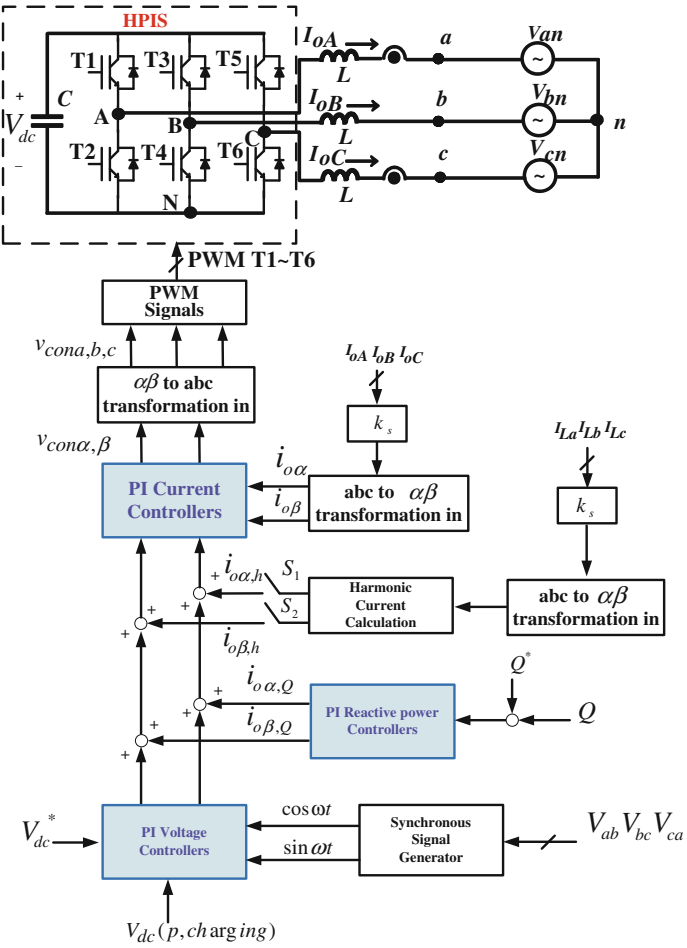


Fig. 2 Simplified MG model with the proposed HPIS and control signals

4 Design of HPIS Control Systems

In this paper, three important control functions are designed for the HPIS; i.e. real power regulation for DG, reactive power compensation for the grid and harmonic currents compensation for the local nonlinear load. These control functions can be activated simultaneously or individually. Because the proposed compensating requirements, i.e. real and reactive power or harmonic compensation, are directly related to the current control, shunt-type connection of DG inverter is a realistic topology as it normally injects currents at PCC. Therefore, this study uses the shunt-type connecting format for the implementation of HPIS control system shown in Fig. 2. With a number of mathematical manipulations, the output voltages and cur-

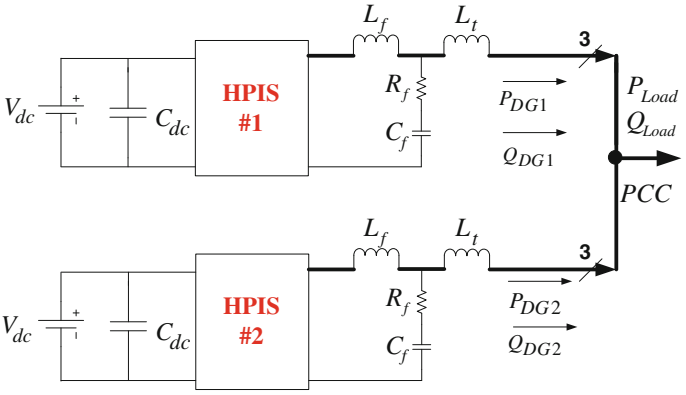


Fig. 3 The system block diagram of a two HPIS modules working in parallel

rent commands of the HPIS can be obtained in two-axis reference frame as briefly addressed below. From Fig. 2, the following voltage and current equations can be easily obtained on KVL.

$$\begin{cases} L\frac{dI_{oA}}{dt} = V_{AN} - V_{an} - V_{nN} \\ L\frac{dI_{oB}}{dt} = V_{BN} - V_{bn} - V_{nN} \\ L\frac{dI_{oC}}{dt} = V_{CN} - V_{cn} - V_{nN} \end{cases} \tag{1}$$

The above (1) can be rewritten into (2) and the equivalent forms in the stationary reference frame as expressed in (3) and (4).

$$\begin{bmatrix} L\frac{dI_{oA}}{dt} \\ L\frac{dI_{oB}}{dt} \\ L\frac{dI_{oC}}{dt} \end{bmatrix} = \frac{2}{3}\begin{bmatrix} 1 & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & 1 & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} & 1 \end{bmatrix}\left(\begin{bmatrix} V_{AN} \\ V_{BN} \\ V_{CN} \end{bmatrix} - \begin{bmatrix} V_{an} \\ V_{bn} \\ V_{cn} \end{bmatrix}\right) \tag{2}$$

$$\begin{bmatrix} L\frac{dI_{o\beta}}{dt} \\ L\frac{dI_{o\alpha}}{dt} \end{bmatrix} = k_{pwm}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} v_{con\beta} \\ v_{con\alpha} \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} v_{s\beta} \\ v_{s\alpha} \end{bmatrix} \tag{3}$$

$$\begin{bmatrix} I_{o\beta} \\ I_{o\alpha} \end{bmatrix} = \frac{k_{pwm}}{sL}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} v_{con\beta} \\ v_{con\alpha} \end{bmatrix} - \frac{1}{sL}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} v_{s\beta} \\ v_{s\alpha} \end{bmatrix} \tag{4}$$

Finally, the three-phase output current signals for the HPIS can be directly derived from (4) by using the inverse Clarke's transformation. In the open literature, many current control methods for three phase inverter systems have been proposed. Among them, a PWM based control scheme using high-frequency switching has been widely used in many applications. This is due to the fact that the design of output filter

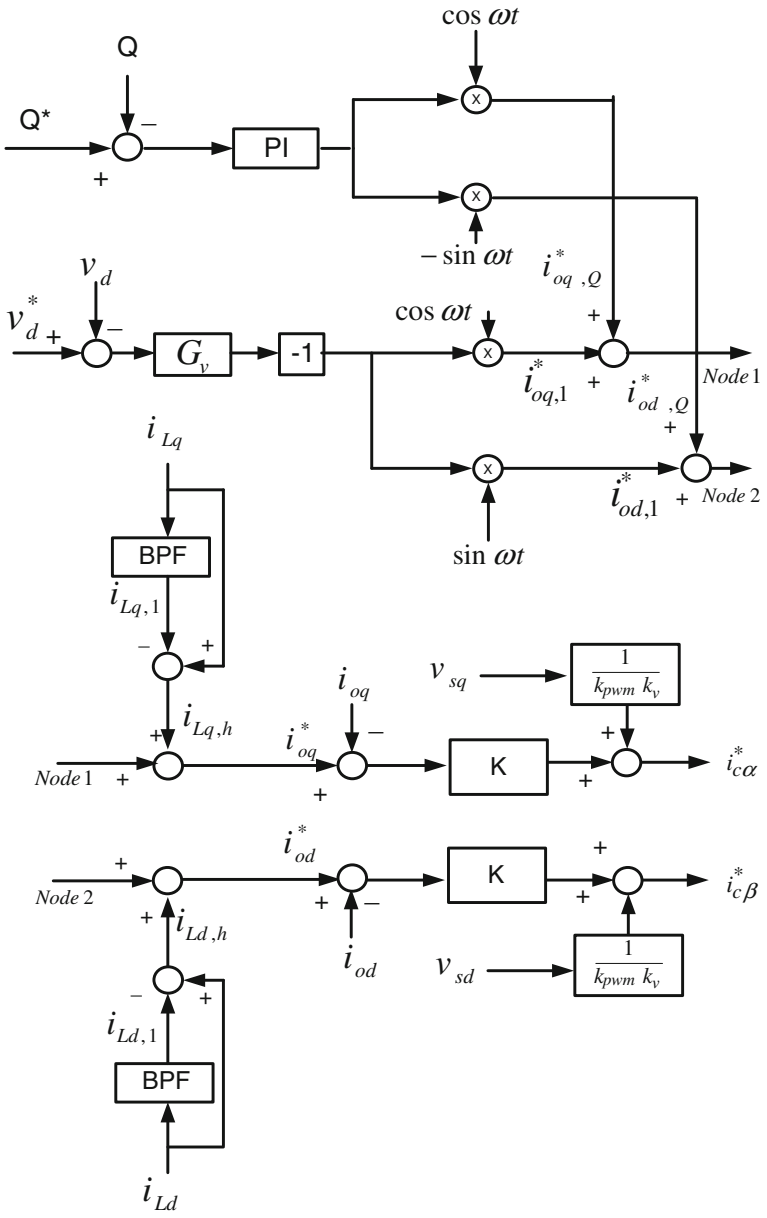


Fig. 4 The signal block diagram of the derived HPIS controller

becomes simple for eliminating the high-frequency current harmonics. Besides, the typical advantages of PWM based current control method are its simplicity in implementation and the high speed of its current control loop. The block diagram of the

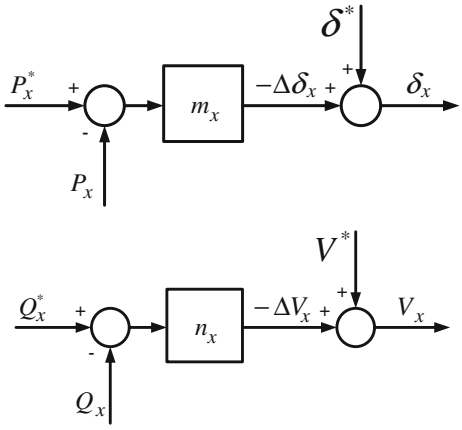


Fig. 5 The signal block diagram of the adopted HPIS droop control controller

proposed current controller based on (4) is shown in Fig. 4. As can be seen from Fig. 4, the input commands to the derived controllers include three parts, i.e. the harmonic compensating current, the reactive power to be compensated, the DC voltage of the HPIS or the real power of the DG if desired. To achieve a better dynamic performance, a feed forward compensating path is also used. To compare dynamic control performances, another design approach on d–q reference frame is also investigated in this paper. Following the mathematical derivations presented previously in (2 - 4), the park's transformation can then be used to obtain the current control signals in d–q reference frame as expressed in (5). After some mathematical manipulations, the equivalent three-phase real and reactive power expressed in d–q reference frame can be expressed as (6). It is clear that the dc voltage on the capacitor in the HIPS can be achieved by regulating the real power (P) coming in and out of the three-phase inverter, while simultaneously control the output reactive power (Q) using the following equation (7) if desired. It follows that the three-phase current command signals for the proposed HPIS can be derived from (7) by using the inverse Park's and Clarke's transformations. To demonstrate the effectiveness of the proposed modular HPIS concept, the widely used droop control method shown in Fig. 5 is adopted for controlling the parallel connected HPIS inverters.

$$\begin{bmatrix} i_{od} \\ i_{oq} \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} i_{o\beta} \\ i_{o\alpha} \end{bmatrix} \tag{5}$$

$$\begin{bmatrix} P^{3-p} \\ Q^{3-p} \end{bmatrix} = \begin{bmatrix} \vec{i} \cdot \vec{v} \\ \vec{i} \times \vec{v} \end{bmatrix} = \begin{bmatrix} v_d i_d + v_q i_q \\ v_d i_q - v_q i_d \end{bmatrix} \tag{6}$$

$$\begin{bmatrix} i^*_{od,P} \\ i^*_{oq,Q} \end{bmatrix} = \begin{bmatrix} i^*_d \\ i^*_q \end{bmatrix} = \begin{bmatrix} P/v_d \\ Q/v_d \end{bmatrix} \tag{7}$$

5 Case Studies and Results

To investigate the detailed dynamics of the proposed HPIS and to validate the proposed two control methods, comprehensive simulation studies are carried out on a MG distribution network connected with a nonlinear DG link in Matlab/Simulink environment. It is considered that for the whole period of simulations the local loads are fed by the main source of the micro-grid feeder. In this study, during the simulation active power which is delivered from DG link is assumed unlimited. This assumption makes it possible to evaluate the capability of HPIS to track the fast change in the harmonic currents and the reactive current components of the reactive power required by the load or the grid independently. To simulate a realistic operation scenario in micro-grid network, the nonlinear loads are connected and disconnected to MG distribution network and the harmonic distortion of current waveform are recorded and compared in various conditions. Since the principle of the proposed current control technique is based on separating active and reactive current components in two reference frames known as the two-axis components, in all conditions only phase-a parameters (voltage and current) are shown. To demonstrate the performance of the proposed modular HPIS to compensate total real and reactive power on variation of load, the R and R-L loads are tested and the output waveforms of the two HPIS modules are shown simultaneously. The switching frequency for the HPIS inverter is set at 20 kHz to achieve satisfactory response. In this study, the DC voltage is controlled at 400 V with the three-phase 220 V, 60 Hz power grid.

Case-1: Connection of nonlinear load link to the MG feeder with the HPIS in operation

In the first simulation case, the HPIS link is connected to the network at t = 0.0 s. A nonlinear load is added to PCC and removed at t = 4.5 s. Figure 6a and b respectively show the related current waveforms of the HPIS performing APF control functions with the proposed two control schemes. As can be seen in Fig. 6a with method-1 and b with method-2, after the connection of HPIS link to feeder (grid) the source current becomes sinusoidal since the harmonic currents are fully provided by the HPIS link. Based on the results shown in Fig. 6a and b, the HPIS link controlled with the proposed method-2 which is derived from the d–q reference frame has better dynamic performance. This is due to the fact that the method-2 with control signals derived from the d–q reference frame is able to eliminate the interactions in control signals between the two control loops for the real power and the reactive power.

Case-2: Connection of the nonlinear load link to the MG feeder with two HPIS modules operated in droop control mode with the proposed d–q control scheme

In this simulation case, two HPIS modules are connected to the network at t = 0.0 s. A nonlinear load is also added to PCC at t = 0.0 s. Two types of loads are arranged; i.e. the pure resistance load and the resistance plus inductance load. In all cases, at the simulation time instant of 0.3 s a step change in active and reactive load power are initiated. Figure 7a–e respectively show the related P-Q tracking results of the two HPIS modules. As can be seen in Fig. 7a an equal real power regulation

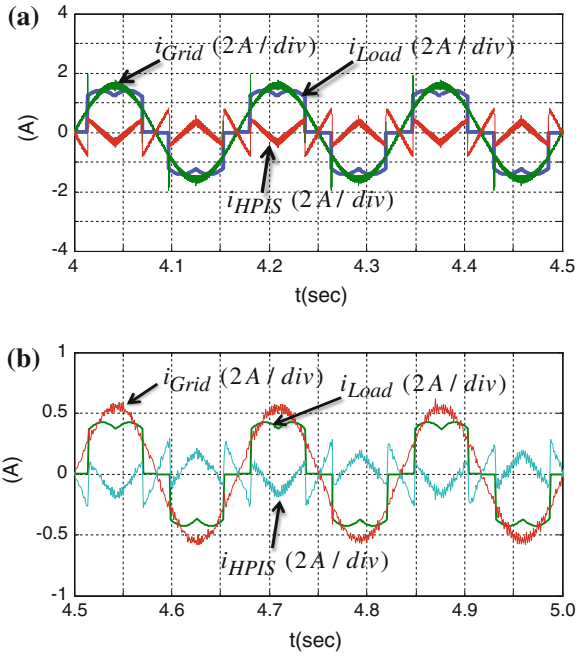


Fig. 6 The related current waveforms in HPIS APF control functions. **a** Simulated results of the related current waveforms (method-1). **b** Simulated results of the related current waveforms (method-2)

result of the two HPIS inverters has been achieved; while the circulating reactive power between the two HPIS modules is negligible. Figure 7b shows the results of using different droop parameters of real power control. Figure 7c shows the results of using a set of identical droop parameters for both real and reactive power control. As can be seen from Fig. 7c both real and reactive powers are equally shared by the two HPIS modules. Figure 7d and e respectively show the results of using different droop parameters in real-power and reactive-power control loops.

6 Conclusion

This paper has demonstrated a novel design of a flexible hybrid power interface system in which any DG inverter system can be utilized as an APF for harmonic currents compensation and a reactive power generator for power factor correction and voltage control at the PCC. Based on the results obtained from the two simulation cases carried out in Matlab/Simulink software environment, the HPIS link with the proposed control schemes exhibits satisfactory functional and dynamic performances. The feasibility and effectiveness of the proposed modular HPIS and the droop control

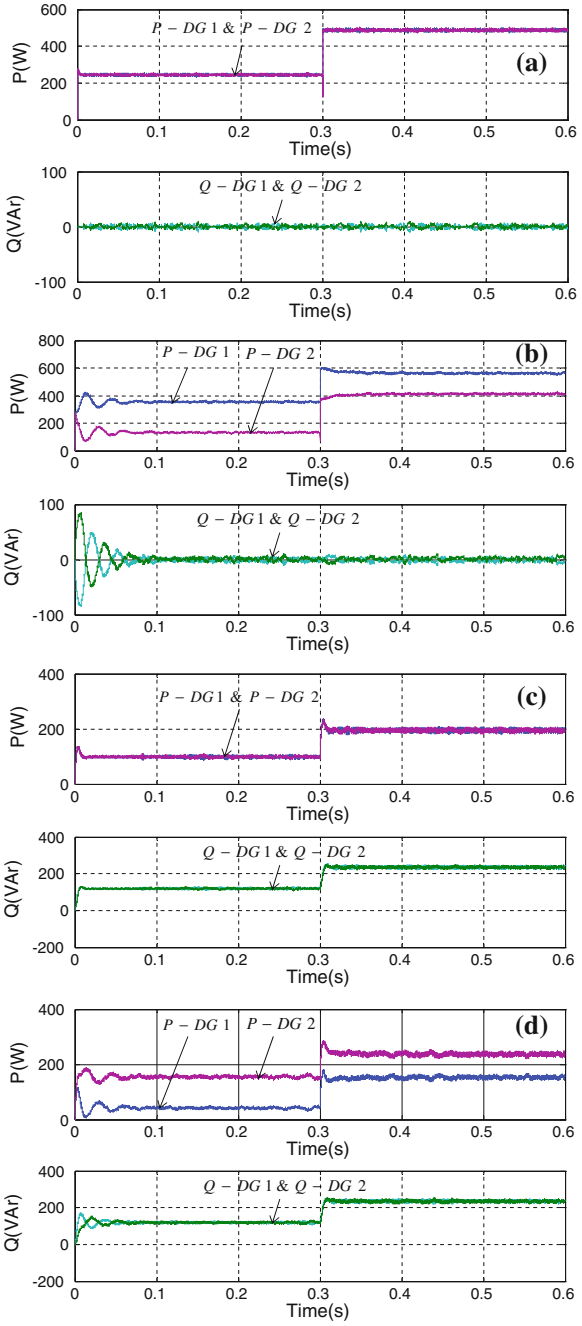


Fig. 7 The results of two HPIS modules operated in droop control mode

scheme have been confirmed. It is important to note that better dynamic response can be achieved if decoupled current controllers are used in the control loops of the HPIS.

Acknowledgments This work was supported in part by the National Science Council of Taiwan, R.O.C. through: NSC 101 - 2221-E -239-031.

References

1. K.M. Syafii and M. A-A. Nor, Steady-State Wind Turbine Generation Model for Three-Phase Distribution Load Flow Analysis, IREMOS, Vol. 4 N. 2, April 2011 (Part B), pp. 772–777.
2. S. Arabi Nowdeh, B. Tousi, A. A. Zoraghchian and M. Hajibeigy, Optimal PV and FC Application in a Hybrid Power System with the Aim of Selling Electrical Energy to Distribution Network, IREMOS, Vol. 4 N. 5, October 2011(Part B), pp. 2392–2401.
3. Aghajani S, Joneidi IA, Kalantar M, Mortezapour V (February 2010) Modeling and Simulation of a PV/FC/UC Hybrid Energy System for Stand Alone Applications. IREMOS 3(1):82–89
4. Grubb M (1995) Renewable Energy Strategies for Europe, The Royal Institute of International Affairs, vol I. Foundations and Context, London, UK
5. Miyamoto Y, Hayashi Y (Oct. 2010) Evaluation of improved generation efficiency through residential PV voltage control of a clustered residential grid-interconnected PV. IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT Europe) 1:1–8
6. H. Laukamp, The new German electric safety standard for residential PV systems, The 25^{th} IEEE Photovoltaic Specialists Conference, 13 May 1996, pp. 1405–1408.
7. J. Carr, J.C.Balda and A.Mantooth, A high frequency link multiport converter utility interface for renewable energy resources with integrated energy storage, Energy Conversion Congress and Exposition (ECCE), 12–16 Sept. 2010 IEEE, pp. 3541–3548.
8. IEEE Standard for Interconnecting Distributed Resources with Electric Power Systems, IEEE Standard 1547–2003, July 2003.
9. Varma RK, Khadkikar V, Seethapathy R (Dec. 2009) NighttimeApplication of PV solar farm as STATCOM to Regulate Grid Voltage. IEEE Trans. Energy Conversion (Letters) 24(4):983–985
10. R.A. Mastromauro, M.Liserre, T.Kerekes and A.Dell'Aquila, A Single-Phase Voltage-Controlled Grid-Connected Photovoltaic System With Power Quality Conditioner Functionality, IEEE Transactions on Industrial Electronics, Vol. 56,Issue 11, 2009, pp. 4436–4444.
11. M. Anwari, M.I.M. Rashid and Taufik, Power quality analysis of grid-connected photovoltaic system with Adjustable Speed Drives, 2010 International Conference on Control Automation and Systems (ICCAS), 2010, pp. 2452–2456.
12. Majumder R, Ghosh A, Ledwich G, Zare F (2010) Power Management and Power Flow Control With Back-to-Back Converters in a Utility Connected Microgrid. IEEE Transactions on Power Systems 25(2):821–834
13. Chen C-L, Wang Y (2010) Jih-Sheng Lai, Yuang-Shung Lee and D. Martin, Design of Parallel Inverters for Smooth Mode Transfer Microgrid Applications, IEEE Transactions on Power Electronics 25(1):6–15
14. Nikkhajoei H, Lasseter RH (2009) Distributed Generation Interface to the CERTS Microgrid. IEEE Transactions on Power Delivery 24(3):1598–1608
15. R.H. Lasseter, A Akhil, C. Marnay, J Stephens, J Dagle, R Guttromson, A. Meliopoulous, R Yinger, and J. Eto, The CERTS Microgrid Concept, White Paper for Transmission Reliability Program, Office of Power Technologies, U.S. Department of, Energy, April 2002.
16. Tsao-Tsung Ma and Tzung-Han Shr, Design of a Cost-Effective Power Interface for Advanced Micro-grid Operation and Control, Lecture Notes in Engineering and Computer Science: Proceedings of The International MultiConference of Engineers and Computer Scientists 2013, 13–15 March, 2013, Hong Kong, pp 651–656.

17. Ma TT (November 2012) Yih-Der Lee, Yung-Ruei Chang and Chin-Lung Hsieh, “Advanced Operation and Control Schemes for a Micro-grid with Battery Energy Storage Systems”. International Journal of Advanced Renewable Energy Research, Vo. 1(6):596–604
18. H. K. Høidalen and R. Sporild, Using Zigzag Transformers with Phase-shift to reduce Harmonics in AC-DC Systems, International Conference on Power Systems Transients (IPST’05) in Montreal, Canada, June 19–23, 2005, Paper No. IPST05 - 44.

Temporal Characteristics of Wavelet Subbands of Epileptic Scalp EEG Data Based on the Number of Local Min–Max

Suparerk Janjarasjitt

Abstract Epilepsy is a chronic brain disorder characterized by recurrent seizures. An electroencephalogram (EEG) which records the electrical activity of the brain can help diagnose seizures. Temporal characteristics of the EEG provide an insight into the states of the brain including epileptic seizures. In this study, the temporal characteristics of epileptic scalp EEG subband signals obtained using the discrete wavelet transform associated with various states of the brain including the pre-ictal, ictal and post-ictal states, are examined using a simple computational measure, referred to as the number of local min–max. From the computational results, it is observed that in any wavelet subband the EEG subband signals associated with different states of the brain exhibit distinguishing characteristics of the number of local min–max. The most remarkable temporal characteristics of EEG subband signals can be observed in the D_1 and A_3 subbands which, respectively, correspond to the highest and lowest frequency components of the EEG signals. In particular, during an epileptic seizure activity the computational results suggest that there is an increase of amplitude regularity of the highest frequency components while there is a decrease of amplitude regularity of the lowest frequency components. Furthermore, the computational results show that the number of local min–max of the D_1 and A_3 subband signals of epileptic EEG can be potentially useful for epileptic seizure classification and detection accompanied with further digital signal processing and analysis.

Keywords Electroencephalogram · Epilepsy · Local min–max · Seizure · Temporal characteristics · Wavelet transform

S. Janjarasjitt (✉)
Department of Electrical and Electronic Engineering, Ubon Ratchathani University,
85 Sathonlamak Road, Warin Chamrap, Ubon Ratchathani 34190, Thailand
e-mail: ensupajt@ubu.ac.th; suparerk.janjarasjitt@case.edu

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 275, DOI: 10.1007/978-94-007-7684-5_5,

1 Introduction

Epilepsy is a common neurological disorder in which clusters of nerve cells or neurons in the brain sometimes signal abnormally [1]. More than 50 million individuals worldwide, about 1 % of the world's population, are affected by epilepsy [2]. In epilepsy, the normal pattern of neuronal activity is disturbed, causing strange sensations, emotions, and behavior, that sometimes include convulsions, muscle spasms, and loss of consciousness [1]. There are many possible causes for seizures ranging from illness to brain damage to abnormal brain development [1]. Epilepsy is characterized by recurrent seizures that are physical reactions to sudden, usually brief, excessive electrical discharges in clusters of nerve cells [3].

The electroencephalogram (EEG) is a signal that quantifies the electrical activity of the brain, usually from scalp recordings. The EEG is commonly used to assess behaviors of the brain and also detect abnormalities of the brain. Also the EEG is crucial for the fundamental diagnosis of epilepsy [1]. In clinical practice, specific features and characteristics of the EEG are identified by visual inspection and analysis to specify the brain state or abnormality of the brain. Therefore, temporal patterns and characteristics of EEG signals provide an insight into the state of the brain and play a significant role in brain examination. Temporal characteristics of EEG signals may be quantified by a variety of signal processing techniques and computational measures. Concepts and computational tools derived from the study of complex systems including nonlinear dynamics gained increasing attractions for applications in biology and medicine [4]. Moreover, epilepsy is an important application for nonlinear EEG analysis [5, 6].

Local minima and local maxima may be ones of elementary features used for quantifying temporal characteristics of EEG signals. Local minima and maxima may be defined in various manners. There are a number of signal processing techniques and algorithms for identifying and detecting the local minima and maxima. Local min–max detection algorithms generally rely on thresholding the amplitude of signal in a specified time window. As originally defined in [7], in this study the local minima are defined as points whose amplitude is less than their neighbors and the local maxima are defined as points whose amplitude is greater than their neighbors. The local minima and maxima thus simply specify extreme points of the signal where its magnitude changes direction [7].

In this study, the temporal characteristics of scalp EEG signals are basically examined in terms of local minima and local maxima. The computational analysis techniques based on quantifying the temporal characteristics and variabilities of EEG signals in terms of the local minima and maxima have been applied to a number of applications in EEG analysis; for example, the epileptic electrocorticogram (intracranial EEG) data [7–9], the sleep EEG data [10–13], the epileptic scalp EEG data [14]. This chapter presents an extended study on the temporal characteristics of wavelet subbands of epileptic scalp EEG data associated with various states of the

brain preliminarily reported in [14]. The temporal characteristics of epileptic scalp EEG signals corresponding to various wavelet subbands are quantified using a local min–max feature, referred to as the number of local min–max.

2 Materials and Methods

2.1 Data and Subject

The scalp EEG data analyzed in this study were obtained from the CHB-MIT scalp EEG database (available online at http://www.physionet.org/pn6/chbmit/) [15]. This database consists of EEG recordings obtained from pediatric subjects with intractable seizures at the Children's Hospital Boston [15, 16]. The scalp EEG data were recorded using a sampling frequency of 256 Hz with 16-bit resolution A/D. Furthermore, these EEG recordings used the international 10–20 systems of EEG electrode positions and nomenclature [15].

A case *chb01* whose EEG recordings were recorded from a female subject with age of 11 years is analyzed. Six scalp EEG recordings that are referred to as chb01_03, chb01_04, chb01_16, chb01_18, chb01_21 and chb01_26 are analyzed. Each recording is a 1-h segment of multi-channel scalp EEG data containing one epileptic seizure event. Only the Fz-Cz channel of the EEG recordings was examined. Note that the lengths of epileptic seizure events occurred in chb01_03, chb01_04, chb01_16, chb01_18, chb01_21, and chb01_26 are 40 s, 27 s, 51 s, 90 s, 93 s, and 101 s, respectively.

Figure 1 shows an EEG signal of the chb01_03 recording. In this recording, an epileptic seizure event occurs between the 2,996th second and 3,036th second as

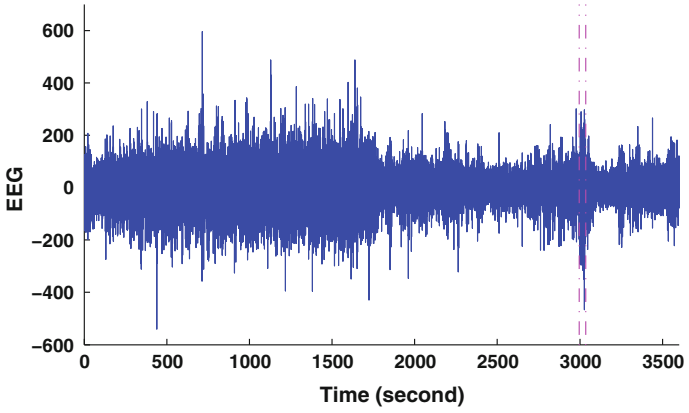


Fig. 1 The chb01_03 EEG signal. The *dashed lines* indicate the beginning and the end of an epileptic seizure event

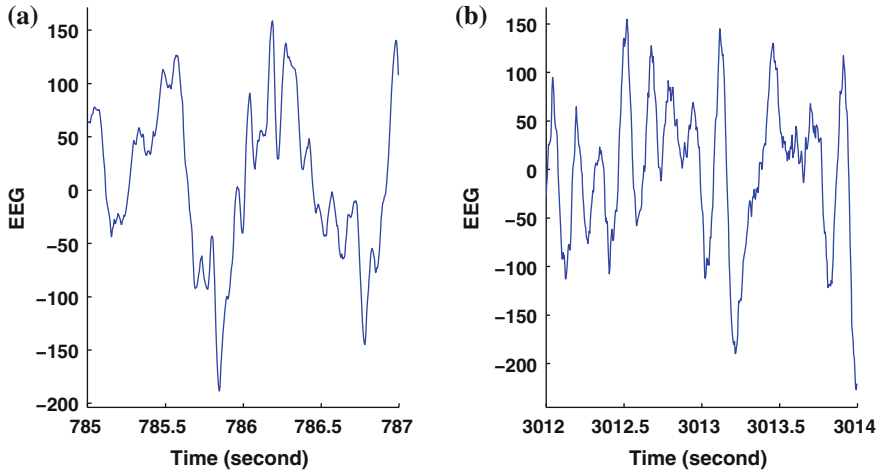


Fig. 2 Segments of the chb01_03 EEG signal corresponding to various states of the brain. **a** During a non-seizure period. **b** During a seizure activity

the dashed lines shown in Fig. 1 indicate the beginning and the end of the epileptic seizure event. In addition, segments of the chb01_03 EEG signal corresponding to a non-seizure period and during an epileptic seizure activity are shown in Fig. 2a and b.

2.2 The Local Min–Max Feature

The local minima of a signal are defined as points whose amplitude is less than that of their consecutive preceding and succeeding points while the local maxima of a signal are defined as points whose amplitude is greater than that of their consecutive preceding and succeeding points. Mathematically, the nth sample of the signal $x[n]$ is called the local minimum if there exist u and v such that satisfies the condition $x[u] > x[n] \wedge x[v] > x[n]$ where $x[u] > x[u+1] = \cdots = x[n]$ and $x[v] > x[v-1] = \cdots = x[n]$ and $u < n < v$, and also the nth sample of the signal $x[n]$ is called the local maximum if there exist u and v such that satisfies the condition $x[u] < x[n] \wedge x[v] < x[n]$ where $x[u] < x[u+1] = \cdots = x[n]$ and $x[v] < x[v-1] = \cdots = x[n]$ and $u < n < v$. In the case that $u < n-1$ or $v > n+1$, the local minimum or the local maximum is specified at $n = \lceil (u+v)/2 \rceil$.

The local minima and the local maxima of an exemplary EEG signal are depicted in Fig. 3. Temporal characteristics of EEG signals can be quantified in various aspects based on the local minima and the local maxima. The temporal characteristics of EEG signals examined in this study are specifically quantified by a total number of local minima and local maxima, referred to as the number of local min–max N_λ.

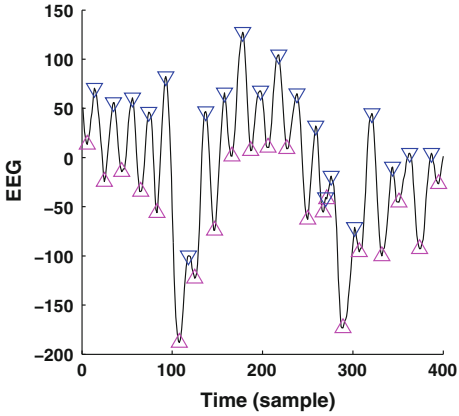


Fig. 3 The local minima specified by '△' and the local maxima specified by '▽' of an EEG signal

2.3 The Discrete Wavelet Transform

The discrete wavelet transform (DWT) is a representation of a signal using a countably-infinite set of wavelets that constitutes an orthonormal basis [17]. From a signal point of view, the wavelet transform can be interpreted as a generalized octave-band filter bank [18, 19] as a wavelet is a bandpass filter [20]. A signal $x[n]$ can be decomposed into approximations and details using the scaling function and the wavelet function that, respectively, correspond to halfband lowpass filter and halfband highpass filter. The signal $x[n]$ that is square summable can be expressed as

$$x[n] = \sum_k a_0[k]\phi_{0,k}[n] \tag{1}$$

$$= \sum_k a_1[k]\phi_{1,k}[n] + \sum_k d_1[k]\psi_{1,k}[n] \tag{2}$$

where the scaling function $\phi_{1,k}[n]$ and the wavelet function $\psi_{1,k}[n]$ are, respectively, an orthonormal basis for the space V_1 and the orthogonal complement of V_1, denoted by W_1, and the space $V_0 = V_1 \oplus W_1$.

The approximation coefficients $a_1[n]$ and the detail coefficients $d_1[n]$ can be obtained by

$$\begin{aligned} a_1[n] &= \sum\nolimits_k a_0[k]h[k-2n] \\ d_1[n] &= \sum\nolimits_k a_0[k]g[k-2n] \end{aligned} \tag{3}$$

where $h[n]$ and $g[n]$ are the impulse response of halfband lowpass filter and halfband highpass filter, respectively. Therefore, for a single-level discrete wavelet decomposition, the approximation coefficients can be obtained by convolving the approximation coefficients $a_0[n]$ with the time-reversed filter of $h[n]$, i.e., $\tilde{h}[n]$, followed by the

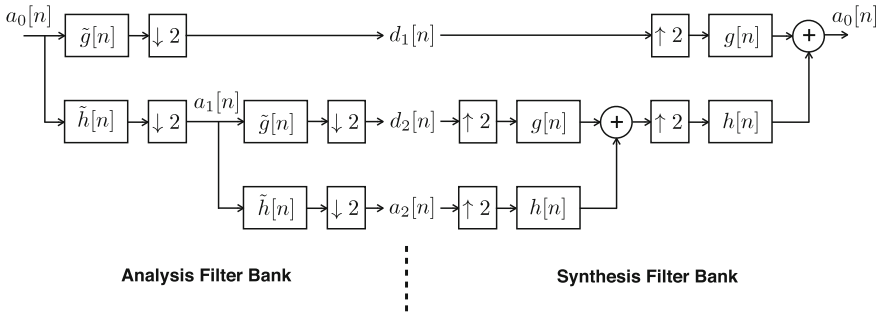


Fig. 4 The discrete wavelet decomposition and reconstruction process

downsampling and, similarly, the detail coefficients can be obtained by convolving the approximation coefficients $a_0[n]$ with the time-reversed filter of $g[n]$, i.e., $\tilde{g}[n]$, followed by the downsampling. This conception forms the two-channel filter bank. The reconstruction of the signal $x[n]$ can be accomplished by the reversed process.

The discrete wavelet transform can thus be realized by successive discrete wavelet decomposition and reconstruction. Figure 4 illustrates a block diagram interepreting the two-level discrete wavelet transform where $\downarrow 2$ and $\uparrow 2$ denote the downsampling and the upsampling, respectively.

2.4 Analytical Framework

The continuous (long-term) epileptic scalp EEG data are divided into epochs using the sliding window technique. The discrete wavelet transform is applied to decompose the EEG epochs into levels, and then reconstruct EEG subband signals from the detail and approximation coefficients using the discrete Meyer wavelets that have a compact support in frequency domain. The corresponding numbers of local min–max of the wavelet subband signals of EEG epochs are consequently determined.

The procedure carried out in this study can be detailed as follows:

1. Applying the sliding window technique to partition the long-term scalp EEG signal into 10-s epochs with a 1-s sliding step.
2. Decomposing an epoch of the scalp EEG signal into three levels, i.e., $j = 1$, 2, and 3, to obtain the detail coefficients $\{d_1[n]\}$, $\{d_2[n]\}$ and $\{d_3[n]\}$, and the approximation coefficients $\{a_3[n]\}$.
3. Reconstructing the wavelet subband signals of the scalp EEG epoch corresponding to the D_1, D_2, D_3 and A_3 subbands from the detail coefficients $\{d_1[n]\}$, $\{d_2[n]\}$ and $\{d_3[n]\}$ and the approximation coefficients $\{a_3[n]\}$, respectively.
4. Identifying the local minima and the local maxima of each scalp EEG subband signal.

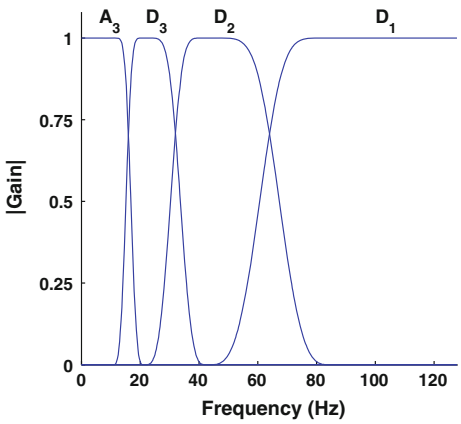


Fig. 5 The spectral subbands of the discrete Meyer wavelets corresponding to D_1, D_2, D_3 and A_3 subbands

5. Determining the corresponding number of local min–max N_λ of each scalp EEG subband signal.

The D_1, D_2, D_3 and A_3 subbands correspond to 64–128Hz, 32–64Hz, 16–32Hz and 0–16Hz spectral bands, respectively. The corresponding frequency response for each wavelet subband is illustrated in Fig. 5. The EEG subband signals depicted in Fig. 2a and b corresponding to the D_1, D_2, D_3 and A_3 subbands are, respectively, shown in Fig. 6a, c, e, and g, and b, d, f, and h.

In addition, the temporal characteristics of the EEG subband signals are examined in terms of the state of the brain. In the computational experiments, the state of the brain is divided into three states: pre-ictal, ictal and post-ictal states. The pre-ictal state in this study is defined as a two-minute period before an epileptic seizure onset while the post-ictal state is defined as a two-minute period after an epileptic seizure activity. The analysis of variance (ANOVA) is used to test whether the numbers of local min–max of the subband signals of scalp EEG data associated with different states of the brain have a common mean.

3 Results and Discussion

3.1 Temporal Characteristics of the chb01_03 Subband Signals

For the chb01_03 EEG signal illustrated in Fig. 1, the numbers of local min–max of the EEG subband signals corresponding to the D_1, D_2, D_3, and A_3 subbands are, respectively, shown in Fig. 7a–d. A shaded strip shown in Fig. 7a–d indicates an epileptic seizure event. It is observed that the EEG subband signals exhibit distinguishing characteristics of the number of local min–max corresponding to different states of the brain in all wavelet subbands. The most intriguing characteristics of the

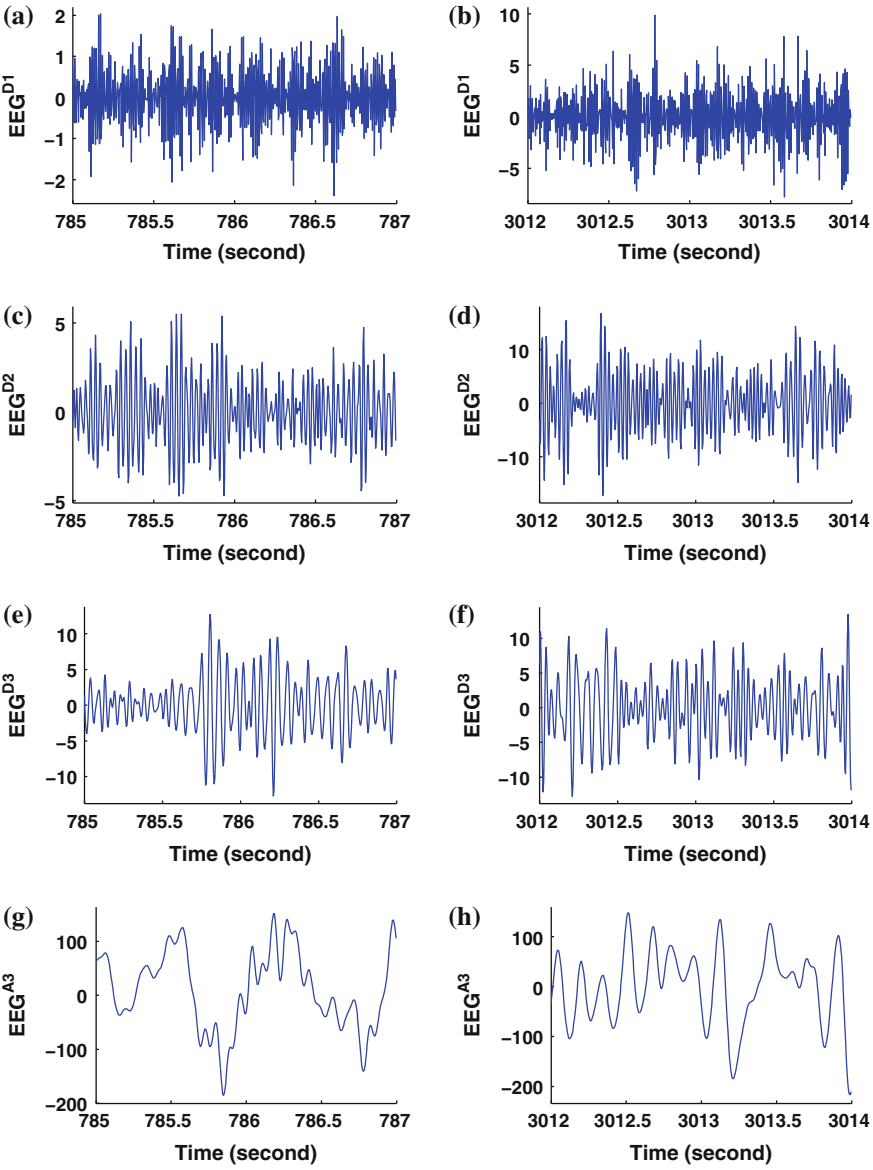


Fig. 6 Subband signals corresponding to the EEG segments depicted in Fig. 2a and b. **a** D_1 subband signal (a non-seizure period). **b** D_1 subband signal (a seizure activity). **c** D_2 subband signal (a non-seizure period). **d** D_2 subband signal (a seizure activity). **e** D_3 subband signal (a non-seizure period). **f** D_3 subband signal (a seizure activity). **g** A_3 subband signal (a non-seizure period). **h** A_3 subband signal (a seizure activity)

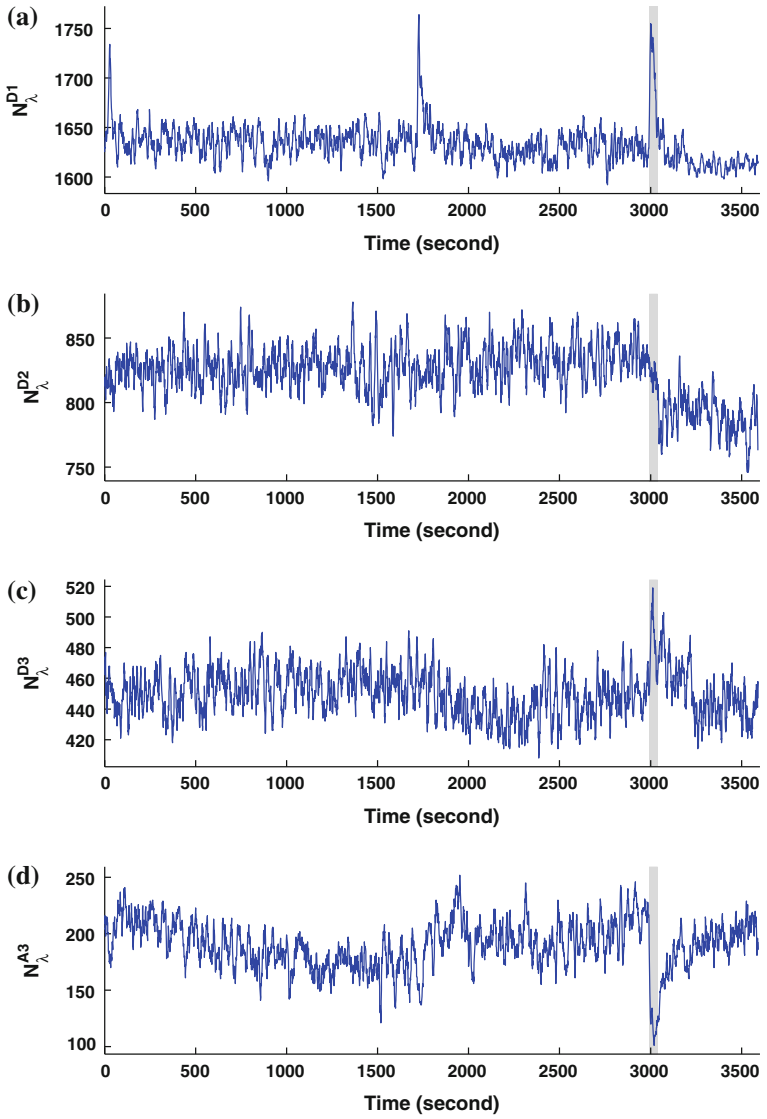


Fig. 7 The numbers of local min–max of the chb01_03 subband signals. **a** D_1 subband. **b** D_2 subband. **c** D_3 subband. **d** A_3 subband

number of local min–max appear in the D_1 and A_3 subbands. In particular, the number of local min–max of the D_1 subband signal instantaneously increases during the seizure onset while the number of local min–max of the A_3 subband signal sharply decreases during the seizure onset.

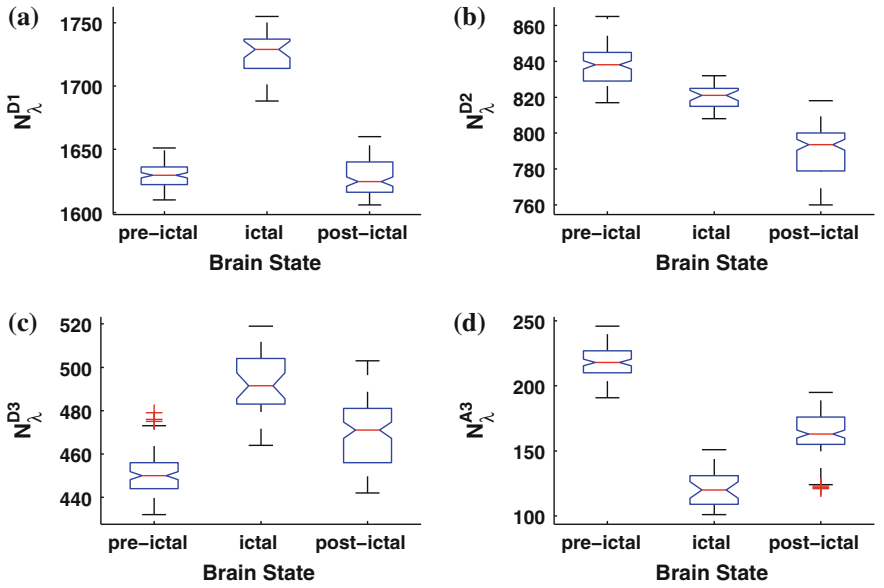


Fig. 8 Comparison of the numbers of local min–max of the chb01_03 subband signals associated with the pre-ictal, ictal and post-ictal states. **a** D_1 subband. **b** D_2 subband. **c** D_3 subband. **d** A_3 subband

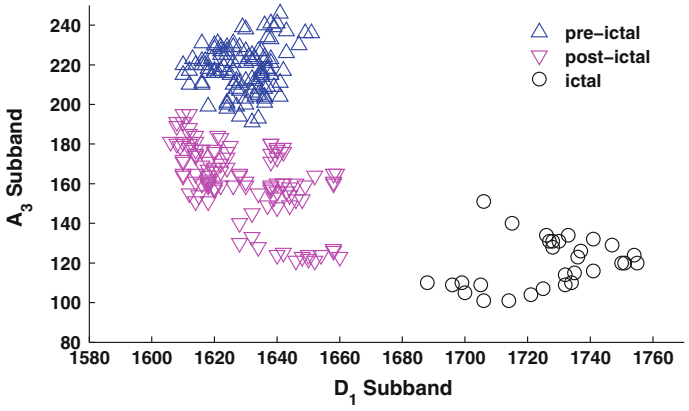


Fig. 9 Comparison of the numbers of local min–max of the chb01_03 subband signals associated with the pre-ictal, ictal and post-ictal states

Figure 8a–d compare the numbers of local min–max of the EEG subband signals associated with different states of the brain, i.e., pre-ictal, ictal and post-ictal states, corresponding to the D_1, D_2, D_3, and A_3 subbands, respectively. Evidently, the numbers of local min–max of the EEG subband signals associated with various states of the brain tend to be substantially different in all wavelet subbands. In the

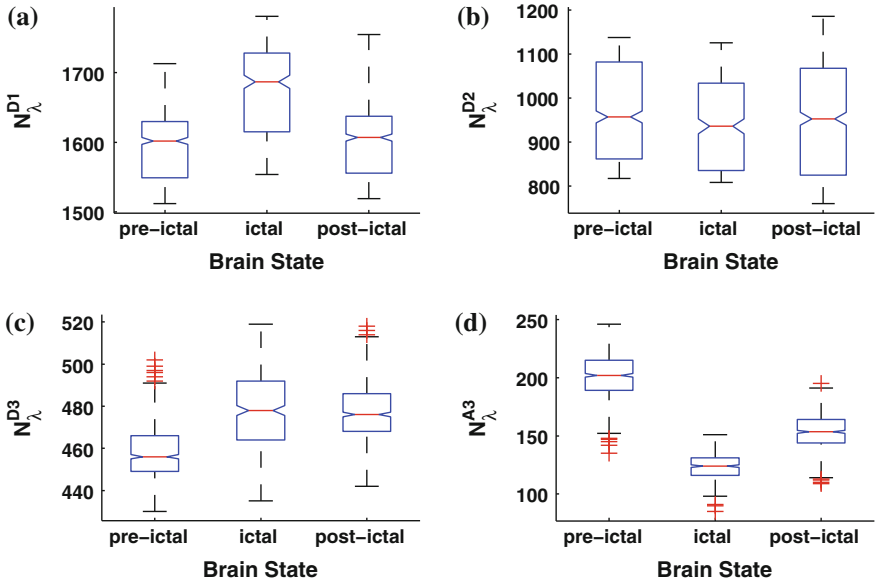


Fig. 10 Comparison of the numbers of local min–max of the EEG subband signals associated with the pre-ictal, ictal and post-ictal states. **a** D_1 subband. **b** D_2 subband. **c** D_3 subband. **d** A_3 subband

D_1 subband, the number of local min–max of the EEG subband signal during an epileptic seizure activity tends to be higher than that of the EEG subband signals associated with the pre-ictal and post-ictal states. On the other hand, the number of local min–max of the EEG subband signal during an epileptic seizure activity tends to be lower than that of the EEG subband signals associated with the pre-ictal and post-ictal states. Also, the number of local min–max of the EEG subband signal associated with the pre-ictal state tends to be higher than that of the EEG subband signal associated with the post-ictal state.

Furthermore, the numbers of local min–max of the EEG subband signals corresponding to only the D_1 and A_3 subbands are simultaneously examined as shown in Fig. 9. This evinces that the EEG signal associated with the pre-ictal, ictal, and post-ictal states can be simply classified using the numbers of local min–max of the D_1 and A_3 subband signals.

3.2 Temporal Characteristics of the Wavelet Subbands of the Epileptic EEG Data

The numbers of local min–max of the wavelet subband signals of all epileptic EEG data associated with the pre-ictal, ictal and post-ictal states of the brain are compared in the box plots shown in Fig. 10a–d, respectively. Similar to that of the chb01_03

subband signals, in the D_1 and D_3 subbands, the numbers of local min–max of the EEG subband signals associated with the ictal state tend to be higher than that associated with the pre-ictal and post-ictal states. However, the numbers of local min–max of the EEG subband signals associated with the ictal state tend to be lower than that associated with the pre-ictal and post-ictal states in the A_3 subband. This suggests that during an epileptic seizure activity there is an increase of amplitude regularity of the highest frequency components of the scalp EEG data (corresponding to the D_1 subband) while there is a decrease of amplitude regularity of the lowest frequency components of the scalp EEG data (corresponding to the A_3 subband). Also, it is observed that in the D_2 subband the numbers of local min–max of the EEG subband signals associated with the ictal state tend to be lower than that associated with the pre-ictal state but higher than that associated with the post-ictal state. The mean and standard deviation of the numbers of local min–max of the wavelet subband signals of all epileptic EEG data associated with the pre-ictal, ictal and post-ictal states are summarized in Table 1.

Table 2 summarizes the mean square, F-statistic and p-value of the ANOVA of the numbers of local min–max of the EEG subband signals associated with various states of the brain in each wavelet subband. From the ANOVA, in any wavelet subbands, the results show that the numbers of local min–max of the EEG subband signals associated with various states of the brain are significantly different (with very small p-values). This implies that the states of the brain have the most significant influence on the temporal characteristics of the epileptic EEG signal as measured using the number of local min–max N_λ in all wavelet subbands. Figure 11 shows the distinction between the numbers of local min–max of the EEG subband signals corresponding to the D_1 and A_3 subbands associated with various states of the brain. It is obviously shown that the EEG signals associated with different states of the brain exhibit distinctive characteristics on the numbers of local min–max in the D_1 and A_3 subbands.

4 Conclusions

In this study, the temporal characteristics of the epileptic scalp EEG signals corresponding to various wavelet subbands are examined using the simple computational measure called the number of local min–max N_λ. The computational results show that the number of local min–max reveals the distinguishing temporal characteristics of epileptic EEG signals associated with various states of the brain corresponding to any wavelet subbands. In particular, the numbers of local min–max of the EEG subband signals corresponding to the D_1, D_2, D_3 and A_3 subbands during an epileptic seizure activity are significantly different from that associated with the pre-ictal and post-ictal states. The different temporal characteristics between the numbers of local min–max of the EEG subband signals associated with the pre-ictal and post-ictal states can also be observed in some wavelet subbands.

Table 1 Statistical values (Mean ± S.D.) of the numbers of local min–max of the epileptic scalp EEG data associated with various states of the brain

Data	Subband	Brain state		
		Pre-ictal	Ictal	Post-ictal
chb01_03	D_1	1628.8182 ± 8.8837	1726.4000 ± 18.1043	1628.7909 ± 15.2003
	D_2	838.1364 ± 11.8867	820.1333 ± 6.2628	790.1636 ± 13.9938
	D_3	450.6455 ± 10.4395	492.5333 ± 14.1122	469.3818 ± 14.0706
	A_3	218.2364 ± 11.6280	119.8000 ± 12.5627	162.0091 ± 18.0689
chb01_04	D_1	1619.2364 ± 10.7958	1750.000 ± 10.5653	1650.3364 ± 30.8962
	D_2	864.9636 ± 13.9020	832.1765 ± 6.2072	828.6091 ± 11.9541
	D_3	455.3909 ± 9.1783	487.3529 ± 8.6958	479.2273 ± 11.0642
	A_3	212.9727 ± 10.5539	112.6471 ± 9.9306	158.7182 ± 15.1186
chb01_16	D_1	1561.4000 ± 26.2186	1742.2927 ± 22.8552	1576.4818 ± 23.8823
	D_2	1052.8455 ± 23.6349	918.6585 ± 17.2534	1040.7636 ± 19.7354
	D_3	460.9182 ± 8.4083	496.6098 ± 11.3905	470.8091 ± 10.7332
	A_3	175.5636 ± 15.5873	117.2195 ± 11.6093	144.0182 ± 15.6428
chb01_18	D_1	1551.0727 ± 13.6164	1644.0750 ± 51.7234	1553.2182 ± 22.0127
	D_2	1093.2364 ± 22.4740	1015.0375 ± 34.8870	1069.7091 ± 25.7302
	D_3	470.1545 ± 12.1513	476.1625 ± 16.7955	475.4636 ± 9.6143
	A_3	191.2545 ± 13.5690	121.0375 ± 11.7865	151.1909 ± 12.4676
chb01_21	D_1	1536.8000 ± 12.2391	1623.3253 ± 49.2612	1544.0182 ± 16.7030
	D_2	1085.2818 ± 17.3560	1040.3253 ± 36.6314	1112.8909 ± 34.8405
	D_3	454.9182 ± 12.8271	467.8916 ± 12.8177	490.6545 ± 12.5949
	A_3	205.9455 ± 12.0266	126.0241 ± 8.6433	153.7727 ± 11.2647
chb01_26	D_1	1649.1636 ± 22.7074	1691.2857 ± 28.3334	1640.8818 ± 25.2901
	D_2	885.3545 ± 22.2598	855.9890 ± 42.7022	860.5091 ± 32.0071
	D_3	455.9455 ± 13.8742	472.9780 ± 19.2954	479.7182 ± 14.7823
	A_3	198.6273 ± 15.3795	127.6044 ± 5.8307	148.2182 ± 8.5288
All data sets	D_1	1591.0818 ± 46.1081	1675.8626 ± 59.2559	1598.9545 ± 48.3989
	D_2	969.9697 ± 110.4626	941.1140 ± 93.1528	950.4409 ± 129.8342
	D_3	457.9955 ± 12.8711	477.7515 ± 18.3233	477.5424 ± 14.1098
	A_3	200.4333 ± 19.4047	123.0117 ± 10.5844	152.9879 ± 15.0979

Table 2 Results of the analysis of variance (ANOVA) of the numbers of local min–max of the epileptic scalp EEG data associated with various states of the brain

Subband	Mean square	F	p
D_1	905185.4396	327.0791	$1.5197e^{-110}$
D_2	304987.8255	25.9382	$1.0269e^{-11}$
D_3	45399.4045	184.3149	$4.3501e^{-69}$
A_3	498740.6501	2472.9429	0

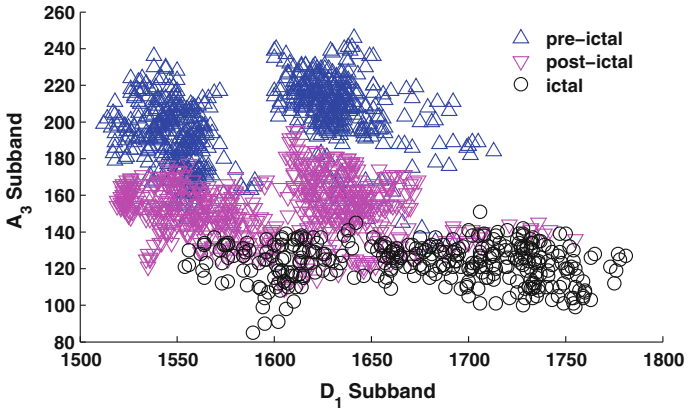


Fig. 11 The numbers of local min–max of the EEG subband signals corresponding to the D_1 and A_3 subbands

In addition, the EEG epochs may be possibly classified into states of the brain, i.e., the pre-ictal, ictal, or post-ictal states, using the numbers of local min–max of the EEG subband signals corresponding to the D_1 and A_3 subbands exhibiting the most remarkable characteristics as the feature vector. By accompanying with further digital signal processing and analysis, the number of local min–max of the wavelet subbands of EEG signal can be potentially useful for epileptic seizure classification and detection. Since this computational analysis technique is simple and requires little computational complexity, it can be implemented for real-time applications in epileptic seizure detection. In future work, in addition to the development for real-time seizure detection, it will be applied to analyze larger sets of epileptic scalp EEG data for more complete outcomes covering a wider range of age and a wider spectrum of types of epilepsy, and also to investigate the effect of region of the brain.

References

1. National Institute of Neurological Disorders and Stroke (2013) Seizures and epilepsy: hope through research. http://www.ninds.nih.gov/disorders/epilepsy/detail_epilepsy.htm. Accessed 8 June 2013
2. Litt B, Echauz J (2002) Prediction of epileptic seizures. Lancet Neurology 1:22–30
3. World Health Organization (2012) Epilepsy. http://www.who.int/mediacentre/factsheets/fs999/en/. Accessed 8 June 2013
4. Goldberger AL (2006) Complex systems. Proc Am Thorac Soc 3:467–472
5. Elger CE, Widman G, Andrzejak R, Arnhold J, David P, Lehnertz K (2000) Nonlinear EEG analysis and its potential role in epileptology. Epilepsia 41(Suppl):S34–S38
6. Elger CE, Widman G, Andrzejak R, Dumpelman M, Arnhold J, Grassberger P, Lehnertz K (2000) Value of nonlinear time series analysis of the EEG in neocortical epilepsies. In: Williamson PD, Siegel AM, Roberts DW, Thadani VM, Gazzaniga MS (eds) Neocortical Epilepsies. Lippincott Williams & Wilkins, Philadelphia, pp 317–330

7. Janjarasjitt S (May 2010) Loparo KA (2010) Temporal variability of the ECoG signal during epileptic seizures. In: The 2010 ECTI International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, Chiang Mai. Thailand 19–21:332–336
8. Janjarasjitt S, Loparo KA (2010) Comparison of temporal variability of epileptic ECoG signals. In: 2010 International Conference on Electronics and Information Engineering, Kyoto, Japan, 1–3 August 2010, pp V2–259-V2–263.
9. Janjarasjitt S, Loparo KA (2010) Investigation of temporal variability of epileptic EEG signals. In: IEEE Region 10 Annual International Conference, Fukuoka, Japan, 21–24 November 2010, pp 359–363.
10. Janjarasjitt S (2011) Investigation of temporal variability of sleep EEG. In: 8th International Conference on Information, Communications and Signal Processing, Singapore, Singapore, 13–16 December 2011, pp 1–5.
11. Janjarasjitt S, Loparo KA, Examination of temporal characteristics of epileptic EEG subbands based on the local min-max, Lecture Notes in Engineering and Computer Science: Proceedings of The International MultiConference of Engineers and Computer Scientists 2012, 14–16 March, 2012, Hong Kong, pp 1095–1099.
12. Janjarasjitt S (2012) Examination of temporal characteristic of sleep EEG subbands based on the local min-max. In: 5th 2012 Biomedical Engineering International Conference, Ubon Ratchathani, Thailand, 5–7 December 2012, pp 1–4.
13. Janjarasjitt S (2013) Classification of temporal characteristics of epileptic EEG subbands based on the local maxima. In: Ao S-L, Chan AH-S, Katagiri H, Xu L (eds) IAENG Transactions on Electrical Engineering, vol 1. World Scientific, Singapore, pp 44–55
14. Janjarasjitt S, Examination of temporal characteristic of wavelet subbands of scalp epileptic EEG based on the local min-max, Lecture Notes in Engineering and Computer Science: Proceedings of The International MultiConference of Engineers and Computer Scientists 2013, 13–15 March, 2013, Hong Kong, pp 657–661.
15. Shoeb AH (2009) Application of machine learning to epileptic seizure onset detection and treatment. Dissertation, Massachusetts Institute of Technology
16. Goldberger AL et al (2000) PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation. doi:10.1161/01.CIR.101.23.e215
17. Mallat S (1998) A wavelet tour of signal processing. Academic Press, San Diego
18. Wornell GW (1993) Wavelet-based representations for the $1/f$ family of fractal processes. Proceedings of the IEEE 81:1428–1450
19. Wornell GW, Oppenheim AV (1992) Estimation of fractal signals from noisy measurements using wavelets. IEEE Trans Signal Processing 40:611–623
20. Vetterli M, Herley C (1992) Wavelets and filter banks: theory and design. IEEE Trans Signal Processing 40:2207–2232

Vibration and Reflection Reduction in Images Captured from Moving Objects Connected Through a Wireless Sensor Network

David Afolabi, Nan Zhang, Hai-Ning Liang, Ka Lok Man, Dawei Liu, Eng GeeLim , T. O. Ting, Yue Yang and Lixin Cheng

Abstract This research explores the use of a computational approach to stabilize image sequences captured by vision sensors mounted on unmanned aerial vehicles (UAV) for road traffic surveillance. The images captured are processed in real-time, and the information necessary to create a feedback to stabilize the flight and enhance the image is computed. The proposed approach is tested on Zigduino, a prototype vision sensor developed for wireless sensor networks used in quadcopter UAV. The results show that our approach can be applied to scenes that have moving objects and dynamic luminance. Although our focus is on UAV for road traffic surveillance, the

D. Afolabi (✉) · N. Zhang · H. N. Liang · K. L. Man · D. Liu · E. G. Lim · T. O. Ting
Xi'an Jiaotong-Liverpool University, Suzhou, People's Republic of China
e-mail: David.Afolabi09@student.xjtlu.edu.cn

N. Zhang
e-mail: Nan.Zhang@xjtlu.edu.cn

H. N. Liang
e-mail: HaiNing.Liang@xjtlu.edu.cn

K. L. Man
e-mail: Ka.Man@xjtlu.edu.cn

D. Liu
e-mail: David.Liu@xjtlu.edu.cn

E. G. Lim
e-mail: EngGee.Lim@xjtlu.edu.cn

T. O. Ting
e-mail: Toting@xjtlu.edu.cn

Y. Yang · L. Cheng
Suzhou Institute of Nano-Tech and Nano-Bionics (SINANO), Chinese Academy of Sciences, Suzhou, People's Republic of China
e-mail: yyang2010@sinano.ac.cn

L. Cheng
e-mail: lxcheng2011@sinano.ac.cn

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 275, DOI: 10.1007/978-94-007-7684-5_6,

results of this research have applications to a variety of moving objects which have mounted cameras to capture images and transmit these wirelessly.

Keywords Glare reduction · Hough transform · LooCI · Traffic surveillance · Vibration reduction · Wireless vision sensor network · Zigduino

1 Introduction

In designing a wireless sensor network (WSN) of unmanned aerial vehicles for traffic surveillance, a number of important issues should be addressed in order to have a robust system that is capable of capturing high quality images regardless of environmental conditions (e.g., wind speed, sunlight intensity / luminosity). Miniature quadcopter drones have been commonly used for surveillance. They can offer added functions over traditional means of stationary mounted cameras such as enabling continuous tracking of multiple vehicles on the road, aiding coordination among multiple drones within a region of interest, and detecting drunk driving and other road traffic anomalies [1].

Due to limited resources in most WSN, such as power, computational capabilities, and transmission bandwidth, the framework needs to be designed based on low-power, low-cost, and commercial off-the-shelf products. To meet these considerations, therefore, the wireless vision sensor network (WVSN) in this research has been developed using the Zigduino board which is an Arduino variant open-source electronic platform that includes an ATmega128RFA1 microcontroller with an in-built 2.4 Ghz IEEE 802.15.4 transceiver. The Zigduino device enables easy troubleshooting of problems with its active developer communities and other current, up-to-date research such as the porting of the Loosely-coupled Component Infrastructure system (LooCI) unto the Zigduino platform. LooCI is a middleware in which distributed component-based WSN application can be built [2, 3].

Whilst a careful selection of hardware components can solve the problem of flight time and bandwidth requirements in the context of aerial road traffic surveillance, the miniature size and elevation of the flying drone makes it vulnerable to vibration and glare. These issues arise from its motor rotation, wind, and reflection of light from the surface of moving vehicles into the vision sensor's field of view (FOV).

2 Problem Statement

The distorted, glary images produced due to the aforementioned environmental conditions and requirements will reduce the ability of a system to automatically identify and distinguish vehicles and other objects. The addition of more hardware to counteract this effect will only further increase the payload of the drone, which is not desirable considering the power needed to carry the load. Therefore, in this research,

we only explore real-time computational approaches which are well suited for the road traffic surveillance applications.

Image stabilization usually involves the use of lens/sensor based stabilization by separating the vision sensor from the main body using vibration isolators, or digital stabilization whereby extra boundary pixels of the scene are captured and used as buffer for portions of the image lost during vibration. A number of solutions to light reflection within the field of view exist such as the use of filters, attaching lens hoods to minimize the scattering of bright light only when it is not within the FOV, and low reflection coated lenses [4]. However, these solutions require the scene to be stationary and therefore are not applicable in this research.

3 System Design and Implementation

The wireless vision sensor network comprises a base station and one or more sensor nodes. In our system, the sensor node is created using the Zigduino microcontroller which operates at a clock speed of 16 MHz on 30 mA during data transmission, and 250 μA when sleeping. A complementary metal-oxide-semiconductor (CMOS) camera module was also connected to the serial port of the Zigduino via a MAX232 chip which converts RS-232 serial signals to TTL digital circuits signals. The base station, as shown in Fig. 1, is a computer which is used to send commands and receive the image or video feeds from the sensor by using an attached Zigduino board to communicate wirelessly over the Zigbee protocol.

3.1 Image Vibration Reduction

Image vibration reduction aims to extract the vibration direction information from images/video feeds that the vision sensor produces and calculates the information required to dampen and reduce the vibration effect. This information can be used as a feedback control to the UAV's motors to fine tuning its roll, pitch, and yaw (See Fig. 2) to reduce the vibration in subsequent images frames.

Fig. 1 Wirelessly paired microcontrollers with a camera module attached

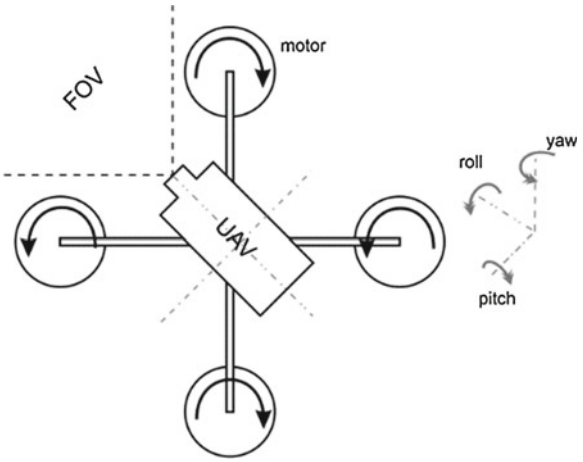


Fig. 2 Diagram of a quadcopter UAV indicating the motions that can cause image distortion

The images captured from a UAV camera in vibration are distorted geometrically by rotation, translation, skewing or scaling. Thus vibration can be classified into *single* and *composite* frequency vibrations [5]. In composite frequency vibrations, the device is subjected to both low and high frequency vibrations. The vibration can be expressed as a sinusoid s(t) as in the following formula:

$$s\,(t) = A\ \sin\left(2\pi f t + \varphi\right) \tag{1}$$

where A is the amplitude of the vibration, f is its frequency, and φ is the phase.

Hough transform is a feature extraction technique that identifies lines in an image by accumulating the parameters ρ (distance) and θ (angle) of data points of the edge detected image into a transformation matrix to determine if each data point and its neighboring points correlate to form a line. Each point has a rho-theta pair defined by Eq. 2 which forms a polar coordinate:

$$r\,(\theta) = x_0 \cdot \cos\theta + y_0 \cdot \sin\theta \tag{2}$$

where r is the distance from the origin to the line and θ is the angle.

The vibration reduction algorithm takes 3 input parameters: (1) a base image (previous frame), (2) the current image, and (3) a de-vibration ratio (i.e. dampening factor). It outputs the amount of vibration detected, a feedback value for a proper image registration, and the corrected image. The following steps describe how the image is processed.

Step 1: Convert base image to grayscale and run canny edge detection to highlight intensity boundaries

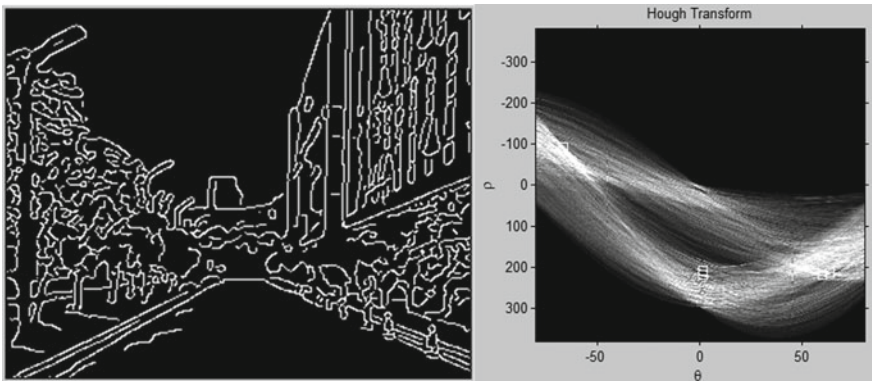


Fig. 3 Hough space graph of the edge detected image (using canny edge detection)

Step 2: Compute the Hough transform over a range of θ (-90 to 89 in steps of 0.25) (See Fig. 3).

Step 3: Select the most prominent line in the image (as illustrated in Fig. 4). If empty, then return the current frame without modification. This is because there are no lines in the image that can be used for vibration reduction.

Step 4: Extract only a subsection in the current image (Fig. 5) using the endpoints from the base image to eliminate the time wasted on computing the other portion of the image.

Step 5: The subsection of current image is also converted to grayscale and canny edge detected to highlight intensity boundaries.

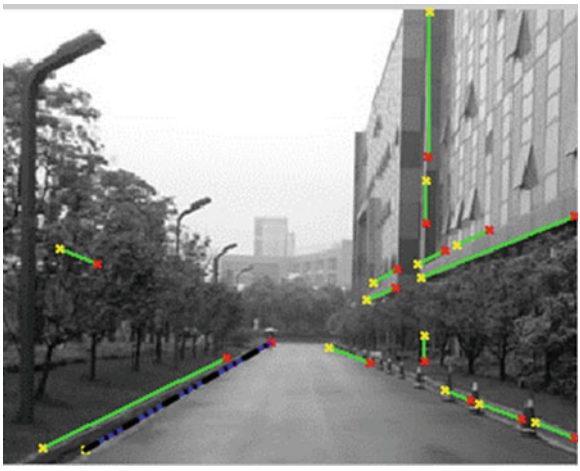


Fig. 4 Image with selected *prominent lines*

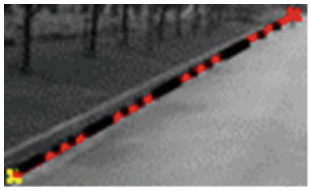


Fig. 5 Subsection of image using endpoints of *longest line*

Step 6: Compute the Hough transform over a range of θ (-90 to 89 in steps of 0.25) of the subsection of the current frame to find the same line from base image.
Step 7: If empty then return the current frame without modification.
Step 8: Calculate the angular difference using the data from the Hough transform matrix of both images.
Step 9: Correct the image's angular vibration by rotating image in an opposite direction of the vibration with a damping factor defined by the de-vibration ratio. (This value should be calibrated so it does not introduce additional vibration).
Step 10: Return the corrected image, feedback value, and amount of vibration detected.

The image vibration reduction algorithm is implemented in MATLAB and the test is carried out with the camera module connected to the USB port. Figure 6 shows the application of Hough algorithm in comparison with MATLAB's built-in image registration function RANdom Sample Consensus (RANSAC) which matches random data points in the base image and subsequent image to detect distortions. The RANSAC function is capable of proper image registration, but it does not perform well when presented with images that have limited information—in this case the subsection of the image.

3.2 Image Glare Reduction

High dynamic range imaging techniques can be applied to reduce glare when taking images of scenes without moving objects, but these techniques are not easily applicable to image acquisition on a UAV in flight. Other related research on active glare removal involves the use of smaller CMOS camera to pinpoint multiple sources of light reflection within the FOV and, in turn, to trigger only selected pixels on an overlaying liquid crystal display (LCD) matrix in order to dim the intensity of the glare sources in real-time [6]. This technique can also be applied to enhance images affected by environmental factors such as water droplets on the lens [7]. Therefore the main goal is to balance areas in the image that have high luminance.

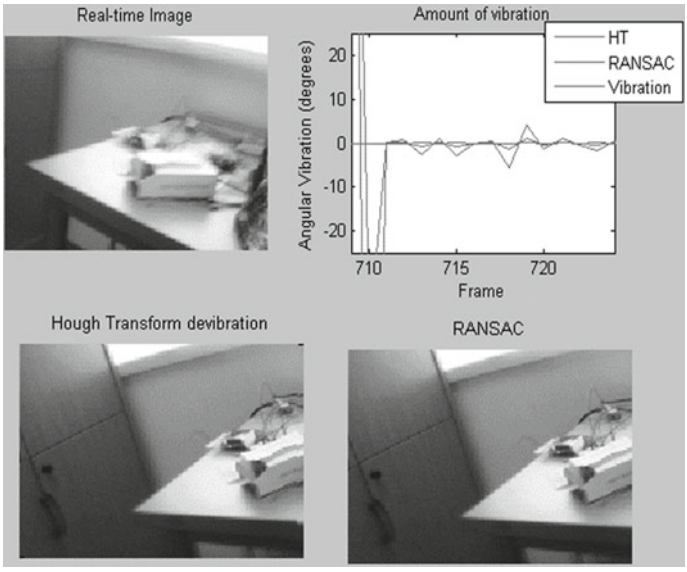


Fig. 6 Application of HT and RANSAC functions to reduce the image vibration of a scene captured with the camera in motion. It shows the input image (*top-left*), the amount of angular vibration detected (*top-right*), and corrected image using both functions (*bottom-left*/*right*)

Luminance is described as the brightness of light reflection on a surface. The mathematical expression of luminance is:

$$L_v = \frac{d^2\Phi_v}{dAd\Omega\cos\theta} \tag{3}$$

where L_v is luminance [candela/m^2], Φ_v is luminous flux, θ is the angular difference between the specified direction and the surface normal, A is the surface area, and Ω is the solid angle.

The luminance component of an image can be extracted and enhanced without affecting the chroma component (i.e. colour information) by transforming the image into the YCbCr colour space (See Fig. 7). The contrast adjustment depends on the amount of glare detected. The Y, Cb, and Cr components of this colour space represent the luma, blue-difference and red-difference respectively, which can be derived from an RGB image using the equations below.

$$\begin{aligned} Y &= 0 + (0.299 \cdot R) + (0.587 \cdot G) + (0.114 \cdot B) \\ Cb &= 128 - (0.16736 \cdot R) - (0.331264 \cdot G) + (0.5 \cdot B) \\ Cr &= 128 + (0.5 \cdot R) - (0.418688 \cdot G) + (0.081312 \cdot B) \end{aligned}$$

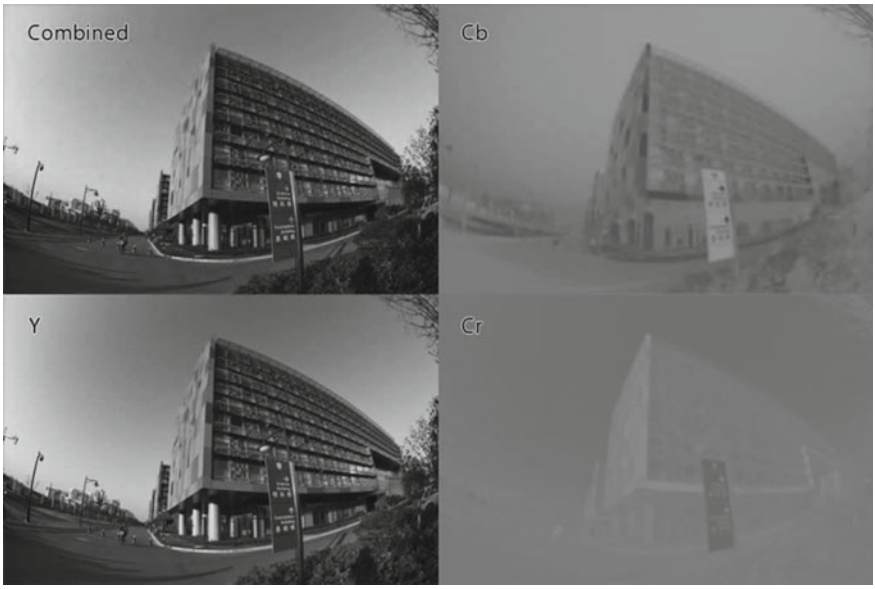


Fig. 7 An image with its Y, Cb, and Cr components

```
Algorithm Reflection reduction
  Input: image
  Output: Corrected image
 1: threshold = mean of dynamic range of luminance
                component from image
 2: for (each pixel)
 3:  if(pixel[i,j]> 1.5*threshold) then
 4:     mark cell and some neighbouring cells
 5:  end if
 6: end for
 7: if (percentage of marked cell>15%)
 8: convert image to YCbCr
 9: correct Y component by balancing luminance
10: return corrected image
```

The glare reduction can be implemented using MATLAB as shown in the algorithm above.

4 Simulation Results

First, the vibration reduction algorithm was tested by subjecting the camera to constant vibration and the result was computed and plotted in real-time (See Fig. 8). In the experiment, the images were acquired at 10–25 frames per second, with a damping factor of 0.75. The graph shows the amount of vibration detected in the original image (i.e. line with large variation) and the reduced vibration in the corrected image

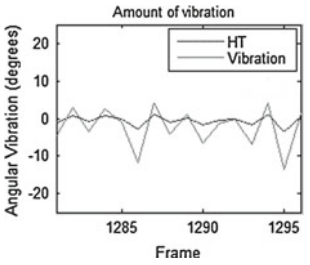


Fig. 8 Hough transform de-vibration graph

(i.e. line with relatively balance trend). The results of the simulation revealed that Hough transform algorithm could potentially be used for as a real-time feedback mechanism to balance the vibration of the UAV through the visual cues from the environment.

Second, both indoor and outdoor tests demonstrated how the glare reduction algorithm was able to detect the dynamic glare when the camera was in motion (See Fig. 9). The 15 % glare threshold was derived heuristically; nonetheless, it could also be adjusted to fit either day or night conditions. Significantly positive results were obtained with most images except for those with high percentage of white region.

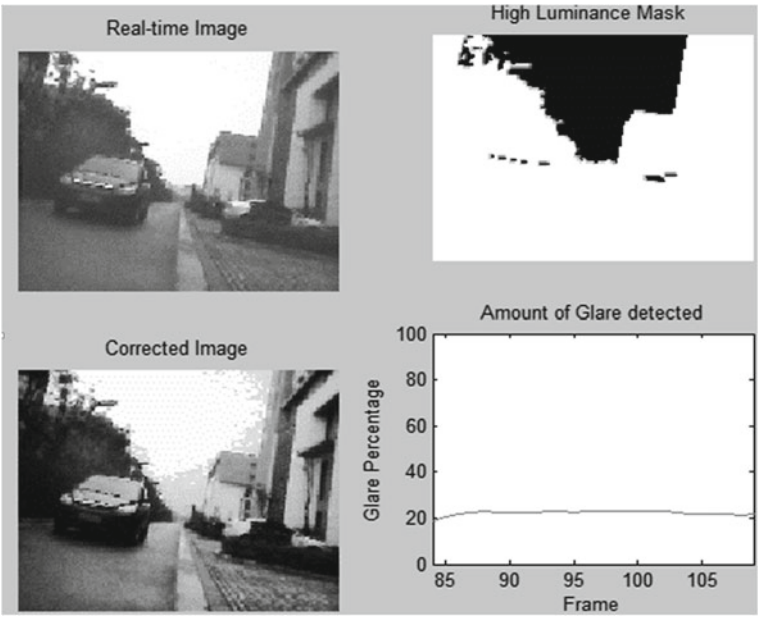


Fig. 9 Outdoor test of glare reduction showing the input image (*top-left*), high luminance spots (*top-right*), corrected image in real-time (*bottom-left*), and percentage of glare detected by the image sensor (*bottom-right*)

5 Conclusion and Future Work

In summary, this research has a twofold purpose: (1) vibration reduction, and (2) glare reduction. Both of them are in the context of images taken from cameras placed on moving objects and capturing images at a distance. The vibration reduction algorithm using Hough transform provides an effective way to reduce the vibration of unmanned aerial vehicles (UAV) in real-time through a feedback mechanism, consequently reducing the vibration that would cause noise in subsequent images. This proposed algorithm also reduces the computational time by removing sections of the image that are not necessary for computing the de-vibration data. The glare reduction algorithm is able to detect the luminance of a dynamic scene and correct it. Thus these two solutions fit the requirement of a real-time image correction system for use in road traffic surveillance with quadcopters.

The use of the Zigduino platform to create the wireless vision sensor network (WVSN) for the network of drones provides easy extensibility by taking advantage of the LooCI middleware which has been successfully ported unto this hardware. Further research based on the use-case proposed in this research and the LooCI system can be focused on environmental monitoring, hazard detection and ambient energy scavenging to power the WSN devices [3]. Although our focus is on UAV for road traffic surveillance, the results of this research have applications to a variety of moving objects which have mounted cameras to capture images of distant objects and transmit these wirelessly within a network.

Acknowledgments This work was partially supported by Xi'an Jiaotong-Liverpool University (Suzhou, China) Research Development Fund under Grants RDF10-01-27, RDF10-02-03, and RDF11-02-06; Xi'an Jiaotong-Liverpool University Summer Undergraduate Research Fellowship under Grant 201328; and Transcend Epoch International Co. Ltd., Hong Kong.

References

1. Afolabi D, Man KL, Liang HL, Lim GL, Shen Z, Lei C, Krilavičius T, Yang Y, Cheng L, Hahanov V, Yemelyanov I (2012) A WSN approach to unmanned aerial surveillance of traffic anomalies: some challenges and potential solutions. In: 10th East-West Design and Test, Symposium 2012 (EWDTS'12)
2. Shen Z, Man KL, Liang H-N, Zhang N, Afolabi D, Lim EG (2013) Porting LooCI component into Zigduino. In: Proceedings of elsevier procedia computer science series of the first international conference on information technology and quantitative management (ITQM 2013), 16–18 May 2013
3. Afolabi DO, Shen Z, Man KL, Liang H-N, Zhang N, Lim EG (2013) Modelling and Analysis of LooCI models in Zigduino. In: Lecture notes in engineering and computer science: proceedings of the international multiconference of engineers and computer scientists 2013, Hong Kong, pp 713–715, 13–15 March 2013
4. Rouf M, Mantiuk R, Heidrich W, Trentacoste M, Lau C (2011) Glare encoding of high dynamic range images. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 289–296

5. Wang X, Yi T, Tang Q, Feng L, Ni G, Zhou L (2010) Simulation and analysis of vibration blurred images. In: 6th international conference on wireless communications networking and mobile computing (WiCOM), pp 1–4, 23–25
6. Bhagavathula K, Titus AH, Mullin CS (2007) An extremely low-power CMOS glare sensor. IEEE Sens J 7(8):1145–1151
7. Hara T, Saito H, Kanade T (2009) Removal of glare caused by water droplets. In: Visual media production (CVMP '09), pp 144–151

Specifying Resource-Centric Services in Cyber Physical Systems

KaiyuWan, VangalurAlagar and Yuji Dong

Abstract A service-oriented view of Cyber-Physical Systems (CPS) is a good platform for managing global supply chain management, service acquisition and service provision. A necessary condition for complex service delivery is that resources required for complex services are of high quality and are available at service execution times. Therefore in a resource-centric service model, both resource quality and service quality using that resource are explicitly stated. In this paper a cascaded specification approach is discussed for describing resource types, services offered by resource, and a cyber configured service (CCS) that package physical services. Energy support from internal-combustion engine is regarded as a resource-centric complex service and discussed as a case study to illustrate our specifying and modeling approach.

Keywords Cyber physical systems · Resource · Resource description · Resource management · Resource specifying · Service model

K. Wan (✉)
Department of Computer Science and Software Engineering, Xi'an Jiaotong-Liverpool University, Suzhou, China
e-mail: kaiyu.wan@xjtlu.edu.cn

V. Alagar
Department of Computer Science and Software Engineering, Concordia University, Montreal, Canada
e-mail: alagar@cs.concordia.ca

Y. Dong
Research Assistant at Department of Computer Science and Software Engineering, Xi'an Jiaotong-Liverpool University, Suzhou, China
e-mail: wildfire1106@gmail.com

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 275, DOI: 10.1007/978-94-007-7684-5_7,

1 Introduction

Cyber-Physical Systems (CPS) [1] is a new research area with a grand vision. This paper is a contribution to formally specify resources and resource-centric services for CPS. The term *resource* is used in a generic sense to denote an entity that is relevant in either producing or consuming a service. In CPS, physical devices are resources, which are hence first class entities. Services may be either generated or consumed by physical devices, which might in turn be consumed by cyber computational resources, such as communication protocols. Software services may be generated by the computational resources that reside either in a static or dynamic host computer in CPS network and may be consumed by other physical devices to make changes in the environment. In general, a CPS resource might offer many services, a CPS service might require several resources, a CPS resource might *use* other resources, and a CPS (complex) service may be produced by combining several services and resources. Thus the service-oriented view of CPS is more complex than the service-oriented view required for traditional business applications, as discussed in SOC literature [2].

In [3] we proposed the three conceptual layers of CPS resources as *physical*, *logical*, and *process* through three-tiered approach, as illustrated in Fig. 1. In this paper, we continue our research and strive for notations that have semantic consistency across these layers. That is, we design the description language for resources, and services. The three important characteristics of the language are: (1) The published

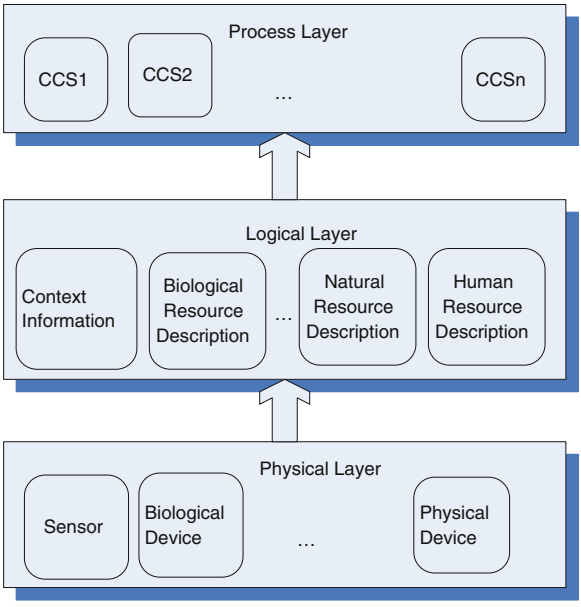


Fig. 1 Three-tiered architecture for CPS

resource or service description, intended for clients, has information completeness, consistency, and correctness. (2) It should be possible to create precise formal descriptions of published service descriptions. Formalized service descriptions are not for public consumption, and are used by the service provider only for the purpose of validating the published descriptions and the demand-response model (DRM) (when their behavior models become available). (3) The service descriptions are modular, and declarative. Complex descriptions are assembled by putting together simpler descriptions, supported by strict semantics. This specification approach imposes some uniformity of resource description across CPS sites.

Throughout the paper we suggest the underlying formalism without being formal. In Sect. 2 we explain our view on service model. In Sect. 3 we discuss resource types, and a generic resource description template. This notation is suggested for modeling resources at the physical layer. The merits in our approach are brought out through a brief comparison with other resource modeling approaches. In Sect. 4 we give resource class specifications, that resemble Larch [4] specifications. A resource class specification will include a resource type description, and it is extensible. This notation is suggested for modeling resources at the logical layer. In Sect. 5 we give a template for service description, which will include the resource description classes for all resources used by the service. We explain the significance of the service description template, and how it can be analyzed for quality claims. This notation is suggested for modeling resources at the process layer. We conclude the paper in Sect. 7 with a brief summary of its significance and our ongoing work.

2 Abstract Service Model

Abstractly, the three major stakeholders in CSP are *Resource Producer* (RP), *Service Provider* (SP), and *Service Requester* (SR). A SP may interact with one or more RPs and one or more SRs. A RP may not be *directly visible* to any SR in the system. So, a SR gets to know about resources used for service composition and delivery only from the service descriptions posted by the SPs. In this abstract CPS model shown Fig. 2, every RP creates a resource model for each resource in its ownership and publishes it to all SPs who subscribe to its services. Thus, it is a comprehensive description of the physical, logical, and process layer needs. This specification will enable the SPs conduct a static analysis of published resource descriptions and request their distribution across CPS nodes in a demand-driven fashion.

Once the resource model is published by a RP, the SPs who are clients of RPs will have an opportunity to independently verify the claims made in service descriptions before selecting it for use in the services created by them.

A SP creates service descriptions for services provided by it. A service description includes the functionality of the service, its non-functional properties, a list of resources used in creating and delivering the service, and a service contract. A SP publishes service descriptions and make them available to SRs who subscribe to its services. The SP guarantees the quality of service through a list of *claims*, which

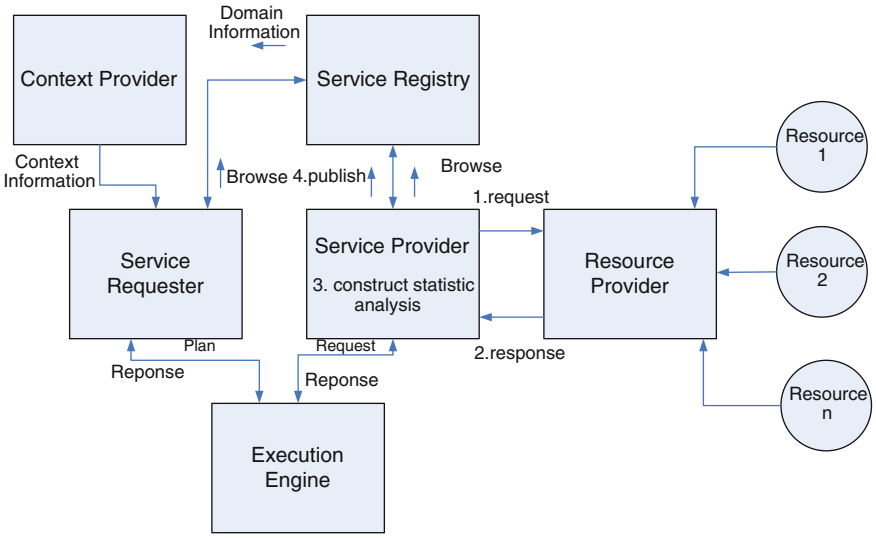


Fig. 2 Resource-centric abstract service model

should be validated by the SP when challenged by the SRs. A SR creates a demand model of service. This model is very much dependent upon the application.

Satisfaction Criteria

Therefore, in order to have matched CPS services the two essential conditions are

- *Provided-by*(RP_q) *SAT Required-by*(SP_q)
- *Provided-by*(SP_q) *SAT Required-by*(SR_q)

where *Provided-by*(X_q) means the 'quality attributes provided by the entity X', *Required-by*(Y_q) means the 'quality attributes required by the entity Y', and *SAT* is the 'satisfaction relation'. So we posit that the resource model should include *Provided-by*(RP_q), and the service model should include *Required-by*(SP_q), *Provided-by* (SP_q), and *Required-by*(SR_q). We assume that a SP, by whichever *Required-by*(SP_q) model it has, will select the resources in order to satisfy the relation *Provided-by*(RP_q) *SAT Required-by*(SP_q). We assume that a SR, by whichever *Required-by*(SR_q) model it has, will select the services in order to satisfy the relation *Provided-by*(SP_q) *SAT Required-by*(SR_q). Thus, the resource description should enable a formal execution of the *SAT* relation. Typical *SAT* relations are *implies* ($\rightarrow$), and *includes* (subset relation $\subset$)). These are resolved using Logic and Set Theory provers.

3 Physical Description Layer

In this section we discuss the attributes for modeling resources in the physical layer. The model that we create is called Resource Description Template (RDT). We may assume that CPS resources are categorized so that all resources in a category are of the same *type*. One such classification is *human resources*, *biological resources*, *natural resources*, *man made resources*, and *virtual resources* [3].

In general, let $\mathcal{RT}$ be a finite set of resource types. The semantics for each resource type is to be understood from its domain. A resource type $T \in \mathcal{RT}$ is a finite collection of resources of that type. As an example, *Metal* is a resource type, and $\{gold, platinum, iron, copper, zinc\}$ are resources of type *Metal*. The description of one resource r_T of type T is a RDT whose structure is shown in Table 1. The RDT table may be extended by adding more element descriptions. We are suggesting that a *model-based formal specification* may be attempted, given the RDT structure, the choice of its description parameters and their types. The tabular RDT format shown in Table 1 is meant for human agents. An XML version of the RDT is automatically generated from the RDT and is used for resource propagation across CPS processing sites. CPS sites will subscribe to the sites of Resource Providers (RP) in order to receive their RDTs and their periodic updates. Published resource types can be searched at service execution times, in order to get the most recent resource that best fits the service requirements.

We have explained the semantics in details in [5], therefore below we explain the semantics briefly.

Table 1 Resource description template

Resource: ⟨*generic description of resource name*⟩
Type: ⟨*resource type*: T⟩
Attribute: ⟨*producer, production facility profile, quality attributes*⟩
Properties: {*physical properties, chemical properties,temporal properties, trustworthiness properties*}
Utility: $\{\langle a_1, u_1\rangle, \langle a_2, u_2\rangle, \ldots, \langle a_k, u_k\rangle\}$
Cost: *cost per unit*
Availability: *available for shipment to all parts of the world or state constraints*
Sustainability: *ratio of demand to supply for the next x years*
Renewability: Reliable period of resource supply
Reuse: *list of applications for reuse of this resource*
Recycling: method names and technology used
Legal Rules for Supply: *URI to a web site*
Other Resources in the Context of Use: *a set of contexts suggesting resource dependencies*
Side Effects: *health and environmental protections*

1. The *Type* of a resource is the *resource category*, as classified earlier or given in industries. We can include more resource types, such as Health (Medical) Resources.
2. The *Attribute* section is used to provide the identity and contact information of resource producer. A general yet concise description of the resource may also be included. Some examples may include *human resources, biological resources, natural resources,man made resources, and virtual resources*.
3. The *Properties* section might include *physical* properties, *chemical* properties, *temporal* properties (persistent or change with time), and *trustworthiness* properties.
4. The utility factor for a resource defines its relevance, and often expressed either as a numerical value $u, 0 < u < 1$, or as an enumerated set of values {*critical*, *essential*, *recommended* }. In the former case, a value closer to 1 is regarded as critical. In the later case the values are listed in decreasing order of relevance. A RP may choose the representation $\{\langle a_1, u_1\rangle, \langle a_2, u_2\rangle, \ldots, \langle a_k, u_k\rangle\}$ showing the utility factor u_i for the resource in application area a_i for each resource produced by it. The utility factors published by a RP are to be regarded as recommendations based on some scientific study and engineering analysis of the resources conducted by the experts at the RP sites.
5. The semantics of *Cost* is the price per unit, where the unit definition might vary with resource type. For example, for natural gas the unit may be 'cubic feet', for petrol the unit may be 'barrel or liter'.
6. The semantics of *Availability* is information under the three categories (1) Measured (provable), (2) Indicated (probable), and (3) Inferred (not certain).
7. The semantics of *Sustainability* is related to *Reserves*, *Contingent*, and *Prospective*. *Reserves* expresses a comparison between the measured amount of resource with the current demand. *Contingent* is an estimate (both amount and time period) of getting the reserves (this is a certainty). *Prospective* specifies the resource quantity determined, and an approximate time scale for its availability.
8. The semantics of *Renewability* is related to the 'perpetual' or 'migratory' nature of the resource. For example, 'solar power' resource can be labeled 'perpetual'; however 'ground water' resource may not be available for ever.
9. The terms *Reuse and Recycling* are well understood both in technology and in environmental applications.
10. The semantics of *Legal Rules* include the business rules of the RP, the government regulations governing the distribution of resources, and international rules regarding quality of resources.
11. The meaning of *Other Resources in the Context of Use* is to express 'resource dependency'. Examples of dependencies may be expressed using *before*, *during*, and *following* temporal operators.
12. The intent of *Side Effects* section is to list the impact and interference effects with environment.

We compare our RDT with the UML modeling approach [6], RDF [7], the Resource Space Model (RSM) [8], the entity-relationship model [9], and Service-Oriented Middleware Architecture (WebMed) [10].

- Modeling resources with UML is the first method proposed with respect to modeling run-time resources for real-time systems. Resource properties and resource dependencies are not part of the model, however resources required for a service can be modeled. The model is service-centric and not resource centric. It might be possible to develop resource-centric UML models at all levels, but we have not attempted this. Given the distributed nature of the resources in CPS it might be hard to manage and use UML models.
- RDF is meant to describe Web resources, which according to our classification are *Virtual resources*. Since RDT is meant for all types of CPS resources we expect that all Web resources can be represented as RDTs. We have not come across RDF examples which include all RDT aspects. In particular, it is not clear as to how Availability, Reuse, Legal Rules for Supply, and Context of Use can be specified in RDF.
- The RSM method considers the resource space as multi-dimensional where each dimension is a resource type. So, essentially RSM produces a model at the logical layer, however it is oriented towards an application. RSM is not resource-centric and its models require a centralized management in order to avoid inconsistencies.
- The Resource-Explicit Service Model (RESM) proposed in [9] is similar to an Entity Relationship (ER) diagram. They consider physical devices as resources, and model resources, the services offered by them, and the service contexts as a bundle in a single ER diagram. This approach suggests that resources and services should be modeled together although the emphasis is on resources, and services offered by the resources always match with the services required by a consumer. A soft real-time application in CPS requires an open market approach. An open CPS network is a loosely coupled system, and it is best to avoid tight coupling between resource and service models. A service provider in CPS should be free to choose the best resources in order to fulfill a service request.
- WebMed is a early conceptual middleware designed with a service-oriented view point to support CPS applications. It enables access to the underlying smart devices and integration of its device specific functionality with other software services through five components such as WebMed node, Web service enabler, service repository, engine, and application development. WebMed tries to provide an easy interface with physical devices. However it didn't support resource modeling and management.

Our modeling approach emphasizes separation of concerns and modularity. An RDT is created by a RP independent of a service that might be created by a Service Provider (SP). A modification to a RDT produces a new RDT which is published by the RP and can be acquired by SPs in the CPS network. The RDT notation is suitable for the physical layer. The logical and process layer modeling include the RDTs, and thus the resource specifications are modular. The Reuse section in RDF adds one more level flexibility by explicitly stating the alternate uses of a resource. The

RDT notation is richer than other notations, because it allows 'user defined types' to be introduced with their semantics and operations. In Table 2 we show the RDT model for *gasoline 97* resource, which is considered a resource for energy support from internal-combustion engine at cars.

4 Logical Layer Description

For the resource-centric CPS model we need to follow the resource-centric service approach. In our approach, the activities in the service are ordered, and the list of activities per single resource are handled taking into account resource dependencies.

Table 2 RDT for gasoline 97

Resource: ⟨*description of unleaded gasoline 97*⟩
- Type: ⟨*man made resource* ⟩
- Attribute:
  - *Opet product*
  - *refuelled at XYZ gas station*
  - *passed ISO 14064-1 quality management system*
- Properties:
  - Appearance: Clear and bright
  - Density(@15°C): 770–775 kg/m^3
  - Vaporization percentage @100°C: 46–71 % volume
  - Vaporization percentage @150°C: 75 % volume
  - Final boiling point: 210°C
  - Distillation residue: 2 % volume
  - Oxidation stability: 360 min
  - Research octane number RON: 97,0
  - Motor octane number MON: 86,0
  - Lead: 5 mg/L
  - Sulfur: 10 mg/kg
- Utility:
  - ⟨*combustion in engine to supply heat, 1*⟩
- Cost: $1.43/litre
- Availability: gas stations supported by Opet
- Sustainability: 100 % in reasonable years
- Renewability: *NO*
- Reuse: *automobile engine*
- Recycling: *NO*
- Legal Rules for Supply: *URI to a web site*
- Other Resources in the Context of Use:
  - *needs related equipments to store and transfer gasoline*
  - *needs an engine to convert chemical energy to kinetic energy and electrical energy*
- Side Effects: *gases from combustion, CO^2 for greenhouse effect, oxide from Lead, Sulfur, etc. for air pollution*

Table 3 Syntax for RCS

Resource Class RC
includes *RDT r*
requires {*RDT* $r_1, \ldots,$ *RDT* r_k}
consumed-by
$\{\tau_1, \ldots, \tau_n\}$
constraints
$\{\sigma_1, \ldots, \sigma_m\}$

A specification for each resource in which the dependencies on other resources and the tasks that can be done with that resource are listed. This is the logical view and we call this specification a Resource Class Specification (RCS). To realize the resource-centric model of CPS it is necessary that every CPS site publishes the RDTs of resources owned (or produced) by it as well as the RDTs acquired from other RPs, develop a mechanism for allocating resources in different service request contexts, and create a RCS.

The structure of RCS, shown in Table 3, resembles a Larch trait [4]. The semantics of Larch is adapted to give semantics to the different clauses in a RCS. Thus, the meaning of the different clauses in it are as follows: (1) The clause *Resource Class* introduces the name RC of the specification. (2) The **includes** clause states that the RDT defining resource r is specified in RC. The effect is that all the information in the included RDT is exported to this specification. (3) The **requires** clause specifies a list of the resources that are *packaged* together with r. These resources are necessary to make r operational. This list may be empty, in which case the resource r is self-sufficient and will be exported to service execution phase. If the list is not empty, the resource r together with all the resources included in this list will be exported as a package to a service. Note that, the resources included in the section "Other Resources in the Context of Use" of the RDT r are required for a service that requires r, in addition to the resources listed in the **requires** clause. (4) The clause **consumed-by** lists the tasks or resources for which r may be needed. Each task listed in this clause is an atomic activity belonging to at least one application domain listed in the Utility section of the RDT r. (5) The **constraints** clause lists *resource* constraints, *compatibility* constraints, and *dependency* constraints. Resource constraints are dependent upon the type of resource r and the context of its use. They may include *minimum* and *maximum* units of resource r that will be available in specific contexts, and a list of byproducts arising from the use of resource r. The compatibility constraint is a relationship between the resource r and the tasks consumed by it. That is, resource r is compatible with two tasks τ_1 and τ_2 if they can share the resource, therefore, both these tasks can be concurrently processed. The dependency constraint can be a relationship between two resources listed in **requires** clause or it can be a relationship between two resource class specifications. In the former case we include the dependency constraints, written $\tau_i \ll \tau_j$, in the **constraints** clause. To describe the later case let us assume that RC_1 and RC_2 are resource class specifications for resources r_1 and r_2. Suppose there exists a context c in which the resource

Table 4 Gasoline 97 resource class specification

Resource Class Gasoline97Class
includes *RDT gasoline97* (Table 2)
requires
{*RDT Storage*; *RDT Transfer*; *RDT Filtered_Air*}
consumed-by
{ *Mix_With_Air*, *Compression_stroke*, *Power_stroke*, *Exhaust_stroke* }
constraints
{
resource constraints:
high quality requirement (industrial standard)
correct mix ratio with clean air
compatibility constraints:
⟨*Mix_With_Air*, *Compression_stroke*, *Power_stroke*, *Exhaust_stroke*⟩
dependency constraints:
Filtered_Air Class $\xleftarrow{combustion}$ *Gasoline_97 Class*
}

r_1 should be used *before* resource r_2 is used, then the class RC_2 is dependent on class RC_1. That is, all tasks listed in RC_1 must be completed before starting the tasks in RC_1. We use the notation $RC_1 \overset{c}{\leftarrow} RC_2$ to show class dependency and include it in **constraints** clause of RC_1. Class dependencies are local to a site where resources are produced. A class specification for robot RDT (Table 2) is shown in Table 4.

5 Process Layer Model

In our resource-centric service model, resource class specifications are included in configuring and composing service specifications. The first step for SP is browsing the sites of those RPs, examining the RDTs published by them, and then selecting the RCSs published by them. The second step is that the SP selects the RPs from whom the RCSs can be bought. The final step for SP is to create services that can be provided by putting together the atomic tasks in the RCSs. We introduce the *CyberConfigured-Service* (CCS) notation for this purpose. In CCS the service with its contract, quality assurances, and other legal rules for transacting business are included. Such configured services are published in the site of the SP. We define a *CyberConfiguredService* (CCS) is a service package that includes all the information necessary that a service requester in CPS needs to know in order to use that service. It will include (1) service functionality, (2) a list of resources used to create the service, together with resource specifications, (3) nonfunctional attributes of service, (4) quality attributes of the service, and (5) contract details. Legal rules, context information on service availability and service delivery, and privacy guarantees are part of contract details. The service and contract parts are integrated in CCS, and consequently no service exists

Table 5 Engine CCS

Service	Function:	*Name:* Gasoline_97 CCS *Pre:* check(quality) $\wedge$ check(quantity) check(equipments_specify) $\wedge$ burn(gas,air,energy) *Post:* Mix_with_FilterAir $\wedge$ MechanicalEnergy
	Resource:	*Resource Class:* Gasoline_97 CCS *Provider Data:* Company Name: Opet
	Non functional:	*Gasoline Cost:* \$1.43/litre *emission:* CO^2
Contract	Trust attributes:	*Resource:* *Safety:* industrial standard *Security:* amount calculating and error monitor *Reliability:* no record of malfunction *Availability:* 99.9999% *Service:* *Security:* intelligent control system *Availability:* 99.99% *Provider:* Consumer rating: $\frac{4.1}{5}$ Organization rating: 5 $\star$ recommendation awarded by BBB
	Legal exceptions:	*Liability insurance:* not covered for intentional injury or Non-Countervailable accident *Renewal of Contract:* not automatically renewable *Maintaining:* must be maintained at the official authorizing place *Refund:* damages and depreciations considered
	Context	*Context Info:* Provider: [LOCATION:Shanghai] Execution: [Time:contract time(date)] *Context Rule (Situations):* Consumer Related: related manufacture must compatible to this engine Delivery Related: free shipping for places within 100 kms from Shanghai for other places the shipping charge should be paid by the consumer

in our model without a contract. The contract part in CCS includes QoS contract *Provided-by*(SP_q) as well as the QoS contract *Provided-by*(RP_q). These contracts must be resolved at service discovery and service execution times. The structure of CCS for Engine CCS is illustrated in Table 5.

Complex CCS Representation We take the energy support from internal-combustion engine as an example and illustrate the CCS specification accordingly. In reality the energy support system from internal-combustion engine in vehicle is a extremely complex service. For simplicity, we model the service involving fuel supply, air supply, working engine to convert energy, electrical system to help engine and get

Table 6 Syntax for creating energy support from internal-combustion engine CCS

Service Name: *EnergySuportMission CCS*

includes *CyberConFiguredServices, FuelCCS, EngineCCS, SensorCCS AirCCS, ElectricalCCS, AgentCCS, TransferCCS, ProtectionCCS*
extensions {
⋮
}
modifications{
⋮
}

energy from engine, related manufactures to transfer fuel, air, heat and electricity, sensors and agent system to measure and make decision, and protection system to deal with error. So, there are eight services required for the energy support. The SP who offers the energy support service creates the eight configured service specifications including *FuelCCS*, *AirCCS*, *EngineCCS*, *SensorCCS*, *ElectricalCCS*, *AgentCCS*, *TransferCCS* and *ProtectionCCS* and puts them together as shown in Table 6. The semantics of the specification in Table 6 is the following: In the **includes** clause the CCSs that are necessary for the complex service are listed. The **extensions** clause will include additions to the non-functional and trust attributes of the included CCSs. The **modifications** clause will list changes and additions to the contract part of the included CCSs. We emphasize that no change will be made to the functionality of the included configured services and the resources used to produce them. In essence, the syntax in Table 6 is intended to be used by SPs in the service execution layer.

We decide to develop them separately. One rationale is *reuse* potential, in that each CCS can be used individually in other service creations. For example, there are many different kinds of fuels used in different kinds of environments while the fuel supply service might be quite similar. The service structures can be used in same way but different places by just relying on different resources. Second, they can be combined in many ways dynamically, as and when a service provision context arises. In the case of the energy support system in vehicle from internal-combustion engine, all the eight CCSs are required. In another situation when energy support system is required at aircraft perhaps these eight CCSs are not sufficient. The SP may need to add more modules such as *BalancingCCS* and *CoolCCS* into the entire energy support system. Thus the SP can create a complex service using the two additional *BalancingCCS* and *CoolCCS* together with the eight CCSs, by modifying the contract part of the structure. Furthermore, *EngineCCS*, *SensorCCS*, *ElectricalCCS*, *AgentCCS*, *TransferCCS*, *ProtectionCCS* and other specific CCSs can be included to model and specify some other CPS systems like intelligent irrigation system, robots rescue systems, intelligent Manufacturing systems etc.

6 Toolkit Implementation

The semantic basis is necessary for developing tools. We are currently developing a Graphical Resource Editor (GRE). The goals are (1) to provide assistance to developers in creating the specifications at the three layers, (2) to automatically generate XML files that can be shared by CPS nodes, and (3) to enable a formal resolution of SAT claims by providing links to other verification tools. The GRE tool that we have completed enables the creation of RDTs and their XML versions.

According to the formal definition and semantics for resources that we proposed, we implemented a graphical user interface (GUI) at each CPS site for humans and systems to interact, share, and yet securely manage resources across CPS [11]. The rationale for GUI is that humans who manage resource will find it user-friendly, and the mechanism that we build behind RDT will faithfully transform resource information in languages that can be shared and communicated securely across the CPS. A Framework for GUI which supports RDT has been discussed in [11]. However Resource management is a multi-step activity. The implementation is only the first step. We may consider Resource Discovery, Resource Acquisition, Resource Modeling, Resource Publication, and Resource Allocation as the distinct layers in resource management. Therefore a complete GUI should be implemented to fulfill resource management.

Some of the benefits and key features in the design of GUI are as follows:

1. Comprehensive and complete visibility of resource availability, resource requests, and resource allocation is possible for the business enterprise.
2. Local and global pool of resources can be assessed for a project, regardless of physical location.
3. Within GUI, it is possible to view and manage resource utilization by graphical plug-ins.
4. Modifications to resource bookings can be monitored and dealt with by real-time reallocation of resources to other projects.
5. Security settings of key aspects of resource knowledge can be distributed across the multiple layer.

7 Conclusion

In this paper we proposed a model-based language to specify resource and resource-centric services through three-tiered approach. The RDT table structure, given its semantics, can be turned into a lightweight formal description. We import the RDT specifications within resource class specifications which are written in Larch style. Cyber configured service specifications are also declaratively written in Larch style. Thus RDT, RCS, and CCS all have set theory and logic semantic basis. Moreover context formalism is also founded on relational semantics. Therefore, the semantic basis in a tier is consistent with the semantic basis in all tiers below. Consequently

formal validation of service claims are possible if they are stated in first order logic. Thus a claim verified in a tier is not contradicted in higher tiers. We have suggested rigorous methods for evaluating three kinds of SAT relations [3]. We are still working on this aspect. To validate the quality claims made for resources themselves we would need scientific evidence and engineering analysis of their respective resource types. For example, the precision and accuracy of energy consumption in a vehicle would need an analysis based on the mechanics behind the design of the energy support system. So, validation issues are hard to tackle; however, the specifications suggested in this paper will enable the claims to be stated formally, a first step towards validation. Clearly, verifying resource claims is a broader challenge which needs investigation by domain experts. Since all the quality claims of resources may not verifiable using software, a tight coupling exists between what experts can do and what machines can be made to do.

Protecting resources, assuring confidentiality in service provision, and privacy of CPS clients are the three challenges to be faced in making CPS survive attacks. In the three-tired architecture that we have proposed these three issues can be addressed separately at each tier. Importing a secure lower layer into the next higher layer enables security verification compositional. As a prerequisite to service layer confidentiality, resource models must be protected. As a simple first step solution the tool enforces access control rights for RPs and SPs. The intent is to ensure the integrity of resource information. SPs can use, but not modify RDTs, and resources allocated to the service bought by a SP are assured to be the resources included in the CCSs viewed by the clients. Thus, deception attacks can be detected, if not prevented, at source. Currently we are working on resource protection issues for other layers.

References

1. C. S. Group (2008) Cyber-physical systems: executive summary. Report, http://varma.ece.cmu.edu/summit/CPS-Executive-Summary.pdf
2. Georgakopolous D, Papazoglou MP (2008) Service-oriented vomputing. The MIT Press, Cambridge
3. Wan K, Alagar V (2013) A resource-centric architecture for service-oriented cyber physical system. In: Proceedings of the 8th international conference on grid and pervasive computing (GPC 2013), May 2013, Korea
4. Guttag JV, Horning JJ, Wing JM (1985) The Larch family of specification languages. IEEE Softw 2(5):24–36
5. Wan K, Alagar V (2013) Modeling resource-centric services in cyber physical systems. In: Proceedings of the international multi conference of engineers and computer scientists 2013, Lecture notes in engineering and computer science, 13–15 Mar 2013, Hong Kong, pp 716–721
6. Selic B (2000) A generic framework for modeling resources with uml. IEEE Comput 33(6):64–69
7. W3C. W3c recommendation. Technical report
8. Zhuge YH, Shi P (2008) Resource space model, owl and database: mapping and integration. ACM Trans. Internet Technol 8(4):20
9. Yen IL, Huang J, Bastani F, Jeng JJ (2009) Toward a smart cyber-physical space: a context-sensitive resource-explicit service model. In: Proceedings of 33rd annual IEEE international computer software and applications conference, IEEE Press

10. Hoang DD, Paik HY, Kim CK (2012) Service-oriented middleware architectures for cyber-physical systems. Int J Comput Sci Netw Secur 12(1):79–87
11. Wan K, Alagar V, Wei B (2013) Intelligent graphical user interface for managing resource knowledge in cyber physical systems, KSEM 2013, LNCS 8041. Dalian, China

Analyzing the Relationship Between B2B E-Marketplace Adoption and E-Business Performance Using NK Simulation Method

Woon Kian Chong, Yan Sun, Nan Zhang and Ka Lok Man

Abstract Business-To-Business electronic marketplace (B2B e-Marketplace), an electronic platform for buyers and sellers, provides a new dimension in facilitating the marketers to work more effectively when making critical marketing decisions. However, small to medium sized enterprises (SMEs) who are keen to compete in the electronic environment remain concerned as how their businesses can benefit from B2B e-Marketplace. We therefore developed a conceptual framework in which multiple facet of e-Marketing services derived from B2B e-Marketplace are linked to e-Business performance. We will use an adjusted NK simulation to test how Demographic Characteristics, Perceived Risk and Online Data Security influence the relationships between B2B e-Marketplace adoption and e-Business performance. The proposed framework will provide a guideline for academics and practitioners and highlights the significant role of each factor in developing and sustaining effective B2B e-Commerce practice for SMEs. Furthermore, SME managers also can derive a better understanding and measurement of marketing activities that appropriately balance between traditional and B2B e-Commerce practice. At the same time, the developed framework can be integrated into the companies to determine the level of e-Business performance in the B2B marketplace.

W. K. Chong (✉) · Y. Sun
International Business School Suzhou,
Xi'an Jiaotong-Liverpool University, Xi'an Jiaotong Suzhou, China
e-mail: woonkian.chong@xjtlu.edu.cn

Y. Sun
e-mail: yan.sun@xjtlu.edu.cn

N. Zhang · K. L. Man
Department of Computer Science and Software Engineering,
Xi'an Jiaotong-Liverpool University, Suzhou, China
e-mail: nan.zhang@xjtlu.edu.cn

K. L. Man
e-mail: ka.man@xjtlu.edu.cn

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 275, DOI: 10.1007/978-94-007-7684-5_8,

Keywords B2B e-Marketplace · Demographic characteristics · e-Business performance · Internet technologies · Online data security · Perceived risk

1 Introduction

B2B e-Commerce, as one of the major business models brought about by the Internet technologies, has made a significant contribution to the e-Marketers [1, 2]. Nonetheless, there are both limitations and gaps on how to explore the opportunities for SMEs in benefiting from the emergent e-Marketing services, derived from the B2B e-Marketplace. Despite much interest from academics and business publications in e-Business, a sharper focus on the B2B e-Marketplace is timely and warranted for several reasons. First, research on B2B e-Marketplace is limited and, given the enormous populations and size of these markets (for instance, Alibaba.com has 40 million registered users), offer considerable opportunities for online marketing activities. Second, we can observe that there is a growing awareness of the contribution of e-Marketing to the global business environment, but the issue of how B2B e-Marketplace can significantly affect firms' e-Business performance remains unclear. Third, B2B e-Marketplace offers significant opportunities to lower costs though global sourcing or local production. Indeed, this knowledge gap has contributed to the occasional misinterpretation of reasons for e-Marketing successes and failures in the region.

However, the initial proliferation of B2B e-Marketplaces proved to be unsustainable and the number of B2B e-Marketplaces in many industry sectors has been considerably few. The major barriers for B2B e-Marketplace adoption remain: lack of understanding of available technology, lack of confidence in electronic-based marketing, lack of technical knowledge and resources and the lack of e-Business recognition in some industry sectors. SMEs may fail to overcome these problems if the necessary capabilities are not evident in the B2B firms. This issue is frequently reported by many academics and practitioners [1, 2] and it limits the progress of SMEs in transforming their businesses into electronic-based entities. However, research into the effects of electronic business is beginning to examine SME involvement in B2B e-Marketplaces.

The reminder of the chapter is organized as follows. Section 2 shows the theoretical development of the study. We then describe the research methodology in Sect. 3 and the results are presented in Sect. 4. In Sect. 5, we present the findings based on the results from NK simulation. Finally, implications, concluding remarks are made and directions for future research are discussed in Sects. 6 and 7.

2 Theoretical Background

The development of B2B e-Marketplaces followed by advances in Internet technologies for marketing purposes has captured the attention of marketers in recent years.

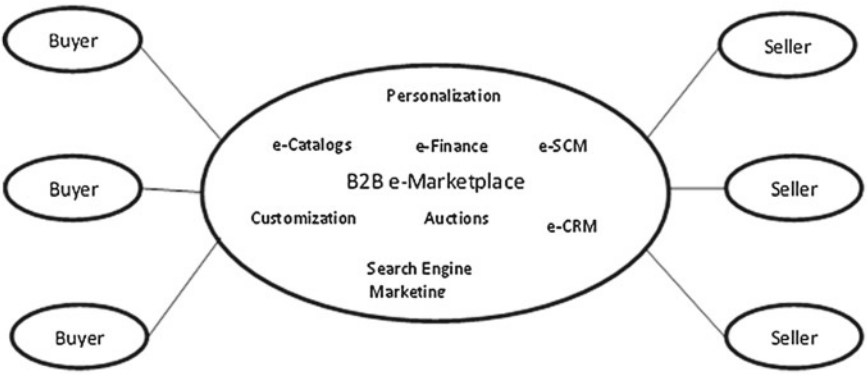


Fig. 1 The B2B e-Marketplace

The advent of B2B e-Marketplace provides new opportunities for companies to deal with their customers and suppliers. The evolution of B2B e-Marketplace technologies reduces the costs of closely integrating buyers and suppliers activities through the Internet technologies. Efficiency is achieved through many activities such as reducing procurement costs by making it easier to find lower price suppliers; and it is much less costly for buyers to place an order by adopting the technologies provided by B2B e-Marketplace (See Fig. 1). Furthermore, electronic intermediaries which provide a platform for B2B transactions will bring the large number of buyers and sellers in one standard gateway to perform their daily business activities.

B2B e-Marketplace has improved the effectiveness and efficiency of B2B processes between buyers and sellers, radically changing traditional procurement activities, and restructuring supply chains, organizations, and industries [3]. B2B e-Marketplace builds value propositions based on three marketing drivers: increased market share, improved channel coordination and new value creation [4]. As B2B e-Marketplace adopt the internet technologies, this will definitely increase the user's market share by helping them to identify their competitors, market position and segments in a dynamic digital environment. In terms of channel coordination, B2B e-Marketplace provides a transparency and visibility facilities across the supply chain such as inventory and production planning and control. New value creation occurs as B2B e-Marketplace promotes collaboration and allows increased information (e.g. market, competition) availability. The benefits of B2B e-Marketplace as reported by many academics and practitioners include:

- Reducing search costs [1, 5, 6].
- Improving production and supply chain ability [7, 8].
- Improving personalization and customization of product offerings [1].
- Enhancing customer relationships management [6].
- Reducing marketing costs [9].
- Operating 24/7 and round the clock on 365 days [10].
- Facilitating global presence [11].

- Exploring new market segments [12].
- Interacting more effectively in terms of services marketing communication [13].

Many authors [1, 2, 14, 15] highlighted several dimensions of the e-Marketing services, however, they do not examine the relationship between the B2B e-Marketplace adoption and e-Business performance. In addition, the relationship between B2B e-Marketplace adoption and e-Business performance can be influenced by Demographic characteristics, Perceived risk and Online data security. In the previous studies, many factors have been analyzed and tested to try to understand the slow rate of adoption. For instance, Howcroft et. al. [16] studied respondents' gender, age, annual income, level of education, ownership of financial products and compared the results with national average for the UK. Besides the demographic characteristics of customers, many studies have focused on customers' attitudes to and behavior towards the adoption of online services. Black et. al. [17] conducted a qualitative study that employed Roger's model [18] to analyze customers' adoption decisions and perceived risk was found to be significant, confirming the results of other studies which indicate that customers are concerned with online security, particularly of online banking [19, 20].

This study intends to develop an e-Business performance framework for SMEs who wish to adopt a proactive approach for business efficiency and competitive advantage, and those who wish to explore the Internet technologies for marketing activities. In addition, the relationship between e-Marketplace adoption and e-Business performance can be influenced by the degree of industrial dynamics or environmental turbulence. Thus, we further test the relationship in a turbulent industrial environment and a relatively stable industrial environment respectively. In satisfying the objective, the study will address a wide range of relevant issues including:

- A critical assessment of the B2B e-Marketplace adoption and their contribution to SMEs' e-Business performances from a marketing perspective.
- To test the moderating effects of Demographic characteristics, Perceived risk and Online data security on the relationship between B2B e-Marketplace adoption and e-Business performance in a turbulent industrial environment and a relatively stable industrial environment respectively.

3 Research Methodology

This study tries to fill in the gap and develop the literature by understanding the role of Demographic characteristics , Perceived risk and Online data security as institutional factors, moderate the relationship between B2B e-Marketplace adoption and e-Business performance. We therefore developed a conceptual framework of e-Business performance (Fig. 2) that focuses on the importance of B2B e-Marketplace adoption. The proposed hypotheses (H) are as below:

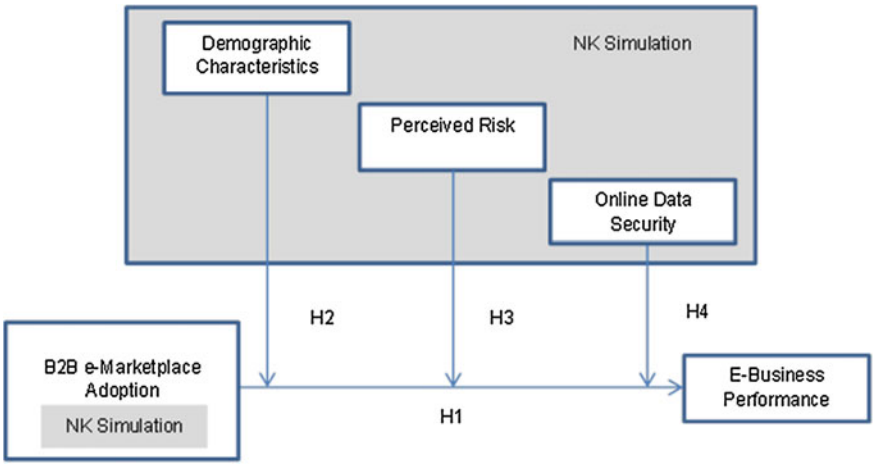


Fig. 2 E-Business performance conceptual framework

H1: The more a firm adopts the B2B e-Marketplace, the higher the firm's e-Business performance.
H2: Demographic characteristics will positively moderate the relationship between e-Marketing services adoption and e-Business performance.
H3: Perceived risk will positively moderates the relationship between e-Marketing services adoption and e-Business performance.
H4: Online data security will positively moderates the relationship between e-Marketing services adoption and e-Business performance.

We employed an adjusted NK simulation to explain the relationships and the moderation effects between the relationships. We examined what individual e-Marketing services are significant in e-Business performance improvement by using the adjusted NK model. It is the first time in the literature that an NK model was applied in interpreting e-Marketing issues. The reason why NK model is adopted is because it can provide a flexible context with different environmental as well as organizational configurations. Thus, it would be easy for scholars to experiment and study how various combinations of organizational external and internal factors influence a firm's performance. Additionally, in reality, there are a lot of difficulties in data collection across industries and it is almost impossible to put a firm in different contexts to examine its performance differences. Consequently, simulating a firm's behaviors in various virtual environments is a realistic alternative and can contribute to theory development [21].

4 Results

In the simulations, we set N = 16 (N= total number of agents in the NK model), while K is the degree of B2B e-Marketplace adoption with three options: 1, 4 and 8, representing low, medium and high. The higher is K, the more rugged the landscape will be. We generated 50 landscapes with 50 firms in each landscape. We let a firm's degree of Demographic Characteristics, Perceived Risk and Online Data Security (SR) range from 1 to 15. We used two different turbulence combinations (VR = 0.2 and VT = 1, VR = 0.05 and VT = 5) to represent a turbulent industrial environment and a relatively stable industrial environment respectively.

A. The effect of Demographic Characteristics, Perceived Risk and Online Data Security on the relationships between e-Marketing services adoption and e-Business performance

We set up two kinds of business environment (VR = 0.2 and VT = 1 versus VR = 0.05 and VT = 5), with three degrees of B2B e-Marketplace adoption (K = 1, 4, 8). In the fast-paced (Fig. 3) as well as the relatively stable business environments (Fig. 4), we allowed firms to face different degrees of knowledge complexity and examined the resulting relationships between the adoption and firm performance. All of the curves show that, as the degree of B2B e-Marketplace adoption increases,

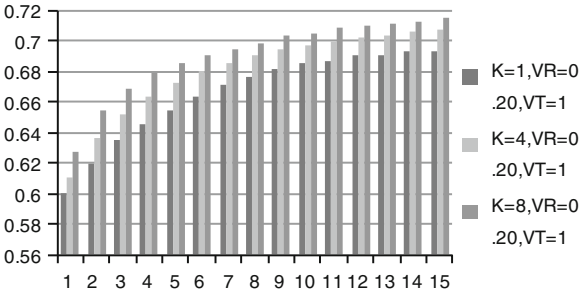


Fig. 3 Turbulent business environment

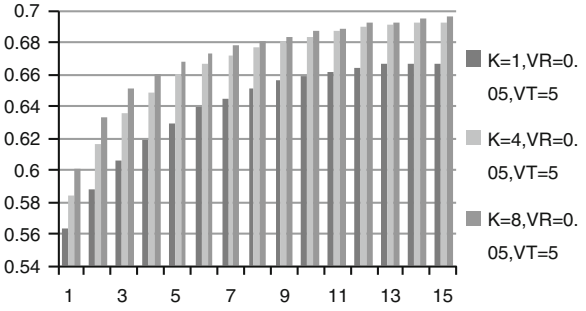


Fig. 4 Stable business environment

e-Business performance also increases, no matter how turbulent the environment or complex the knowledge. However, the slope of each curve decreases as the degree of B2B e-Marketplace adoption increases. In other words, a firm can use B2B e-Marketplace to improve its e-Business performance, but the marginal effects of doing so are higher when the firm increases its degree of Demographic Characteristics, Perceived Risk and Online Data Security from a low level (e.g. SR = 1 or 2) to a medium level (SR = 3, 4, 5) than when it moves from a medium to a high level (SR = 8, 9, 10 . . . 15). Hence, our hypotheses are supported.

5 Discussions

This study focuses on the adoption of B2B e-Marketplace and e-Business performance, particularly in SMEs. Supported by the findings above, positive relationships were found between the adoption and firm performance using three factors/simulators: demographic characteristics, perceived risk and online data security.

Demographic characteristics used to be keys to business success back to old times when internet hasn't stepped into our daily life. Especially for traditional business organizations focusing only on the market in the physical world, demographic factors almost reached every aspect of business process including accessibility for customers, opening/service hours, delivery speed, transportation cost etc. Surprisingly however, the first virtual bank, the Security First National Bank was launched in America in 1995 [22] and set up specifically to provide financial services through the internet. Other virtual banking operations, like Egg, First Direct in the UK soon followed.

Many major traditional banks such as HSBC, Barclays and RBS also offer online services in addition to their branch banking activities. An argument [23] was once raised in the UK due to service quality of overseas call centre. Considering the time difference and demographic limits, overseas call centre helps business organizations provide the seamless service for customers. Studied by Robinson et. al. [24], call centre management was successfully delivered a combination of customer service and financial budgets.

Being the downsides of doing business in cyber world, perceived risk (technical aspect) and online data security have started to draw people's attention gradually. According to the survey conducted by the UK Payment Association [25], the fraud losses involving the Internet and bank cards in the UK alone reached £609.9 million pounds in 2008. However, this does not stop business organizations from shifting their businesses to e-marketplace due to significant benefits, such as cost reduction. Suggested by Zhang [26], the cost of branch banking in USA was $1.07 while the cost of online banking only reached $0.1. The significant difference of transaction cost keeps driving business organizations moving to the e-marketplace/marketspace. On the other hand, the transaction cost/delivery cost would only take a tiny percentage

if business organizations actually provide service/knowledge instead of physical products/goods.

Specific questions were designed to understand the difference of business performance at "marketplace—before" and in "marketplace—after". The simulation results of this study showed positive attitudes to e-marketplace for B2B from various aspects and they believed that e-marketplace is helpful to improve business performance. In other words, the future of B2B is quite promising in e-marketplace.

6 Implications

Based upon an understanding of the analysis of the simulation, this study demonstrates the positive and significant correlations between B2B e-Marketplace adoption and e-Business performance. All the arguments are supported by the simulation analysis results. The findings identified the degree of demand needed to push a firm to adopt B2B e-Marketplace is moderated by demographic characteristics, perceived risk and online data security. Furthermore, this study has also demonstrated that an examination of B2B e-Marketplace is relevant to SMEs operating in digital environments. The exploitation of e-Marketing services derived from B2B e-Marketplace used by SMEs can cause profound changes to their business performance [27, 28]. If SMEs know that their business performance could by enhanced by adopting B2B e-Marketplace, they would feel the urge to work with B2B e-Marketplace service providers to improve stakeholders coordination. The findings are particularly useful for supply chain players by guiding them identify the value of B2B e-Marketplace to make right marketing decisions. Since B2B e-Marketplace is becoming increasingly critical for marketers, it is managerially crucial to establish a coordination mechanism between SMEs and B2B e-Marketplace.

The findings of this study are of great help for business organizations that are switching to the e-Marketplace, in particularly for SMEs, who are trying to increase e-business performance and competitive abilities. According to the data analysis and interpretation, managerial implications were learnt in the following three factors: demographic characteristics, perceived risk and online data security.

For business organizations worldwide today, embracing the e-marketplace seems to become the must-do instead of to-do strategy in terms of efficiency and cost of running business. Taking those three factors into consideration would benefit the business organizations at all stages, from the market entry to market domination stage.

This study focuses on B2B, which involves significant transaction volume on the internet comparing to B2C. Therefore business information at top confidential level is inevitably required in the online process. A solid understanding of perceived risk and appropriate solutions in term of online data security would put positive impacts on building customer relationship.

When data bleach occurred to TJ Maxx in 2007, [29] a panic was stirred among leading retailers worldwide and estimated 94 million bank card accounts were under

the security risk. Suggested by Berg et. al. [29], the monetary loss was expected to be $4.5 billion and even worse, the damage made to the business reputation of TJ Maxx would never be assessed accurately.

Among the three factors that have been tested significantly positive in the simulation, perceived risk and online data security have become the key indicators in terms of adoption of B2B e-Marketplace and e-business performance improvement because demographic limits were conquered by advanced information technology nowadays.

Besides the practical implications for business organizations, this study would bring benefits to local authorities directly with respect to policy making and market regulating. Particularly for e-Marketplace which provides a platform for cross-border transactions, collaboration among different countries and areas is another layer of assurance to build safe trading environment on the internet and maintain professional business performance.

7 Conclusion and Future Research

This study provides a distinct stream of literature incorporating a simulation analysis in order to create a rich and deep understanding of the fields of B2B Marketing, e-Marketing, and B2B e-Marketplace. In addition to the specific contributions outlined above, this study provides an example of how NK simulation model methods can be combined in a consistent and complementary fashion to provide a holistic understanding of e-Business performance.

The B2B e-Marketplace has a profound impact on the global business environment. However, it raises many unsolved questions for marketers especially in from the e-Marketing perspective. This research is related to one of the most important topics of e-Marketing, which contributes to the B2B e-Marketplace literature by further demonstrating the B2B e-Marketing activities conducted by SMEs [30].

The ideas presented in this study offer a complementary perspective to many existing theories advocated by practitioners. Current studies indicate that SMEs are still investigating whether or not they should implement e-Business. Based on the need for a dynamic framework for e-Business from the literature, this study is significant to SMEs, marketers, IT practitioners and other stakeholders that use the Internet and other electronic means for B2B marketing purposes.

The limitation of the study may be argued to be the validity of the simulation results of the study. Similarly, one could also reason that due to the homogeneous characteristics of SMEs, the study does provide a credible and useful source of reference. Future studies may place further emphasis in empirical perspective with a focus on large scale organizations for result comparisons. Nonetheless, this study contributed towards the systematic adoption of B2B e-Commerce on SMEs' business performances, with implementing the model being another area for future research.

References

1. Bakos YJ (1998) Towards friction-free markets: the emerging role of electronic marketplaces on the Internet. Commun ACM 41(8):35–42
2. Chaffey D, Ellis-Chadwick F, Mayer R, Johnston K (2009) Internet marketing: strategy, implementation and practice, 4th edn. Prentice-Hall, Harlow
3. Balocco R, Perego A, Perotti S (2010) A classification framework to analyse business models and critical success factors. Ind Manag Data Syst 110(8):1117–1137
4. Rohm AJ, Kashyap V, Brashear TG, Milne GR (2004) The use of online marketplaces for competitive advantage: a Latin American perspective. J Bus Ind Mark 19(6):372–385
5. Kandampully J (2003) B2B relationships and networks in the Internet age. Manag Decis 41(5):443–451
6. Kaplan S, Sawhney M (2000) E-hubs: the new B2B marketplaces. Harvard Bus Rev 78(3):97–103
7. Barua A, Ravindran S, Whinston AB (1997) Efficient selection of suppliers over the internet. J Manag Inf Sys 13(4):117
8. Albrecht C, Dean D, Hansen J (2005) Marketplace and technology standards for B2B e-commerce: progress, challenges, and the state of the art. Inf Manag 42(6):865–75
9. Sculley AB, Woods WA (2001) B2B exchanges: the killer application in the business-to-business internet revolution. Harper and Collins, NY
10. Ngai EWT (2003) Internet marketing research (1987–2000): a literature review and classification. Eur J Mark 37(1/2):24–49
11. Laudon KC, Laudon JP (2002) Management information systems: managing the digital firm. Prentice-Hall Inc, Upper Saddle River, NJ
12. Murtaza MB, Gupta V, Carroll RC (2004) E-marketplaces and the future of supply chain management: opportunities and challenges. Bus Process Manag J 10(3):325–335
13. Petersen KJ, Ogden JA, Carter PL (2007) B2B e-marketplaces: a typology by functionality. Int J Phys Distrib Logistics Manag 37(2):4–18
14. Cannon JP, Perreault WD (1999) Buyer-seller relationships in business markets. J Mark Res 36(4):439–460
15. Kalyanam K, McIntyre S (2002) The e-marketing mix: a contribution of the e-tailing wars. J Acad Mark Sci 20(4):487–499
16. Howcroft B, Hamilton R, Hewer P (2002) Consumer attitude and the usage and adoption of home-based banking in the United Kingdom. Int J Bank Mark 20(3):111–121
17. Black NJ, Lockett A, Winklhofer H, Ennew C (2001) The adoption of internet financial services: A qualitative study. Int J Retail Distrib Manag 29(8):390–398
18. Rogers EM (1962) The diffusion of innovations. Free Press , NY
19. Jayawardhena C, Foley P (2000) Changes in the banking sector–the case of internet banking in the UK. Internet Res Electron Netw Appl Policy 10(01):19–30
20. Rotchanakitumnuai S, Speece M (2003) Barriers to internet banking adoption: a qualitative study among corporate customers in Thailand. Int J Bank Mark 21(6/7):312–323
21. Davis JP, Eisenhardt KM, Bingham CB (2007) Developing theory through simulation methods. Acad Manag J 32(2):480–499
22. Zeng R (2006) Risk analysis of the Internet banking in China. China Acad J 7.
23. HSBC moves jobs to overseas call centres, 'saving and banking' (7 Feb 2013), http://www.thisismoney.co.uk/money/saving/article-1670745/HSBC-moves-jobs-to-overseas-call-centres.html
24. Robinson G, Morley C (2006) Call centre management: responsibilities and performance. Int J Serv Ind Manag 17(3):284–300
25. APACS (2009) Fraud, the facts 2009. APACS, The UK Cards Association
26. Zhang Q (2007) Introduction to online banking services in China. China Financial Publishing House, Beijing
27. Chong WK, Shafaghi M, Tan BL (2011) Development of a business-to-business critical success factors (B2B CSFs) framework for Chinese SMEs. Mark Intell Plann 29(5):517–533

28. Chong WK, Shafaghi M, Woollaston C, Lui V (2010) B2B e-marketplace: an e-marketing framework for B2B commerce. Mark Intell Plann 28(3):310–329
29. Berg GG, Freeman MS, Schneider KN (2008) Analysing the TJ Maxx data security fiasco: lessons for auditors. CPA J 78(8):34–37
30. Chong WK, Sun Y, Zhang N, Man KL (2013) The role of demographic characteristics, perceived risk and online data security on E-business performance. In: Lecture notes in engineering and computer science: proceedings of the international multiConference of engineers and computer scientists, Hong Kong, pp 722–725, 13–15 Mar 2013

Computing Implied Volatilities for Exchange-Traded Options on Dividend-Paying Stocks

Nan Zhang and Ka Lok Man

Abstract We present an algorithm and its software implementation that computes implied volatilities for exchange-traded stock options. The stocks underlying the options are assumed to pay out dollar cash dividends. The Leisen–Reimer (LR) binomial tree is used for the option pricing that is at the core of the root-finding procedure in searching for the implied volatilities. The Brent's method is used for the root-finding, and the option pricing code is optimised for performance. Tests were made on call and put options traded on the stocks of Microsoft Corporation and Apple Inc. In 0.046 and 0.226 s, respectively, the implemented software program completed the computation for 154 Microsoft and 823 Apple call options.

Keywords Brent's method · Dividend-paying stocks · Implied volatility · LR binomial tree · Option pricing · Volatility surface

1 Introduction

In the Black-Scholes-Merton (BSM) [1, 7] option pricing model volatility is the standard deviation of the continuous compounding return of the underlying stock in one year's time. It is a measure of the uncertainty about the returns provided by the stock. In the settings of the original BSM model volatility of a stock is assumed to be constant during the lifetime of an option traded on the stock.

N. Zhang (✉) · K. L. Man
Department of Computer Science and Software Engineering, Xi'an Jiaotong-Liverpool University, Suzhou, China
e-mail: nan.zhang@xjtlu.edu.cn

K. L. Man
Baltic Institute of Advanced Technologies, Vilnius, Lithuania
e-mail: ka.man@xjtlu.edu.cn

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 275, DOI: 10.1007/978-94-007-7684-5_9,

Given a stock and an option on the stock the volatility implied by the option price and the BSM pricing model can be computed using an iterative root-finding procedure. Contrary to the assumption made by the BSM model that the volatility is constant, the computed implied volatilities often vary with strike prices and expirations. In practice, different volatilities are used to price options having different strikes and expirations. The general pattern of the variation with strike in implied volatility for stock options is referred to as a volatility skew. The volatility used to price a low-strike option is higher than that used to price a high-strike option. The changing of implied volatilities with time to expiration is referred to as a volatility term structure. Combining volatility skews for different expirations generates a surface of implied volatilities that tabulates the volatilities appropriate as determined by the market for pricing options with certain strike and expiration.

Because option price is monotonically increasing in volatility, implied volatility computed from option price is often used as a proxy for option value. To compare the relative value of two options an investor needs only to look at their implied volatilities. As a proxy for option value, implied volatility presents the market's expectation of a stock's future price moves. A significant change in implied volatility means that there may be a shift in the expectation of the market towards a stock's future price. This provides to investors a tool for predicting the direction of a stock's future price moves.

Prices of exchange-traded stock options change in real time. The computation of the implied volatilities for exchange-traded options therefore should be performed with minimum delay. To this purpose we have developed a software that computes the implied volatilities from prices of exchange-traded options for different strikes and expirations. The underlying algorithm is able to handle multiple dollar cash dividends that take place within the lifetime of an option. To accelerate the computation the Leisen–Reimer (LR) [6] binomial tree is used for the underlying option pricing, because it has a much faster convergence speed than the commonly-used Cox–Ross–Rubinstein (CRR) [3] tree. Brent's method [2] is used as the root-finding algorithm which takes the pricing procedure at its core and searches for solutions on an iterative basis. Tests were made on an Intel quad-core 3.4 GHz Corei7-2600. Using a single thread the software finished in 0.22 s generating implied volatilities for 823 stock options. A preliminary conference presentation for this work is found in [8].

Organisation of the rest of this book chapter

Section 2 gives a brief background on exchange-traded stock options. Section 3 discusses the volatility computation algorithm in detail. This includes explanations on handling dollar cash dividends, the LR tree pricing method, the root-finding procedure and the optimisations to the option pricing procedure. Section 4 discusses how common range of strike prices are selected for processing the options in the tests. Section 5 presents the test results. Section 6 draws the conclusion.

Table 1 Option chains on Microsoft's stock on July 13, 2012. All these options expire on August 17, 2012

Calls					Puts			
Price	Change	Bid	Ask	Strike	Price	Change	Bid	Bid
–	–	9.40	9.65	19.00	0.04	0.00	0.01	0.03
–	–	8.50	8.70	20.00	0.04	0.00	–	0.02
8.15	0.00	7.60	7.70	21.00	0.04	0.00	0.01	0.04
8.00	0.00	6.60	6.70	22.00	0.02	0.00	0.02	0.03
5.85	−0.40	5.60	5.70	23.00	0.03	0.00	0.03	0.04
4.94	−0.96	4.60	4.70	24.00	0.07	+0.03	0.06	0.07

2 Exchange-Traded Stock Options

Most stock options are traded on exchanges. Such exchanges in the United States include the Chicago Board Options Exchange (www.cboe.com) and NASDAQ OMX (www.nasdaqtrader.com). Exchange-traded options are American in style, which means they can be exercised at any time before or on the expiration date. For American options there is no closed-form pricing formula exist. Their prices must be computed using numerical procedures. In this work we use the binomial tree method.

Within the lifetime of many exchange-traded options cash dividends will be paid out by the underlying stocks. Because exchange-traded options are not adjusted for cash dividends we need to take the effect of cash dividends on option prices into consideration in the computation for implied volatilities. Information about exchange-traded options are easily accessible from the Internet. Exchange-traded options are often organised into chains. A chain consists of options with different strike prices but the same expiration date. Table 1 shows a segment of an option chain on vanilla call and put options traded on the stock of Microsoft Corporation.

3 Computing Implied Volatilities

Computing the implied volatility of an option is to find the right value for the volatility parameter which if fed into an option pricing model will produce the option's traded price. The computation starts with an initial estimation for the volatility. This initial estimation is fed into the option pricing model to compute the option price under the estimation. This price is then compared with the option's traded price to calculate a closer estimation according to the difference. This procedure repeats until the option's price computed under an estimation for volatility becomes close enough to the option's traded price. At this point the estimation is treated as the implied volatility of the option.

3.1 Binomial Pricing with Cash Dividends

Since exchange-traded stock options are American-style their prices must be computed by numerical procedures. For this purpose we use the binomial method. Most of exchange-traded options see within their lifetimes cash dividends paid out by their underlying stocks. For this reason we follow the method presented by John Hull [5] which incorporates dollar dividends into the binomial pricing. The advantage of this method is that it preserves the recombining structure of binomial trees.

The method calculates the present value of all cash dividends known to be paid out by the underlying stock within the lifetime of the option. An recombining binomial tree is then built where the root node corresponds to the stock's present price less the present value of all future cash dividends. After the whole tree is built, discounted cash dividends are added upon nodes of the tree at appropriate time steps. In what follows, we discuss this method in detail assuming a single dividend within the lifetime of the option. But this method can be easily generalised to handle multiple cash dividends.

Suppose D is the amount of cash dividend that is to be paid out at the ex-dividend time τ (measured in years). We use r to denote the annual continuously compound interest rate. The present value of the cash dividend is $De^{-r\tau}$. If S_0 denotes the present stock price, the uncertain component S_0^* of the price is $S_0^* = S_0 - De^{-r\tau}$. We then build a recombining binomial tree that models the dynamics of S^* – the uncertain component of the stock price process S. This tree is rooted at S_0^*. We use u and d to denote the up-move and down-move factors of the tree, respectively. On such a binomial tree a jth node at ith time step corresponds to the stock price $S_0^* u^j d^{i-j}$, where $j \in \{0, 1, 2, \ldots, i\}$. Note that at ith time step the number of nodes is $i + 1$. Both indexes i and j start from zero. Now based on this tree that models S^* we convert it to another tree that models S. Since there is a single cash dividend D paid out at time τ we add D's discounted value onto all nodes whose time horizon proceeds τ. At time step i the nodes in the modified tree correspond to the stock prices $S_0^* u^j d^{i-j} + De^{-r(\tau - i\Delta t)}$, $j = 0, 1, 2, \ldots, i$ when $i\Delta t < \tau$, and $S_0^* u^j d^{i-j}$, $j = 0, 1, 2, \ldots, i$ when $i\Delta t > \tau$. The quantity Δt is the length of time represented by one step of the tree. If there are multiple cash dividends nodes at time step i need to be adjusted by the sum of the discounted values of all future dividends. Figure 1 shows an example where there are two cash dividends within the lifetime of an option. The dashed lines show the tree that models S^*, and the solid lines show the modified tree that models S. The tree that models S^* is generated by the LR binomial tree method.

3.2 LR Tree Versus CRR Tree

In the computation for the implied volatilities the root-finding procedure may need to run many iterations. In each iteration the option price is evaluated under different estimations of the volatility. It is therefore very important to use an efficient option

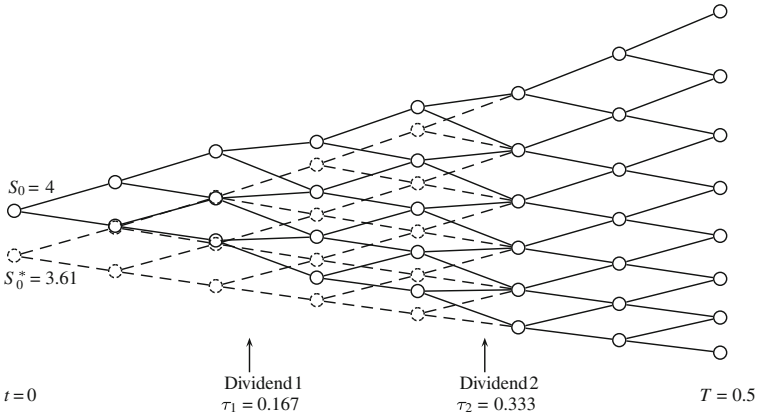


Fig. 1 Binomial trees for an American call with two dividends. The parameters were set as: current stock price $S_0 = 4$, strike price $K = 4$, interest rate $r = 0.1$, volatility $\sigma = 0.2$, expiration $T = 0.5$, time steps $N = 7$, ex-dividend times $\tau_1 = T/3$, $\tau_2 = 2T/3$ and cash dividends $D_1 = D_2 = 0.2$

pricing procedure as it will be called many times during the computation. In our work we use the LR binomial tree [6] rather than the more common CRR tree [3]. The LR tree converges much faster than the CRR tree. Figure 2 shows a comparison between the LR tree and the CRR tree using a deep in-the-money European put option as an example. It can be seen that the price computed by the LR tree converges smoothly to the Black-Scholes (BS) price of the option without the oscillation pattern demonstrated by the CRR tree. Note that the LR-tree method works only on odd numbers of time steps.

As before we denote the annual continuously compound interest rate by r, the option's expiration time by T, the number of time steps by N, the strike price by K

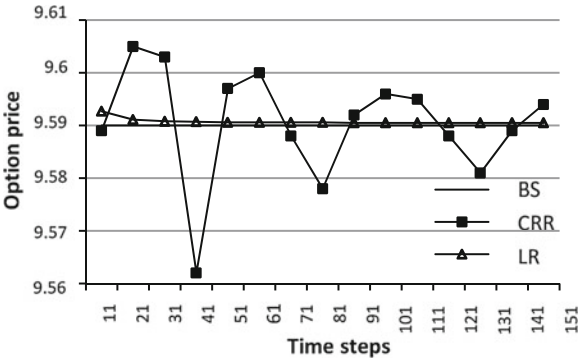


Fig. 2 Convergence comparison between the LR tree and the CRR tree using an European put option. The parameters were $S_0 = 40$, $K = 50$, $r = 0.1$, $\sigma = 0.4$ and $T = 0.5$

and the uncertain component of the initial stock price by S_0^*. The up-move probability p, up-move factor u and down-move factor d used with the LR-tree method are set through the following formulas.

$$d_1 = \frac{\ln(S_0^*/K) + (r + \sigma^2/2)T}{\sigma\sqrt{T}}$$

$$d_2 = \frac{\ln(S_0^*/K) + (r - \sigma^2/2)T}{\sigma\sqrt{T}}$$

$$h^{-1}(z) = \frac{1}{2} + \frac{\text{sgn}(z)}{2}\sqrt{1 - \exp\left[-\left(\frac{z}{N + \frac{1}{3}}\right)^2\left(N + \frac{1}{6}\right)\right]}$$

$$\text{sgn}(z) = \begin{cases} 1 & z > 0 \\ 0 & z = 0 \\ -1 & z < 0 \end{cases}$$

$$p\prime = h^{-1}(d_1)$$

$$p = h^{-1}(d_2)$$

$$u = \exp(rT/N)\frac{p\prime}{p}$$

$$d = \frac{\exp(rT/N) - pu}{(1 - p)}$$

3.3 The Root-Finding Procedure

For the root-finding purpose we use Brent's method [2] which builds on an earlier algorithm proposed by Dekker [4].

3.4 Optimisations

To minimise the runtime required by the binomial pricing procedure we applied several source level optimisations to the code. A binomial tree conceptually is a two-dimensional structure. In main memory we store it in an one-dimensional array because explicit construction of a whole binomial tree is unnecessary. During the backward computation that starts from the leaf nodes the one-dimensional array stores only the option values represented by the nodes that are being processed. When the computation proceeds from the ith time step to the $(i - 1)$th step the values represented by the nodes at the ith step are overwritten. This saves the time and space that are required by constructing a whole binomial tree.

On a that models S^*, the uncertain component of the stock price, at the ith time step the nodes represent the stock prices $S_0^* u^j d^{i-j}$, $j = 0, 1, 2, \ldots, i$. But we did not use these expressions to compute the stock prices because using these expressions will cause lots of repetitive work. Instead, we only use this expression once when computing the stock price S_i^{*0} represented by the 0th node at the ith time step, and we have $S_i^{*0} = S_0^* d^i$ because $j = 0$. When computing the stock price S_i^{*1} we set it to be the product of S_i^{*0} and the constant u/d, and so $S_i^{*1} = S_i^{*0}(u/d) = S_i^* u d^{i-1}$, $j = 1$. When computing S_i^{*2} we set it to be $S_i^{*2} = S_i^{*1}(u/d) = S_i^* u^2 d^{i-2}$, $j = 2$. Each new stock price is obtained by multiplying the constant u/d onto the price that has just been computed. This avoids the repetitive evaluation of the power expressions u^j and d^{i-j}, and therefore significantly shortens the runtime of the binomial pricing procedure.

The same binomial pricing procedure in our program works both for European and American options and for call and put options. But we did not use branching statements to distinguish these situations. This is to avoid the penalty brought about by the potential mis-predication on the branching statements. To distinguish European and American options a lookup table of two positions was used. The first position of the table always stores zero, and the second stores the payoff from an immediate exercise. In the case of an European option an expected value is compared with the first element of the table, and in the case of an American option it is compared with the second. These two situations are unified by using a lookup index whose value is 0 for European options and 1 for American. To unify call and put options we calculate their payoffs by the formula $\max(I(S - K), 0)$, where I is 1 for calls and -1 for puts, and S and K are the stock and strike prices, respectively.

4 Implementation

The programs in our work were written in C/C++. The compiled generator takes as input a text file that lists information about a stock, the option chains based on the stock and interest rate term structure. The output of the generator is a text file contains the computed implied volatilities arranged in a tabular form with columns being the expiration dates and rows the strike prices. The text script input file containing information for Microsoft call option chains is displayed in Fig. 3. In the text file the option chains are numbered starting from zero. Only a fraction of the first option chain is displayed in the figure.

Options offered by the exchanges are organised into chains for different expiration dates. For a particular expiration date the call or put option chains contains a series of options. With all other parameters being the same the options in a chain differ in their strike prices. In most cases, the increment in strike prices between each pair of successive options is a fixed amount, which depends on the present price of the stock. To build an implied volatility surface for a stock that includes all the options in all the chains we need to choose common strike prices for all the options. However,

```
Company          MSFT
Date             23-05-2012
Stock-price      29.11
Option-type      call
Option-style     American

Number-of-dividends 2
Dividend-values      0.2  0.2
Dividend-times       0.231  0.481

Number-of-expirations   6
Expiration-dates     23  58  86  121  149  240
Interest-rates       0.002397  0.003458  0.004588  0.005754  0.006525  0.008559

Strikes  13   47  1  (minimum  maximum  increment)

Option-prices " (rows for strike, column for expiration)"

Option-chain         0

        price-strike        13.15        16
        price-strike        12.95        17
        price-strike        11.95        18
        price-strike        10.95        19
        price-strike        9.15         20
        price-strike        8.15         21
        price-strike        7.15         22
        ...
```

Fig. 3 Excerpt of the input file containing information about the option chain traded on May 23, 2012 for Microsoft stocks

for exchange-traded options the range of strike prices are often different in different chains that expire on different dates. Moreover, the strike price increment in different chains can be different as well. All these irregularities pose problems for the volatility computation.

To solve these problems we choose the minimum and maximum strike prices that are found in all the option chains on a stock as the lower and upper bounds of the range of strike prices. The increment between each pair of strikes we choose is a value that will include most of the options in all the chains in the computation. In choosing the range of the strikes and the increment call and put option chains are dealt separately. Once a range of strikes and the increment have been chosen the generator will go through this range starting from the smallest strike price. For each strike price in the range and an option chain the generator will find the option with that strike price in the chain and compute the volatility implied by its exchange-traded price. The generator repeats this process until all the option chains listed in the input text file have been processed. If a strike price in the range is not found in an option chain the generator will by-pass this strike price and proceed to the next. In the output text file the generator will write a special notation (####) in the entry that corresponds to the un-found strike price. If in any option chain, an option's strike price is not found in the range the option will be ignored by the generator and no information will be output for that option. This can happen if in some option chain the strike

```
S0          29.11
callORput       1
Early exercise  1
Interest rates  0.0024  0.0035  0.0046  0.0058  0.0065  0.0086
Dividends       0.2000  0.2000
Dividend times  0.2310  0.4810
Number of time steps        120

Option prices (strikes for rows, expirations for columns)

        23      58     86     121     149     240
13.00  ####   17.90   ####    ####    ####    ####
14.00  ####   16.80   ####    ####    16.50   ####
15.00  ####   15.70   ####    ####    14.55   14.20
16.00  13.15  13.80   ####    ####    13.50   13.60
17.00  12.95  12.35   ####    ####    12.30   ####
18.00  11.95  11.25   ####    ####    11.45   ####
19.00  10.95  10.50   ####    ####    10.40   10.45
20.00  9.15    9.20   9.20    ####    10.00   9.35
21.00  8.15    8.20   8.25    ####    8.50    8.55
22.00  7.15    7.20   7.30    7.35    7.40    ####
23.00  6.15    6.25   6.35    6.40    6.50    ####
24.00  5.20    5.30   5.40    5.50    5.60    5.85
...
```

Fig. 4 Excerpt of the output file containing the implied volatilities for the the option chains shown in Fig. 3

prices follow an irregular increment pattern. An excerpt from an output file is shown in Fig. 4.

5 Tests

We made tests using options traded on the stock of Microsoft Corporation (stock code MSFT) and Apple Inc. (stock code AAPL). The information about the options were collected on May 23, 2012, at which time Microsoft's stock was traded at $29.11 per share and Apple's stock at $570.56 per share. For each of the two stocks we selected 6 call option chains and 6 put option chains. The options in these chains expired in 23, 58, 86, 121, 149 and 240 days, respectively. These were converted to years when implied volatilities were computed, assuming 365 days a year. The interest rates corresponded to each of the expiration dates were 0.2397, 0.3458, 0.4588, 0.5754, 0.6525 and 0.8559 %. The range of strike price for the Microsoft options were minimum $13, maximum $47 and increment $1, and the range for the Apple options were $135, $960 and $5, respectively. At the time the data were collected no dividend information was available for Apple's stock, but Microsoft would pay out $0.2 at two future dates within the lifetime of the options that had the longest expiration. The distances of the two days to May 23, 2012 were 0.231 and 0.481 years, respectively.

The tests were made on an Intel quad-core 3.4 GHz Corei7-2600 processor under Ubuntu Linux 10.10 64-bit version. The programs were compiled by Intel C/C++ compiler icpc 12.0 for Linux with -O3 and -ipo optimisations. The compiled generator finished in 0.046 s in computing the implied volatilities for the 154 Microsoft call options, and in 0.045 s for the same number of Microsoft put options. The runtimes in processing the 823 Apple call and 823 Apple put options were 0.226 and 0.222 s, respectively. The timing results were obtained by running the generator using a single thread. In all the computations the number of time steps was fixed to 121 in the LR binomial pricing.

From the computed implied volatilities we generated plots. Figure 5 shows the plots for the volatility skews (Fig. 5a–d) and the volatility term structure (Fig. 5e–h). Figure 6 shows the surfaces which combine the volatility skew and term structure.

Volatility skew describes the relationship between implied volatility and strike price. The typical pattern for volatility skew, as it is shown by the curves in Fig. 5c, is that implied volatility decreases as strike price increases. This means that the volatilities used to price options of low strike prices (i.e., deep in-the-money calls and deep out-of-the-money puts) are higher than that used to price options of high strike prices (i.e., deep out-of-the-money calls and deep in-the-money puts). Three curves are plotted in each of the Fig. 5a–d. They are for options expiring in 86, 121 and 149 days, respectively.

The volatility term structure describes the relationship between implied volatility and time to expiration. Implied volatility tends to be increasing as time to expiration increases when short-dated volatilities are historically low. This is the situation demonstrated by the curves in Fig. 5g–h. Similarly, implied volatility tends to be decreasing as time to expiration increases when short-dated volatilities are historically high. This seems to be the situation described by the solid-line curve in Fig. 5e. The curves plotted in each of the Fig. 5e–h are for options with different strike prices.

6 Conclusion

We have presented an implied volatility generator that computes implied volatilities for exchange-traded call and put options. We use the binomial method as the underlying option pricing model and implemented it using the LR trees. The pricing process handles multiple dollar cash dividends by separating a stock's price into a certain component and an uncertain component. The certain component consists of the discounted value of all cash dividends. The uncertain component is the stock's price less this certain component. To price an option of the stock a LR tree is first built using the present value of the uncertain component as root and then the discounted value of the dividends are added upon nodes of the tree at appropriate time steps. The Brent's algorithm was used as the root-finding method.

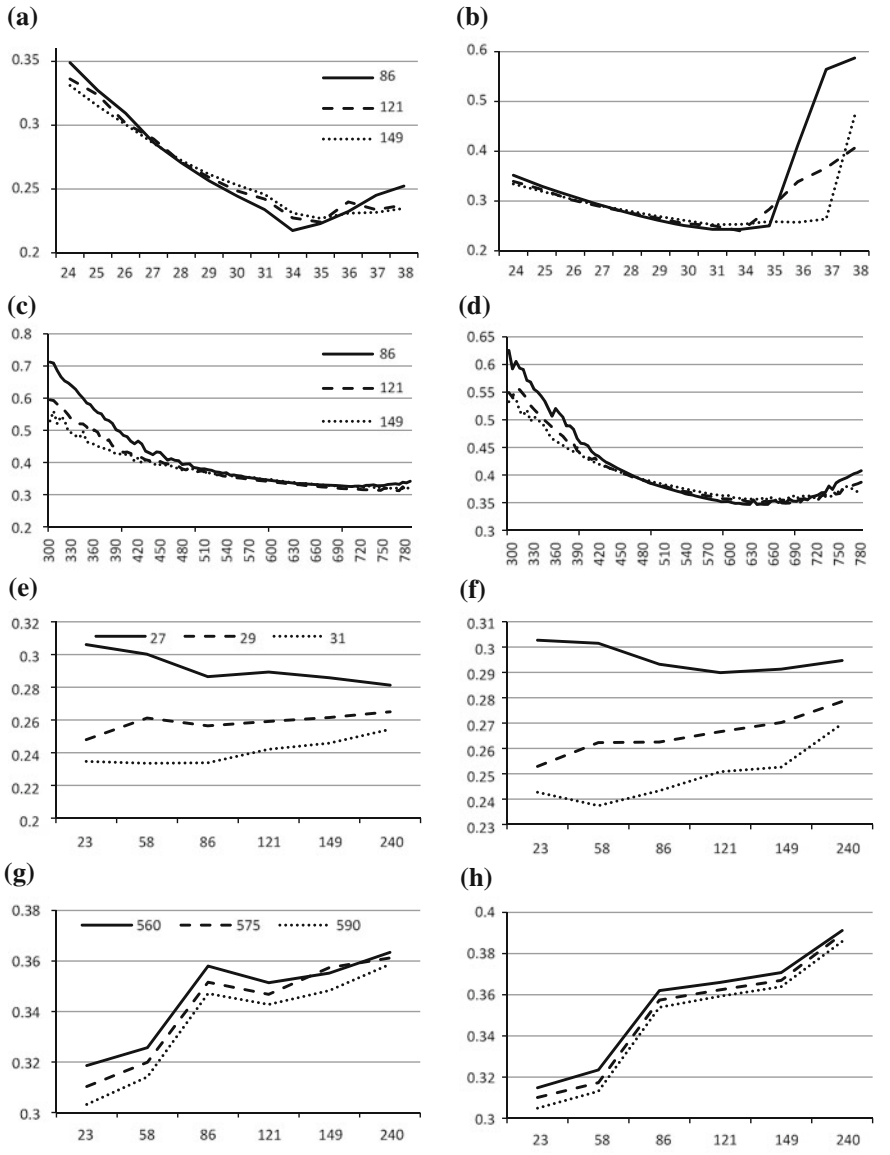


Fig. 5 Volatility skew and term structure for MSFT and AAPL call and put options. The x-axes in (**a**–**d**) are labelled by strike price, and in (**e**–**h**) are labelled by expiration date. **a** MSFT call volatility skew for different expirations. **b** MSFT put volatility skew for different expirations. **c** AAPL call volatility skew for different expirations. **d** AAPL put volatility skew for different expirations. **e** MSFT call volatility term structure for different strikes. **f** MSFT put volatility term structure for different strikes. **g** AAPL call volatility term structure for different strikes. **h** AAPL put volatility term structure for different strikes

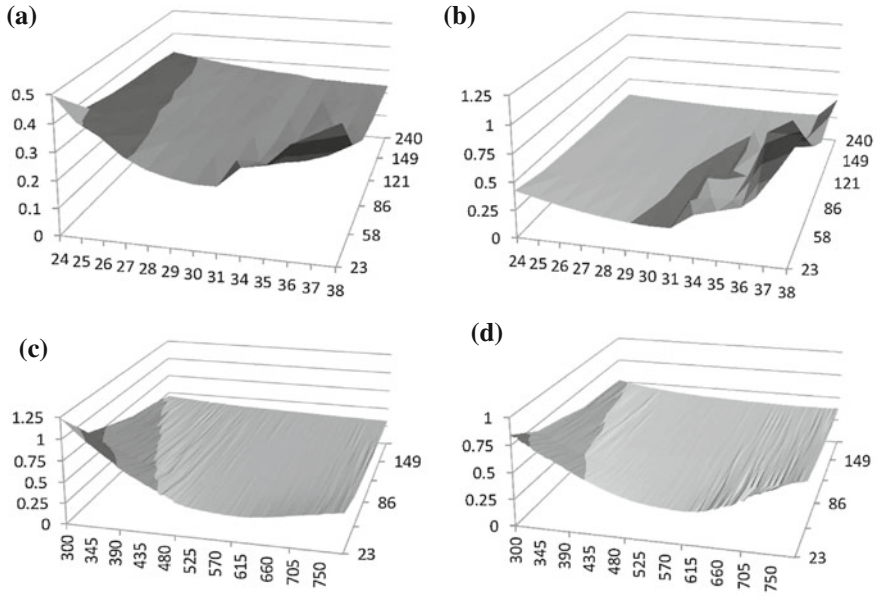


Fig. 6 Implied volatility surface forMSFT and AAPL call and put options. The x-axes are labelled by strike, y-axes by expiration and z-axes by implied volatility

Source code level optimisation techniques were applied to the option pricing procedure to minimise its runtime. Tests were made on options traded on stocks of Microsoft Corporation and Apple Inc. From the computed results implied volatility skew, term structure and surface were plotted and presented. Some of the plots confirm to the well-known pattern for volatility skew or term structure.

Acknowledgments This work was supported by the XJTLU Research Development Fund under Grant 10-03-08.

References

1. Black F, Scholes M (1973) The pricing of options and corporate liabilities. J Polit Econ 81(3):637–659
2. Brent RP (1973) Algorithms for minimization without derivatives, Chap. 4. Prentice-Hall, NJ.
3. Cox JC, Ross SA, Rubinstein M (1979) Option pricing: a simplified approach. J Financ Econ 7(3):229–263
4. Dekker TJ (1969) Finding a zero by means of successive linear interpolation. In: Dejon B, Henrici P (eds) Constructive aspects of the fundamental theorem of algebra. Wiley-Interscience, London
5. Hull JC (2012) Options, futures, and other derivatives, Chap. 20.3, 8th edn. Prentice Hall, NJ.
6. Leisen D, Reimer M (1996) Binomial models for option valuation-examining and improving convergence. Appl Math Finan 3:319–346

7. Merton R (1973) Theory of rational option pricing. Bell J Econ Manag Sci 4:141–183
8. Zhang N, Man KL (2013) Fast generation of implied volatility surface for exchange-traded stock options. In: Lecture notes in engineering and computer science: proceedings of the international multiConference of engineers and computer scientists. Hong Kong, vol 2, pp 741–746

Investigating Polling Cycle Time with Waited-Based DBA in GPONs

I-Shyan Hwang, Jhong-Yue Lee and Tzu-Jui Yeh

Abstract The Gigabit Passive Optical Network (GPON), have emerged as one of the most promising access network technologies for future last-mile solutions. To prevent data collision and ensure efficient transmission, the point-to-multipoint topology of GPONs requires a time-division media access control (MAC) protocol to allocate the shared resource of a common upstream transmission medium. Therefore, Dynamic Bandwidth Allocation (DBA) is an open and hot topic in the GPON. However, most proposed DBA plans to ignore the impact of the maximum cycle time for Quality of Service (QoS) ensures maximum delay uplink bandwidth utilization, and drop probability in GPON. In this paper, we propose a Waited-based Dynamic Bandwidth Allocation called WDBA which is predict the arriving real time packet based on the proportion of waiting time for multiple services over GPONs. In addition to ensuring the quality of QoS, our work focus on the fundamental problem of trading-off between upstream channel utilization and maximum polling cycle time with different proportion of Traffic Containers (T-CONTs) traffic in a GPON with multiple ONUs and verify the accuracy of the analysis with simulations. Overall, our numerical results indicate that the packet delay, throughput and drop probability performance is better when the polling cycle time is longer; the fairness is better when the polling cycle time is shorter.

Keywords DBA · GPON · MAC · Polling cycle time · QoS · WDBA

I-S. Hwang (✉)
Department of Information Communication, Department of Computer Science and Engineering, Yuan-Ze University, Chung-Li 32003, Taiwan
e-mail: ishwang@saturn.yzu.edu.tw

J-Y. Lee and T-J. Yeh
Department of Computer Science and Engineering, Yuan-Ze University, Chung-Li 32003, Taiwan
e-mail: jylee@saturn.yzu.edu.tw

T-J. Yeh
e-mail: s966047@mail.yzu.edu.tw

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 275, DOI: 10.1007/978-94-007-7684-5_10,

1 Introduction

The Gigabit Passive Optical Networks (GPONs) [1] have been widely considered as the best candidate for next-generation access networks since it represents the high bandwidth, increased flexibility, broad area coverage, higher splitting ratios, and economically viable sharing of the expensive optical links. GPON consists of an optical line terminal (OLT) located at the provider central office (CO) and connect to a number of optical network units (ONUs) at the customer premises by a single splitter/ODN, as illustrated in Fig. 1.

Currently, GPON supports several bit rates in both channels such as asymmetric or symmetric combinations, from 155 Mb/s to 2.5 Gb/s. In downstream, The GPON OLT connects all ONUs as a point-to-multipoint (P2MP) architecture, the OLT transmits encrypted user traffics over the shared bandwidth by broadcasting through the 1:N splitter/ODN on a single wavelength (e.g. 1,490 and 1,550 nm). In upstream, a GPON is a multipoint-to-point (MP2P) network. All ONUs transmit their data to the OLT on a common wavelength (e.g. 1,310 nm) through the 1:N passive combiner. The main problem in the link layer of PON networks is occurred in upstream direction, as all users share the same wavelength, and a medium access control protocol (MAC) is necessary to avoid collisions and efficiently allocate uplink access between packets from different ONUs [2]. Therefore, the time division multiple access (TDMA) [3] is be used to provide shared high-bit-rate bandwidth. In a TDMA scheme, time is divided into periodic cycles, and these cycles are divided into as many time slots as the number of ONUs which shares the channel. As a result, each slot is dedicated to one ONU and every cycle is organized in such a way that one slot transports

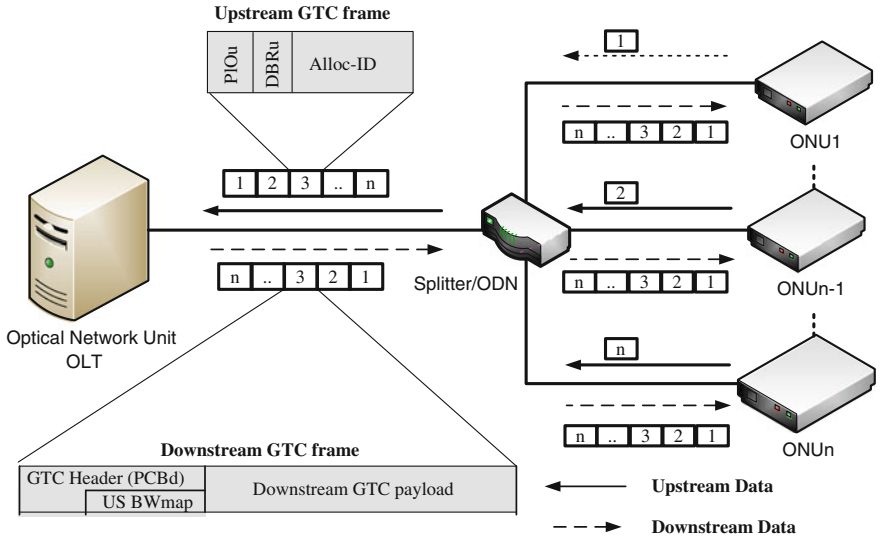


Fig. 1 GPON architecture

packets from one ONU periodically. Meanwhile, GPON OLT supports dynamic broadband algorithm (DBA), making the distribution of available bandwidth to ONU more flexible. They adapt network capacity to the traffic conditions by changing the distribution of the bandwidth assigned to each ONU depending on the current requirements.

The FSAN once seek to control each different traffic stream by means of the MAC protocol to be able to affect the SLA and provide the required quality per user and stream for the QoS support. To the end, logically separate queuing is employed for each flow in each different ONU down to a fine level of resolution. The QoS class is determined by assigning each queue, such as Alloc-ID, to one of five traffic containers (T-CONTs) that follow different service policies [4]: T-CONT1 is based on unsolicited periodic permits granting fixed payload allocations. This is the only static T-CONT not serviced by DBA. T-CONT2 is intended for VBR traffic and applications with both delay and throughput requirements. The availability of bandwidth for the service of this T-CONT is ensured in the SLA, but this bandwidth is assigned only on request to allow for multiplexing gain. T-CONT3 is intended for better than best effort services and offers service at a guaranteed minimum rate; any surplus bandwidth is assigned only on request and availability. T-CONT4 is intended for purely best effort services, and as such is serviced only on bandwidth availability up to a provisioned maximum rate. T-CONT5 is a combined class of two or more of the other four T-CONTs to remove from the MAC controller specification of a target T-CONT when granting access.

Some papers have been investigated the general properties of GPON. The greatest attention is devoted to QoS guarantee by dynamic bandwidth allocation in upstream. Nevertheless, service providers almost always utilized dynamic bandwidth allocation. There is not sufficient attention paid to investigation of an influence of this bandwidth allocation on various period cycle time behaviors. Therefore, we decided to investigate the properties of maximum polling cycle time in case of dynamic bandwidth allocation. Concretely, some key QoS parameters will be observed, i.e. packet delay, jitter, throughput, drop probability and fairness.

Here we focus on an impact of different maximum polling cycle time on QoS parameters in GPON. The allocated bandwidth will be shared between difference maximum polling cycle time and proportion of T-CONTs traffic. The rest of the paper is structured as follows. Section 2 introduces the Dynamic Bandwidth Allocation for GPON system and Sect. 3 presents the Waited-based Dynamic Bandwidth Allocation (WDBA) with theoretical explanations. Performance evaluation and detailed analyses are presented in Sect. 4. Final section concludes the paper and defines future works.

2 Dynamic Bandwidth Allocation for GPON System

In the GPON system, there are two forms of DBA process algorithms, which are Non Status Reporting (NSR) and Status Reporting (SR) operation, the GPON standard provides the tools to implement DBA and leaves the actual bandwidth allocation

scheme open to different implementations. In NSR DBA, the OLT continuously monitor idle frames and surmise traffic status to allocate a small amount of extra bandwidth to each ONU. ONUs do not provide explicit queue occupancy information. Instead, the OLT estimates the ONT queue status, typically based on the actual transmission in the previous cycle. For example, if the OLT observes that a given ONU is not sending idle frames, it increases the bandwidth allocation to that ONU, otherwise, reduces its allocation accordingly. Therefore, NSR ONUs underutilize link capacity, since they do not inform queue occupancy to the OLT as well as traffics in the access network are bursty. NSR DBA has the advantage that it imposes no requirements on the ONU, and the disadvantage that there is no way for the OLT to know how best to assign bandwidth across several ONUs that need more. In SR DBA, All ONTs report their upstream data queue occupancy, to be used by the OLT calculation process. Each ONT may have several T-CONTs, each with its own traffic class. By combining the queue occupancy information and the provisioned SLA of each T-CONT, the OLT can optimize the upstream bandwidth allocation of the spare bandwidth on the PON. For all of these reasons, an efficient DBA algorithm should be SR DBA for GPON system. Therefore, in this paper, the status reporting is employed, which deals with the bandwidth allocation providing more powerful advantages.

In GPON, the scheduling of bandwidth is inherently connected to the 125 μs periodicity of the GTC supper frames and effective DBA algorithms must be tailored for GPON. The connection between the DBA and the frame structure is due to the way DBA messages, Status Reports (SRs), and upstream bandwidth maps (BW maps) are integrated into the GTC frames [5]. In the upstream direction, the upstream frame has the same length and can contain information of several ONUs. Each upstream frame contains the Physical Layer Overhead upstream (PLOu) field. Besides the payload, it may also contain the physical layer operation, administration and management upstream (PLOAMu), the power leveling sequence upstream (PLSu), and the dynamic bandwidth report upstream (DBRu) sections. Furthermore, the GPON specification defines three ways in which one ONU can inform the OLT about its status: sending piggy-backed reports in the upstream DBRu field, using status indication bits in the PLOu field, or including an optional ONU report in the payload. On the other hand, the downstream frame, called GRANT, consists of the physical control block downstream (PCBd) field, the ATM partition, and the GPON encapsulation method (GEM) partition. The OLT sends pointers in the PCBd field, that each of them indicating the time at which each ONU starts and ends its upstream transmission. This granting process performance allows that only one ONU can access the shared channel at the same time. Figure 2 shows a schematic view of the GPON SR DBA process with the OLT-ONU communication resulting in an updated bandwidth assignment.

In general, because of the non-negligible DBA execution time and round-trip time leading the idle period problem in GPON. As shown in Fig. 2, the *idle period* is given by

$$Idle\ Period = T_{DBA} + RTT, \tag{1}$$

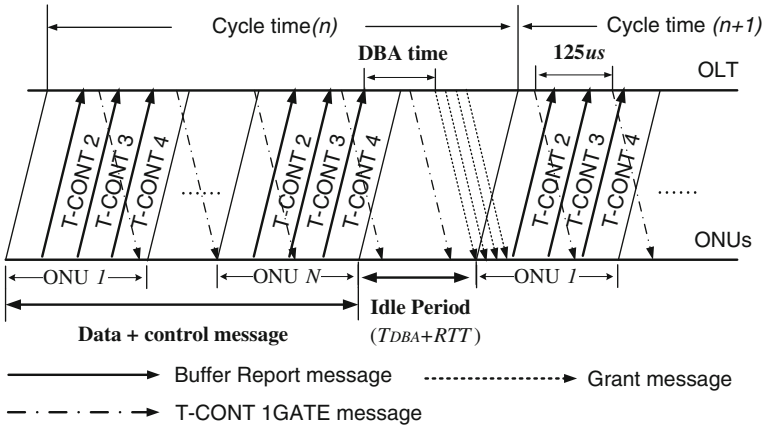


Fig. 2 Idle period issue in a GPON status reporting DBA polling scheme

where *RTT* is the round-trip time from ONU to OLT and T_{DBA} is the processing time of the DBA algorithm. To elaborate, the idle period is sum of computation time of DBA and the round trip time between OLT and each ONU (N. B. ONUs cannot transmit data during the idle period). Hence, for the DBA scheme, reducing the idle period becomes one of the main challenges issues to address in order to improve bandwidth utilization. Moreover, status report information from an ONU may be outdated will causing another problem—*queue state inconsistency* due to packets that continue to arrive during this waiting time. In a detailed, each ONU experiences a waiting time between sending the REPORT message and sending the buffered frames. Consequently, packets that arrive during the waiting time and transmission time, have to be delayed to the next transmission cycle, potentially leading to stupendous packet delay and delay jitter. The above problems are even worse in Next Generation PON (NG-PON) owing to increased upstream rates and distances between OLT and ONUs. In order to decrease packet delay and improve overall fairness, predictive schemes can be used so that traffic arrival during the waiting time and transmission time is taken into consideration.

To date, in order to decrease packet delay and improve overall fairness, various prediction-based DBA algorithms have been proposed for GPON networks. In the prediction-based DBA algorithms, the extra bandwidth will be allocated to each ONU for the next cycle $(N + 1)$, the current bandwidth demand in cycle N should be takes into consideration. The assigned bandwidth for the next cycle is calculated by ONU request bandwidth demand plus a prediction term based on the algorithms of each scholars, it may be the constant credit-based [6, 7], SLA-based [8–10], class-based [11, 12] etc. However, most of the above proposals are aware of the fact that delay sensitive traffic should be treated in a specialized manner within the OLT DBA stage. Nevertheless, these schemes do not address the investigation of an influence of this bandwidth allocation on various period cycle time behaviors. In this paper we propose a Waited-based Dynamic Bandwidth Allocation (WDBA)

mechanism that uses a different granting scheme for GATE messages to improve QoS support by predict the arriving packet based on the proportion of waiting time and history request bandwidth for multiple services over GPONs. Moreover, the WDBA is adapted to variable scheduling frame size and guarantees a minimum bandwidth for each ONU in every polling cycle and observe that the impact between upstream channel utilization and maximum polling cycle time with different proportion of T-CONT traffic in a GPON.

3 Waited-Based Dynamic Bandwidth Allocation Mechanism

The motivation of this paper is to resolve the *idle period problem* and *queue state inconsistency* to improve the uplink bandwidth utilization, reduce the packet latency, and provide better QoS guarantee, regardless of the environment that whether the uplink is under different traffic load and proportion of T-CONT traffic. Moreover, this work focus on the fundamental problem of trading-off between upstream channel utilization and maximum polling cycle time in a GPON system. To achieve this goal, the proposed WDBA scheme, combined the waited-based prediction scheme with Limit Bandwidth Allocation (LBA) and Excess Bandwidth Allocation (EBR) scheme, applies service level agreement (SLA) scheduling policy [7] at each ONU to guarantee Quality of Service (QoS) ensures maximum delay and jitter of real-time traffic, uplink bandwidth utilization, drop probability and fairness in GPON upstream transmission. At the same time, an interleaved scheduling is also introduced, which is our previous work [13, 14] to support different services and the different classes of service require differential performance bounds. The interleaved scheduling can not only resolve the idle period problem caused by MPCP protocol scheduling policy-"grant after report", but also reduce the packet delay and increase the fairness between heavily-loaded ONUs and lightly-loaded ONUs. Moreover, this paper focus on the relationship between the maximum cycle time and the system performance, instead of only to the QoS as the traditional DBA scheme does. The WDBA mechanism is detailed as follow.

3.1 Interleaved Scheduling of WDBA

The interleaved scheduling of WDBA is proposed to resolve the idle period problem and improve bandwidth utilization by using interleaved scheduling transmission. The interleaved scheduling algorithm divides one transmission cycle time into two groups and dynamically adjusts the bandwidth between the first group and the second group to execute interleaved transition to resolve the traditional idle period problem. The first group and the second group will be performed in accordance with the number of the ONU evenly. Moreover, the T-CONT 1 traffic is transmitted with guaranteed fixed bandwidth allocation for time-sensitive applications, and the T-CONT 2 traffic

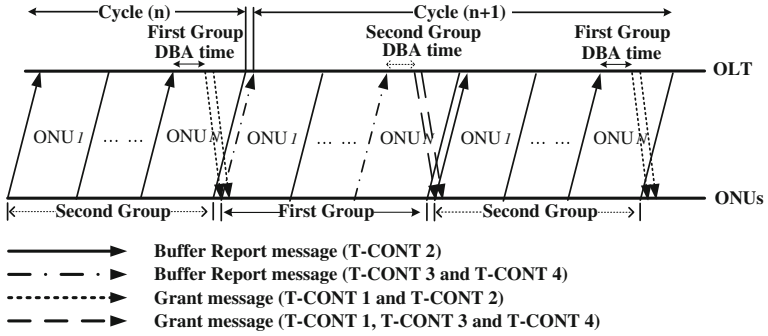


Fig. 3 Interleaved scheduling for waited-based dynamic bandwidth allocation

is transmitted with prediction mechanism for the guaranteed assured bandwidth allocation and not time-sensitive applications to alleviate queue state inconsistency problem; recycling the remaining bandwidth from the first group for the second group to obtain maximum performance. Figure 3 shows a schematic view of the interleaved scheduling for Waited-based Dynamic Bandwidth Allocation process in the OLT-ONU upstream communication.

3.2 Dynamic Bandwidth Allocation of WDBA

The WDBA mechanism is proposed to resolve the idle period problem, and enhance the QoS for differentiated services and improve bandwidth utilization by using prediction, LBA and EBR in the GPON system. The flowchart of the WDBA mechanism is illustrated in Fig. 4, after receiving whole REPORT messages from each ONU, the total available bandwidth B_{av} can be calculated as $r \times (T_{Cycle}^{Max} - N \times T_g) - N \times 16$, where r is the transmission speed of the GPON in bits per second, T_{Cycle}^{Max} is the maximum cycle time, N is the number of ONUs, T_g is the guard time and the control message length is 16bits (2Byte) [1] for the GPON system. Initially, the available bandwidth for each group and the minimum guaranteed bandwidth ($B_{min} = B_{av}/N$) for ONU$_i$ are evenly distributed. After calculating minimum guaranteed bandwidth threshold, the WDBA executes the prediction mechanism based on the proportion of waiting time, historical and current traffic status information, which is expressed in Eq. (2).

$$\begin{cases} P_{T2} = R_i^{T2} + (\overline{H_i^{T2}} \times (\frac{T_{waiting,i}}{T_{cycle,i}})) \\ P_{T3} = R_i^{T3} + (R_i^{T3} - \overline{H_i^{T3}}),\ R_i^{T3} - \overline{H_i^{T3}} > 0 \\ P_{T4} = R_i^{T4} + (R_i^{T4} - \overline{H_i^{T4}}),\ R_i^{T4} - \overline{H_i^{T4}} > 0 \end{cases}, \tag{2}$$

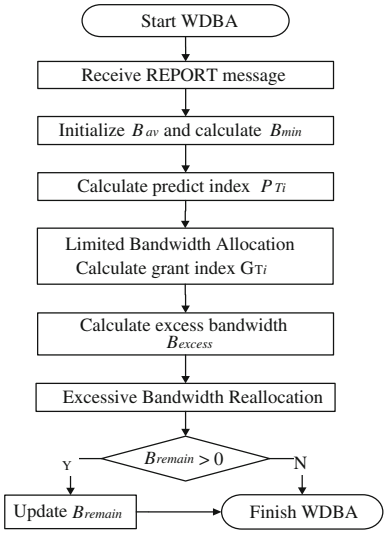


Fig. 4 Flowchart of waited-based dynamic bandwidth allocation

where P_T represents predict index R_i^T represents bandwidth request of each T-CONT of ONU$_i$, and $\overline{H_i^T}$ is the average bandwidth requirements of the history ten cycle of each T-CONT of ONU$_i$, where T-CONT 1, T-CONT 2, T T-CONT 3, T-CONT 4. For the T-CONT 2 and T-CONT 3 traffic, the predict index can be update when the $R_i^{T3} - \overline{H_i^{T3}}$ is bigger than zero, otherwise, the predict index is equals to request bandwidth.

During dynamic allocation, the allocated timeslot will be adapted to the requested bandwidth. To prevent the bandwidth wasted, the limited bandwidth allocation mechanism follow SLA and compares the minimum guaranteed bandwidth threshold with the predicted index of each ONU to get the grant bandwidth index (G_{Ti}), which is expressed in Eq. (3).

$$\begin{cases} G_{T2} = \min(P_{T2}, B_{\min}) \\ G_{T3} = \min(P_{T3}, B_{\min} - P_{T2}) \\ G_{T4} = \min(P_{T4}, B_{\min} - P_{T2} - P_{T3}) \end{cases}. \tag{3}$$

In the Excessive Bandwidth Reallocation (EBR) mechanism, the excess bandwidth can be collected from lightly-loaded ONUs and redistributed among the heavily-loaded ONUs. The sum of underutilized bandwidth of lightly-loaded ONUs is called excessive bandwidth (B_{excess}), which can be expressed as follow:

$$B_{excess} = \sum_{j \in L} (B_{\min} - \sum_{i=2}^{4} G_{Ti}), \ B_{\min} > \sum_{i=2}^{4} G_{Ti}, \tag{4}$$

where L is the set of lightly-loaded ONUs and j is a lightly-loaded ONU in L. In the end, a heavily-loaded ONU obtains an additional bandwidth based on the EBR mechanism. If the bandwidth has not yet been distributed to the heavily-loaded ONUs after B_{excess} has been allocated, the remaining available bandwidth (B_{remain}) can be reserved for the next group of ONUs for DBA. The B_{remain} is expressed in Eq. (5) as follows:

$$B_{remain} = B_{excess} - \sum_{j \in H} (\sum_{i=2}^{4} G_{Ti} - B_{\min}), \; B_{\min} < \sum_{i=2}^{4} G_{Ti}, \tag{5}$$

where H is the set of heavily-loaded ONUs and j is a heavily-loaded ONU in H. Therefore, the WDBA can support QoS and enhance system performance for differential services and efficiently reallocates excessive bandwidth in GPON.

4 Performance Evaluation

In this section, the system performance of WDBA mechanism is compared between various maximum polling cycle time: 1, 1.5 and 2 ms in terms of the throughput, end-to-end delay, drop probability, jitter and fairness for 16 ONUs. The GPON simulation model, set up by the OPNET modeler network simulator, the upstream/downstream link capacity is 1.24 Gbps, the OLT-ONU distance is 10–20 km, the buffer size is 10 MByte, the guard time is 1.8 μs and the computation time of DBA is 10 μs. The service policy follows the first-in first-out (FIFO) principle. The T-CONT 1 traffic has the deterministic efficacy with limits is anticipated. For the traffic model considered, an extensive study has shown that most network traffic can be characterized by self-similarity and long-range dependence (LRD) [15]. The packet size generated each time for T-CONT 2, T-CONT 3 or T-CONT 4 traffic is 64, 500, 1500 bytes with probability of 60, 20 and 20 %, respectively [16]. The traffics with minimum assured bandwidth and with additional non-assured bandwidth of T-CONT 3 are assumed to distribute evenly. In order to observe the effective of high priority traffic, the proportion of traffic profile is analyzed by simulating the four significant scenarios in (T-CONT 1, T-CONT 2, T-CONT 3, and T-CONT 4) with (10 %, 60 %, 20 %, 10 %, 1,621), (10 %, 40 %, 30 %, 20 %, 1,432), (40 %, 30 %, 20 %, 10 %, 4,321) and (40 %, 40 %, 10 %, 10 %, 4,411), respectively. The simulation parameters are summarized in Table 1.

4.1 Throughput

Figure 5 shows the throughput comparisons of average, T-CONT 2, and T-CONT 3 of different maximum pooling cycle time in 16 ONUs with different proportions of traffic profile for different traffic loads. In the simulation results show that the

Table 1 Simulation scenario

Item	Parameter			
Number of ONUs in the system	16			
Upstream/downstream link capacity	1.24 Gbps			
OLT-ONU distance (uniform)	20 km			
Buffer size	10 MB			
Maximum transmission cycle time	1 ms, 1.5 ms, 2 ms			
Guard time	1.8 μs			
Computation time of DBA	10 μs			
Traffic proportion	T1	T2	T3	T4
	10 %	40 %	30 %	20 %
	10 %	60 %	20 %	10 %
	40 %	30 %	20 %	10 %
	40 %	40 %	10 %	10 %

proportion of traffic will lead greater impact to average and T-CONT 2, and T-CONT 3 throughput. When the proportion of high priority traffic is higher, the average throughput is better than the others proportion ratio, the main reason is that the LBA in WDBA follow the SLA policy and the T-CONT 2 has highest priority to get the bandwidth. Moreover, the traffic flow of T-CONT2 is higher result in higher bandwidth throughput of T-CONT2. Moreover, the maximum polling cycle time will affect the system throughput when the traffic loading exceeding 50 %, and result in greater impact of higher proportion of high-priority traffic scenario. Moreover, the throughput of T-CONT 3 will increasing when the maximum polling cycle time increasing, because of the more frequent communication between OLT and ONU wiil lead to the control messages and guard time increases, resulting in bandwidth wasted, this situation will tends to be more obvious when the number of ONU increase.

4.2 Packet Delay

Figure 6 shows the packet delay comparisons of average, T-CONT 2, and T-CONT 3 packet delay of WDBA in 16 ONUs with different proportions of traffic profile for different traffic loads. The packet delay d is equal to $d = d_{poll} + d_{grant} + d_{queue}$, packet delay ($d$) consists of polling delay (d_{poll}), granting delay (d_{grant}) and queuing delay (d_{queue}). Simulation results show that the average packet delay of WDBA in 1 ms maximum polling cycle time have relatively poor performance because of they can not transmit more data in short cycle time. In the different proportion of traffic, the scenario 1621 has the worst performance and begin to increase when the traffic load exceeding 50 %, the scenario 4321 has the best performance, the reason is that the T-CONT 2 in WDBA will be transmitted early, so that the T-CONT 2 can get more resource and higher service level but has worse performance in other T-CONT

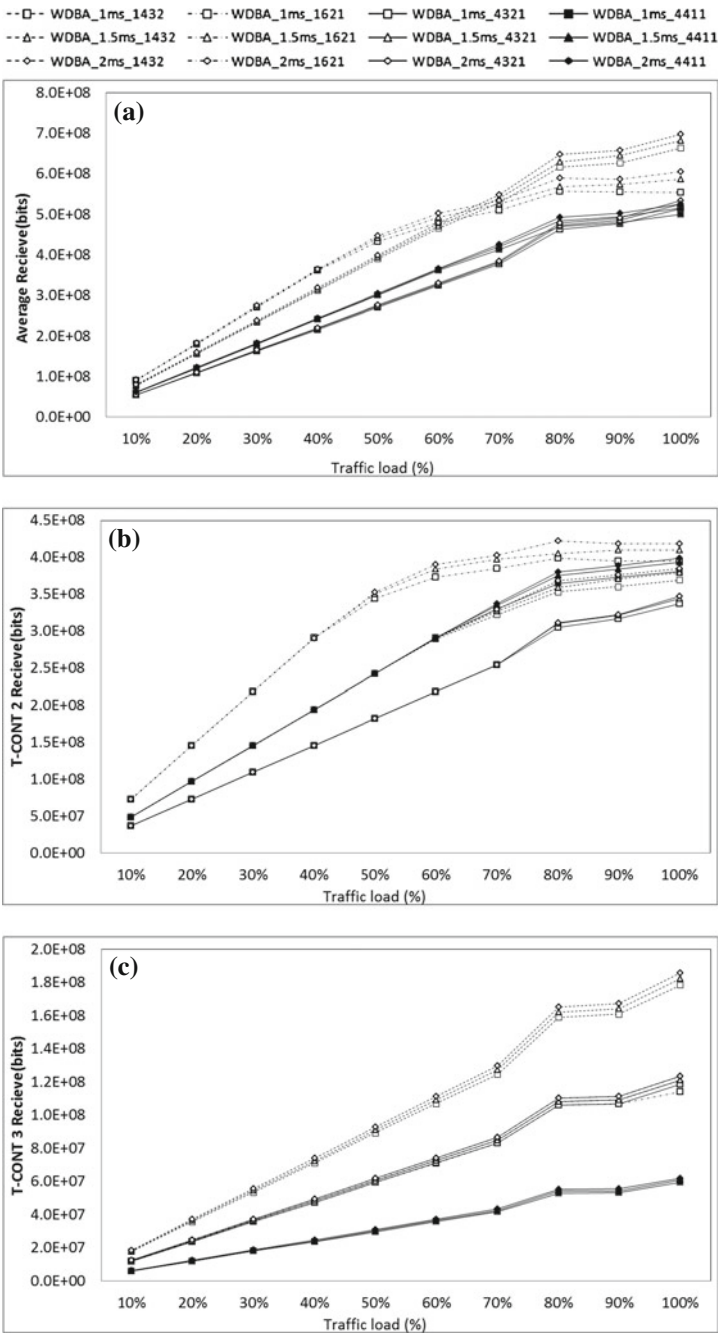


Fig. 5 **a** Average, **b** T-CONT 2, **c** T-CONT 3 throughput

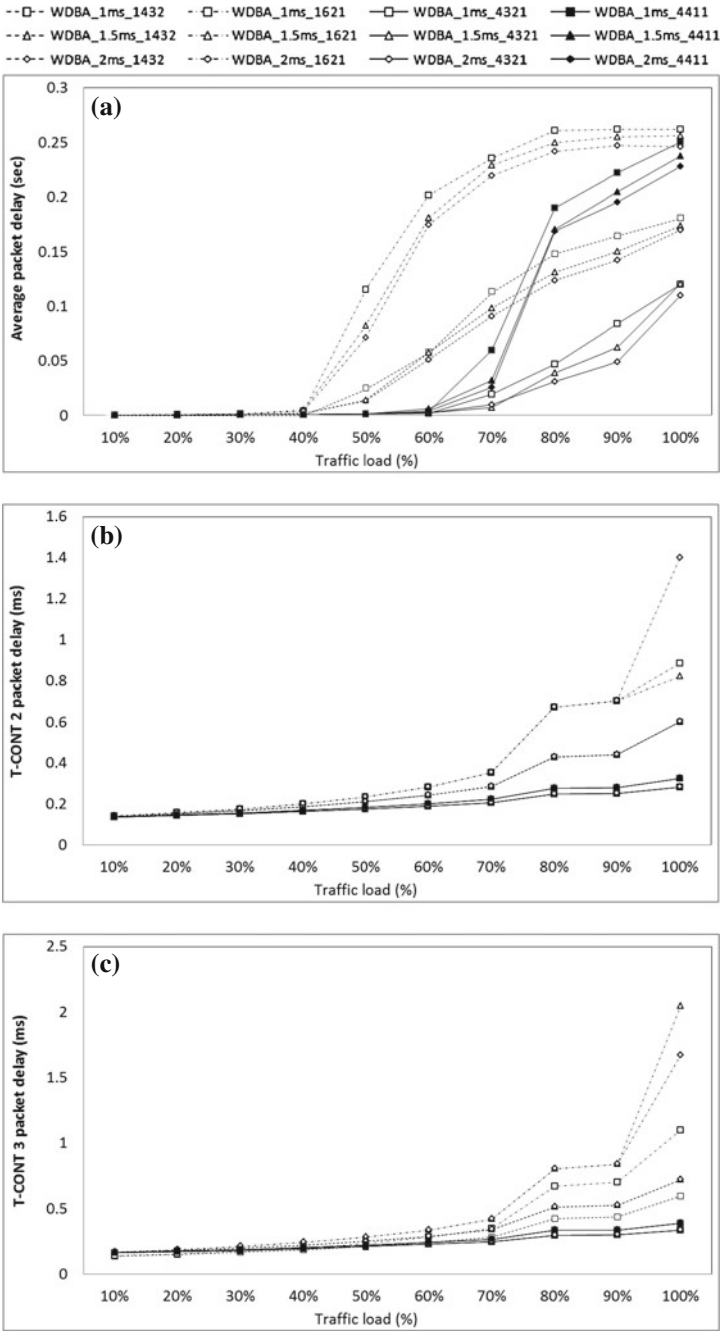


Fig. 6 **a** Average, **b** T-CONT 2, **c** T-CONT 3 packet delay

traffic. Moreover, the packet delay of high priority traffic can be guaranteed but the packet delay of low priority traffic will increasing when the maximum polling cycle time is shorter, because of the high priority traffic can get the service more faster. On the other hand, the packet delay of low priority traffic will decreasing when the maximum polling cycle time is longer, because of the low priority traffic can get more chance to transmit in a single polling cycle time.

4.3 Drop Probability, Jitter and Fairness

Figure 7 compares the drop probability, jitter and fairness of the WDBA in 16 ONUs with different proportions of traffic profile for different traffic loads. Simulation results show that the blocking probability of WDBA scheme in scenario 1432 is worst and the scenario 4321 has the best performance except traffic loading is 100 %, the reason is that scenario 1432 has lower total throughput and the allocated bandwidth of T-CONT 3 and T-CONT 4 must wait for the remaining unused bandwidth given from T-CONT 2. Therefore, the traffic data of T-CONT 3 and T-CONT 4 will be queued in the buffer especially for T-CONT 4.

The delay variance σ^2 is calculated as $\sigma^2 = \sum_1^N (d_i^{T2} - \overline{d})^2/N$, where d_i^{T2} represents the delay time of T-CONT 2 packet i and N is the total number of received T-CONT 2 packets. Simulation results show that the delay variance for T-CONT 2 traffic increases as the traffic load increases especially in scenario 4411. The T-CONT 2 jitter of WDBA is increasing when the traffic load exceeding 50 % for scenario 1621, and exceeding 70 % for the others. The reason is that the transmission order of each ONU is sequential and that the T-CONT 2 jitter of WDBA is depends on the cycle time when the proportion of high priority traffic is higher the jitter getting worse.

The global fairness index $f (0 \leq \mathrm{f} \leq 1)$ has been addressed [17] which is defined as:

$$f = \frac{\left(\sum_{i=1}^{N} G_{[i]}\right)^2}{N \sum_{i=1}^{N} G_{[i]}^2}, \tag{6}$$

where N is the total number of ONUs and $G[i]$ is the ratio between the granted bandwidth of ONUi and requirement of ONUi. Simulation results show that the proportion of high priority traffic is higher and the shorter maximum polling cycle time will leading good fairness. The reason is that the high proportion of VBR traffic will cause significant changes in the system data flows compared with the high proportion of CBR traffic.

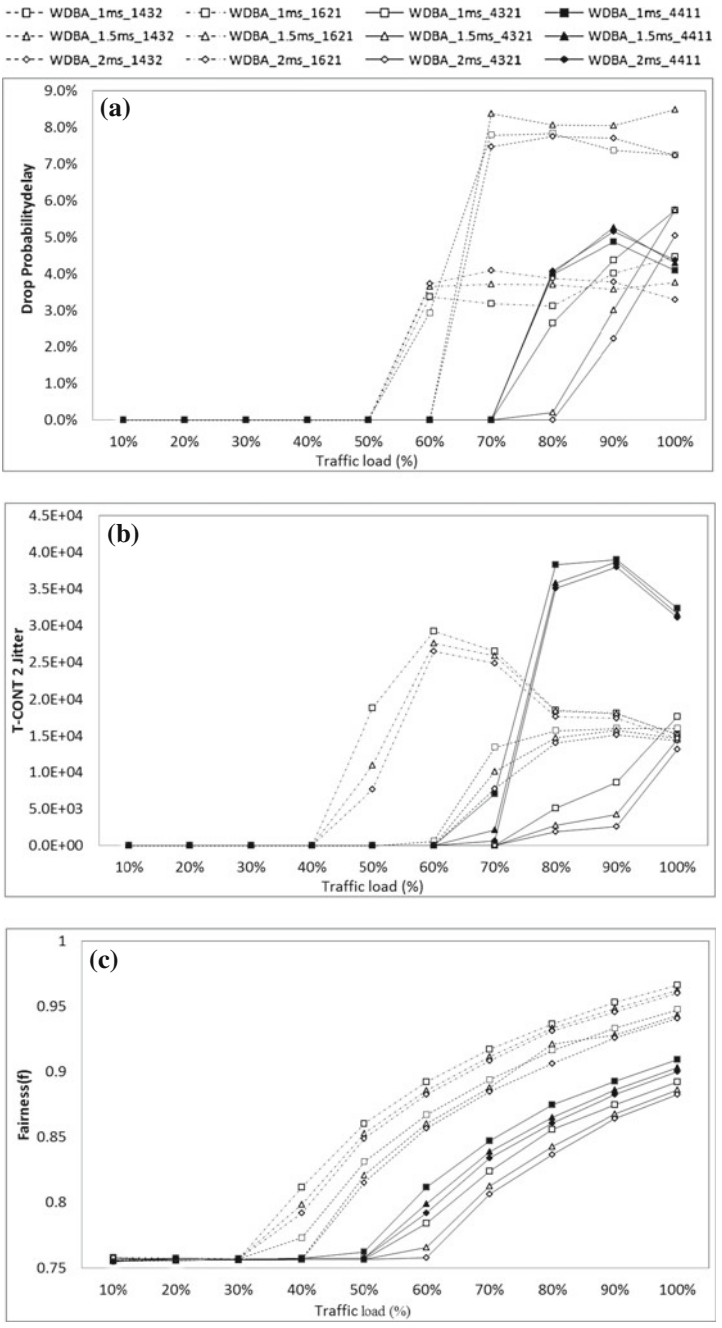


Fig. 7 **a** Drop probability, **b** T-CONT 2 Jitter, **c** Global fairness

5 Conclusion

In this study, important factors that can improve the performance of GPON are discussed and evaluated. The WDBA mechanism executes an interleaved transmission process to automatically adjust cycle time to resolve the *idle period problem* for traditional DBA scheme, enhancing the system performance to reduce end-to-end packet delay and improving the throughput. In performance evaluation, the proportion of T-CONTs and the different maximum polling cycle time will lead different simulation result. The packet delay, throughput and drop probability performance is better when the polling cycle time is longer; the fairness is better when the polling cycle time is shorter. Moreover, when the maximum polling cycle time is longer will reducing the control massage and guard time in the upstream transmission and result in better throughput efficiency and packet delay performance for low priority traffic; on the other hand, when the maximum polling cycle time is decreasing will be able to quickly communication between the OLT and ONU result in better fairness performance, packet delay and jitter performance for high priority traffic. The situation described above will become more apparent when the ONU number increases.

References

1. ITU-T standardization (2008) G.984.1: Gigabit-capable passive optical networks (GPON): General characteristics. Available at http://www.itu.int/rec/TREC-G.984.1-200803-P/en/
2. López A, Merayo N, Martínez JJ, Fernández P (2012) Fiber to the home through passive optical networks, WDM systems and networks optical, Networks, 2012, pp 337–372
3. Brunnel H (Jul. 1986) Message delay in TDMA channels with contiguous output. IEEE Trans Commun 34(7):681–684
4. Angelopoulos JD, Leligou HC (2004) Theodore Argyriou and Stelios Zontos, Efficient transport of packets with QoS in an FSAN Aligned GPON, IEEE communications magazine vol 42, Issue 2, 2004
5. Iniewski K (2010) Dynamic bandwidth allocation in EPON and GPON, in convergence of mobile and stationary next-generation networks, John Wiley and Sons, pp 226–251
6. Hakjeon B, Sungchang K, Dong-Soo L, Chang-Soo P (Feb. 2010) Dynamic bandwidth allocation method for high link utilization to support NSR ONUs in GPON. Int Conf Adv Commun Technol (ICACT) 1:884–889
7. Chang CH, Kourtessis P, Senior JM (Sep. 2006) GPON service level agreement based dynamic bandwidth assignment protocol. IET Electron Lett 42(20):1173–1174
8. Berisa T, Bazant A, Mikac V (Apr. 2009) Bandwidth and delay guaranteed polling with adaptive cycle time (BDGPACT): A scheme for providing bandwidth and delay guarantees in passive optical networks. J Opt Netw 8(4):337–345
9. Jiang J, Senior JM (Aug. 2009) A new efficient dynamic MAC protocol for the delivery of multiple services over GPON. Photonic Netw Commun 18(2):227–236
10. Hamada A, Mohamed A (2011) An end-to-end QoS mechanism for GPON access networks. IEEE GCC conference and exhibition (GCC), pp 513–516
11. Skubic B, Chen B, Jiajia C, Ahmed J, Wosinska L (2009) Improved scheme for estimating T-CONT bandwidth demand in status reporting DBA for NG-PON, Communications and photonics conference and exhibition (ACP), pp 1–6
12. Kanonakis K, Tomkos" I (Aug. 2009) Offset-based scheduling with flexible intervals for evolving GPON networks. J Lightwave Technol 27(15):3259–3268

13. Hwang IS, Lee JY, Liem AT (Feb. 2012) A bi-partitioned dynamic bandwidth allocation mechanism for differentiated services to support state report ONUs in GPON. J Comput Inf Syst 8(2):675–682
14. Hwang IS, Lee JY, Yeh TJ (2013) Polling cycle time analysis for waited-based DBA in GPONs. International conference on communication systems and applications: proceedings of the international multi conference of engineers and computer scientists, Hong Kong, 13–15 Mar 2013, pp 608–613
15. Willinger W, Taqqu MS, Erramilli A (1996) A bibliographical guide to self-similar traffic and performance modeling for modern high-speed networks. Stochastic networks: theory and applications, royal statistical society lecture notes series, vol 4, Oxford University Press
16. Fraleigh C, Moon S, Lyles B, Cotton C, Khan M, Moll D, Rockell R (Nov. 2003) Packet-level traffic measurements from the Sprint IP backbone. IEEE Netw 17(6):6–16
17. Jain R, Durresi A, Babic G (1999) Throughput fairness index: an explanation. http://www.cs.wustl.edu/jain/atmf/ftp/af_f.X

Island-Model-Based Distributed Modified Extremal Optimization for Reducing Crossovers in Reconciliation Graph

Keiichi Tamura, Hajime Kitakami and Akihiro Nakada

Abstract To determine the mechanism of molecular evolution, molecular biologists need to carry out reconciliation work. In reconciliation work, they compare the relation between two heterogeneous phylogenetic trees and the relation between a phylogenetic tree and a taxonomic tree. Phylogenetic trees and taxonomic trees are referred to as ordered trees and a reconciliation graph is constructed from two ordered trees. In the reconciliation graph, the leaf nodes of the two ordered trees face each other. Furthermore, leaf nodes with the same label name are connected to each other by an edge. To perform reconciliation work efficiently, it is necessary to find the state with the minimum number of crossovers of edges between leaf nodes. Reducing crossovers in a reconciliation graph is the combinatorial optimization problem that finds the state with the minimum number of crossovers. In this chapter, a novel bio-inspired heuristic called island-model-based distributed modified extremal optimization (IDMEO) is presented. This heuristic is a hybrid of population-based modified extremal optimization (PMEO) and the distributed genetic algorithm using the island model that is used for reducing crossovers in a reconciliation graph.

Keywords Bio-inspired heuristic · Combinatorial optimization · Extremal optimization · Island model · Reconciliation graph · Reducing crossovers

K. Tamura (✉) · H. Kitakami · A. Nakada
Graduate School of Information Sciences, Hiroshima City University,
3-4-1 Ozuka-Higashi, Asa-Minami-Ku, Hiroshima 731-3194, Japan
e-mail: ktamura@hiroshima-cu.ac.jp

H. Kitakami
e-mail: kitakami@hiroshima-cu.ac.jp

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 275, DOI: 10.1007/978-94-007-7684-5_11,

1 Introduction

Molecular biologists need to carry out reconciliation work [1–3] in order to determine the mechanism of molecular evolution. In reconciliation work, the relation between two heterogeneous phylogenetic trees and the relation between a phylogenetic tree and a taxonomic tree are compared. To compare two heterogeneous trees, a graph called a reconciliation graph that consists of two heterogeneous phylogenetic trees or a phylogenetic tree and a taxonomic tree are constructed. In a reconciliation graph, phylogenetic trees and taxonomic trees are referred to as ordered trees. The leaf nodes of these ordered trees face each other and leaf nodes with the same label name are connected to each other by an edge.

To perform reconciliation work efficiently, it is necessary to find the state with the minimum number of crossovers of edges between leaf nodes in the reconciliation graph. For example, in Fig. 1, phylogenetic tree 1 and phylogenetic tree 2 are inferred from different molecular sequences with four identical species "a", "b", "c" and "d." The leaf nodes of phylogenetic tree 1 and those of phylogenetic tree 2 face each other. Moreover, leaf nodes representing the same species are connected to each other. The reconciliation graph shown in Fig. 1a has two crossovers. If node "1" and node "d" are replaced, the optimal reconciliation graph shown in Fig. 1b, which has no crossovers, is obtained.

Reducing crossovers in a reconciliation graph is the combinatorial optimization problem that finds the state with the minimum number of crossovers. The number of combinations increases exponentially as the number of leaf nodes increases. Thus, some heuristics, which use a genetic algorithm (GA) [4], were proposed [5, 6]. Recently, alternatives to GA-based heuristics have been proposed and they use extremal optimization (EO) [7–9], modified EO (MEO) [10], and population-based MEO (PMEO) [11]. EO is a general-purpose heuristic inspired by the Bak-Sneppen model [12] of self-organized criticality from the field of statistical physics. MEO improves the methodology of alternation of generations. Although EO selects a neighbor solution randomly at an alternation of generations, MEO selects the best

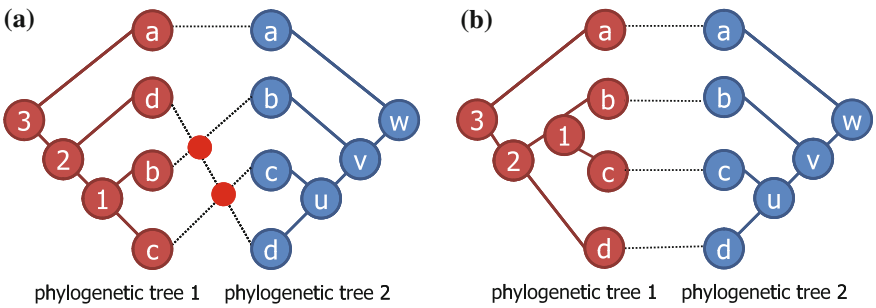


Fig. 1 Examples of reconciliation graphs. **a** shows a reconciliation graph that has two crossovers, and **b** shows a reconciliation graph that has no crossovers

solution in multiple neighbor solutions. The performance of MEO is better than that of EO, however, it depends on an initial individual. To overcome this difficulty, PMEO utilizes the population-based approach.

In this chapter, a novel extremal optimization model called island-model-based distributed modified extremal optimization (IDMEO) [13] for reducing crossovers in a reconciliation graph is proposed. IDMEO is a hybrid of PMEO and the distributed genetic algorithm (DGA) [14] using the island model [15]. The performance of PMEO is better than that of MEO, however, it is difficult to maintain diversity at the end of alternation of generations. In the island model, a population is divided into two or more sub-populations called islands and each island evolves individually. Each island can maintain different types of individuals at the end of alternation of generations. Therefore, IDMEO can maintain diversity at the end of alternation of generations.

Many studies [7–9, 16–19] have proposed the EO-based heuristics for combinatorial optimization problems such as the traveling salesman problem, graph partitioning problem, and image rasterization. Recently, some studies [20–22] have focused on integrating the population-based approach in EO. To the best of our knowledge, there is no study on PMEO involving the distributed genetic algorithm using the island model. To evaluate IDMEO, we implemented IDMEO for reducing crossovers in a reconciliation graph. Moreover, we evaluated IDMEO using two actual data sets for experiments. Experimental results shows that IDMEO outperforms PMEO. Moreover, we compared IDMEO with another population-based heuristic based on genetic algorithm with minimal generation gap (MGG) [23], which is the one of the best generation alternation models. The performance of IDMEO also is better than that of MGG.

The rest of the chapter is organized as follows. In Sect. 2, the problem definition is presented. In Sect. 3, related work is reviewed. In Sect. 4, we explains MEO and PMEO. In Sect. 5, the details of IDMEO is described. In Sect. 6, experimental results are presented, and Sect. 7 is the conclusion of this chapter.

2 Problem Definition

A reconciliation graph (RG) consists of two ordered trees, $OT_1 = (V_1, E_1)$ and $OT_2 = (V_2, E_2)$, where V_1 and V_2 are finite sets of nodes and E_1 and E_2 are finite sets of edges. A node that has no child nodes is a leaf node. The leaf node sets of OT_1 and OT_2 are denoted by $L_1 \in V_1$ and $L_2 \in V_2$, respectively. If the number of species is n, the number of leaf nodes is n. A leaf node has a label name, which is a species' name. The label name set is denoted by L_{leaf}.

In the reconciliation graph, OT_1 and OT_2 are located face to face. If a leaf node of OT_1 has the same label name as that of OT_2, then the two leaf nodes are connected to each other. In Fig. 2, phylogenetic tree 1 is OT_1 and phylogenetic tree 2 is OT_2. The leaf node set L_1 has four nodes, v_{14}, v_{15}, v_{16}, and v_{17}. Similarly, L_2 has four nodes,

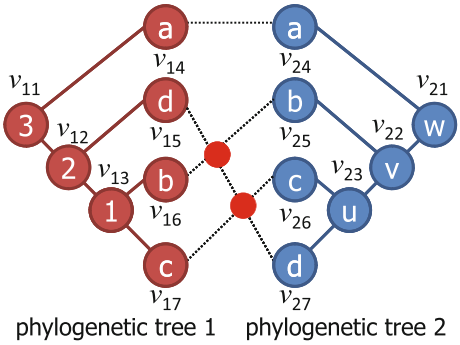


Fig. 2 Problem definition (OL_1 is given by $OL_1 = [v_{14}, v_{15}, v_{16}, v_{17}]$ and OL_2 is given by $OL_2 = [v_{24}, v_{25}, v_{26}, v_{27}]$)

v_{24}, v_{25}, v_{26}, and v_{27}. There are four label names in L_{leaf}, "a," "b," "c," and "d." Two leaf nodes v_{14} and v_{24} are connected because they have the same label name "a."

Let OL_1 and OL_2 be the order lists of leaf nodes:

$$OL_1 = [ol_{1,1}, ol_{1,2}, \cdots, ol_{1,n}](ol_{1,i} \in L_1, \mathscr{L}(ol_{1,i}) \in L_{leaf}),$$
$$OL_2 = [ol_{2,1}, ol_{2,2}, \cdots, ol_{2,n}](ol_{2,i} \in L_2, \mathscr{L}(ol_{2,i}) \in L_{leaf}),$$

where function $\mathscr{L}$ returns the label name of an input node. The function $C(M)$ returns the number of crossovers:

$$C(M) = \sum m_{j,\beta} m_{k,\alpha} [1 \le j < k \le n, 1 \le \alpha < \beta \le n], \tag{1}$$

where $m_{i,j}$ is (i, j)th-element of the connection matrix M that is defined as

$$m_{i,j} = \begin{cases} 1 & if\ \mathscr{L}(ol_{1,i}) = \mathscr{L}(ol_{2,j}), \\ 0 & otherwise. \end{cases} \tag{2}$$

In Fig. 2, OL_1 is given by $OL_1 = [v_{14}, v_{15}, v_{16}, v_{17}]$. Similarly, there are four leaf nodes in phylogenetic tree 2, $ol_{2,1} = v_{24}$, $ol_{2,2} = v_{25}$, $ol_{2,3} = v_{26}$, and $ol_{2,4} = v_{27}$. Therefore OL_2 is given by $OL_2 = [v_{24}, v_{25}, v_{26}, v_{27}]$. For example, the $(0, 0)$th-element $m_{0,0}$ is 1 because $\mathscr{L}(v_{14})$ equals $\mathscr{L}(v_{24})$. Similarly, the $(1, 1)$th-element $m_{1,1}$ is 0 because $\mathscr{L}(v_{15})$ does not equal $\mathscr{L}(v_{25})$.

$$M = \begin{array}{c} \\ a \\ d \\ b \\ c \end{array} \begin{array}{c} \begin{array}{cccc} a & b & c & d \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

The task of reducing crossovers in the reconciliation graph is defined as follows:

min: $C(M)$,
subject to: (1)M is the connection matrix of the RG,
(2) There are no crossovers on edges between non-leaf nodes in the RG

There should be no crossovers on edges between non-leaf nodes in the reconciliation graph. For this constraint, we need to change order of leaf nodes by changing the order of child nodes in intermediate nodes. We cannot change the order between v_{15} and v_{17} (Fig. 2) because it will lead to the presence of crossovers on edges between non-leaf nodes. If we want to change the order between v_{15} and v_{17}, it is necessary to replace v_{15} and v_{13}, which are child nodes of v_{12}. If we replace v_{15} and v_{13}, the number of crossovers in the reconciliation graph becomes zero, and OL_1 is changed to $OL_1 = [v_{14}, v_{16}, v_{17}, v_{15}]$.

3 Related Work

Molecular biologists used to perform reducing crossovers in reconciliation graphs manually. However, with increase in the number of nodes in a reconciliation graph, it is very difficult to make it manually. Hence, some computational heuristics to reduce crossovers automatically in a reconciliation graph were proposed. The most simplest computational heuristic was proposed in [6]. This simplest heuristic could obtain only a local optimal solution with a kind of local search. To improve the performance, a GA-based heuristic was proposed in [5]. There are two steps in the GA-based heuristic. First, the GA-based heuristic searches quasi-optimal solutions with simple GA. Second, the GA-based heuristic finds more better solutions from quasi-optimal solutions by using local search.

The GA-based heuristic can reduce crossovers automatically, however, it has a performance issue. This is because that it is difficult to design efficient crossover functions. One of the main performance issues is that the speed of convergence slow. Therefore it need huge computation time to get optimal solutions. To overcome this difficulty, modified Extremal Optimization (MEO) [10], which is a EO-based heuristic, was proposed. The EO mechanism [7–9] follows the spirit of the Bak-Sneppen model, updating variables that have among the worst values in a solution and replacing them by random values without ever explicitly improving them. In other word, EO evolves a single individual by making local modifications to the worst components in the individual. MEO improves the methodology of alternation of generations The experimental results show that MEO outperforms EO. Moreover, MEO is good performance compared with the GA-based heuristic.

The performance of MEO is good, however, its performance depends on an initial individual. To address this issue, proposed population-based modified extremal optimization (PMEO) was proposed [11]. PMEO is a combination of the population-based approach and MEO. Recently, there are some studies [20–22] on population-based EO. These heuristics are based on EO. Our approach uses MEO for changing

state of individual. Multiple individuals are only used in [21, 22]. In our approach, not only multiple individuals are used but also restrictive crossover is performed between individuals. Our approach is most similar with the approach of [20]. However, [20] is a hybrid of partial swam optimization (PSO) and EO. This hybrid approach is only performing EO as mutation. On the other hand, PMEO repeats alternation of generations by using MEO.

4 Population-Based Modified Extremal Optimization

EO follows the spirit of the Bak-Sneppen model, updating variables that have one of the worst values in a solution and replacing them by random values without ever explicitly improving them. Algorithm 1 shows the details of processing steps of EO. In EO, an individual I consists of n components O_i $(1 \leq i \leq n)$. Let λ_i be the fitness value of O_i. First, EO selects O_{worst}, which has the worst fitness value. Second, the state of component O_{worst} is changed randomly. Henceforth, selection and change state of a component are repeated. The component with the worst fitness value has a high possibility that the fitness value of it will become better by changing state. Consequently, the fitness value of the individual also gets better because the fitness value of the component with worst fitness value gets better.

Algorithm 1 EO

1: Generate initial individual I randomly.
2: $I_{best} \leftarrow I$
3: $m \leftarrow 0$
4: **while** $m < max\_of\_generations$ **do**
5: Evaluate fitness value λ_i of each component O_i.
6: Select O_{worst} with the worst fitness value.
7: Change the state of O_{worst} randomly.
8: **if** $\mathbf{F}(I) > \mathbf{F}(I_{best})$ /* The function which returns the fitness value of an individual is denoted as **F**. */ **then**
9: $I_{best} \leftarrow I$
10: **end if**
11: $m \leftarrow m + 1$
12: **end while**

MEO generates two or more neighbor individuals as candidates for the next generation individual. The best neighbor individual in the candidates is selected as the next generation individual. Moreover, MEO uses roulette selection to select a component. Algorithm 2 shows the details of processing steps of MEO. First, MEO selects $O_{selected}$ with roulette selection. The selection rates of roulette selection are reciprocals of fitness values with components. Second, MEO generates new individual I' from I by changing the state of $O_{selected}$. Third, the generated I' is stored into $Candidates$. Finally, MEO selects the best individual from $Candidates$.

PMEOintegrates the population-based approach in MEO. There are two or more individuals in a population. Alternation of generations is repeatedly performed for every individual by using MEO. To improve the search efficiency, individuals copy a sub-structure of an individual that has good sub-structures at each alternation of generations. This operation resembles the crossover operation in genetic programing (GP). However, one side only copies a sub-structure of another side. Copying of good sub-structures leads to a high probability of generation of a good individual.

Algorithm 2 MEO

```
1: Generate initial individual I randomly.
2: I_best ← I
3: m ← 0
4: while m < max_of_generations do
5:   Evaluate fitness value λ_i of each component O_i.
6:   Candidates ← φ
7:   n ← 0
8:   while n < num_of_candidates do
9:     Select O_selected with roulette selection (selection rates are the reciprocal of fitness values
       with components).
10:    Generate new individual I' from I by changing the state of O_selected.
11:    Candidates ← Candidates ∪ I'
12:    n ← n + 1
13:  end while
14:  I ← BEST(Candidates)
15:  if F(I) > F(I_best) then
16:    I_best ← I
17:  end if
18:  m ← m + 1
19: end while
```

5 Island-Model-Based Distributed Modified Extremal Optimization

This section explains the alternation of generations model of Island-Model-based DMEO (IDMEO) and the algorithm of IDMEO for reducing crossovers in a reconciliation graph.

5.1 Alternation of Generations Model

IDMEO is a hybrid of PMEO and DGA using the island model. IDMEO divides the entire population into two or more sub-populations, as many islands. Each island

has a sub-population and the sub-population evolves individually using PMEO. In the island model, from each island, some individuals are selected and transferred to another island. In return, the same number of migrants are received from another island. Each sub-population in a island converges to the separate best solution. Each island evolves individually, the island model can maintain diversity at the end of alternation of generations.

IDMEO repeats the following two steps:

(1) Sub-populations in islands are evolved through one or more generations using PMEO.
(2) Some individuals in islands are migrated to other islands.

5.2 Individual and Component

A reconciliation graph is defined as an individual. A component of an individual is defined as a pair of leaf nodes with the same label name:

$$O_i = \{ol_{1,i}, ol_{2,\delta(i)}\} \quad (\mathscr{L}(ol_{1,i}) = \mathscr{L}(ol_{2,\delta(i)})). \tag{3}$$

Let $ol_{1,i}$ be a leaf node of OL_1 and $ol_{2,\delta(i)}$ be a leaf node of OL_2. The function $\delta(i)$ returns the subscript number of an element of OL_2 whose label name is the same as the label name of $ol_{1,i}$. To change the state of O_i, it is necessary to change the order of child nodes of ancestor nodes of $ol_{1,i}$ or $ol_{2,\delta(i)}$. Here, $\mathscr{AS}(T, lname)$ is a set of ancestor nodes of a leaf node in T that has the label name $lname$. For example, $\mathscr{AS}(I, T_1, O_2)$ returns $\{v_{12}, v_{11}\}$ in Fig. 2.

5.3 Definition of Fitness

The number of crossovers between $ol_{1,i}$ and $ol_{2,\delta(i)}$ is denoted by $\mathscr{C}(M,i)$. The following are the definitions of $\mathscr{C}(M,i)$ and the fitness value λ_i of O_i:

$$\lambda_i = \frac{\mathscr{C}(M) - \mathscr{C}(M,i)}{\mathscr{C}(M)}, \tag{4}$$

$$\mathscr{C}(M,i) = \sum_{l=i+1}^{n} \sum_{m=1}^{\delta(i)-1} \frac{m_{l,m}}{2} + \sum_{l=1}^{i-1} \sum_{m=\delta(i)+1}^{n} \frac{m_{l,m}}{2}. \tag{5}$$

In Fig. 2, there are four components, $O_1 = \{ol_{1,1}, ol_{2,1}\}(= \{v_{14}, v_{24}\})$, $O_2 = \{ol_{1,2}, ol_{2,4}\}(= \{v_{15}, v_{27}\})$, $O_3 = \{ol_{1,3}, ol_{2,2}\}(= \{v_{16}, v_{25}\})$, and $O_4 = \{ol_{1,4}, ol_{2,3}\}$ $(= \{v_{17}, v_{26}\})$, with $\delta(1) = 1$, $\delta(2) = 4$, $\delta(3) = 2$, and $\delta(4) = 3$. The fitness values of the components are $\lambda_1 = 1$, $\lambda_2 = 1/2$, $\lambda_3 = 3/4$, and $\lambda_4 = 3/4$.

5.4 Algorithm

The algorithm of IDMEO for reducing crossovers in a reconciliation graph consists of two steps: (1) Evolution Step and (2) Migration Step (Algorithm 3). First, an initial population divided to p sub-populations (p is the number of sub-populations). Sub-population $SubP_i$ is located in an island $ILND_i$. In the Evolution Step (step 6), the sub-populations in all the islands are made to evolve through m generations by using the function **PMEO**($SubP_i, m$) (m is migration interval). In Migration Step (step 7), some individuals of a sub-population in an island are migrated to another island. Finally, the best individual is selected from all the islands (step 8 and step 9).

Algorithm 3 IDMEO

1: Generate initial population P_{init} randomly.
2: $I_{best} \leftarrow$ **BEST**(P_{init})
3: Divide P_{init} into p sub-populations $SubP_i$.
4: Store sub-populations $SubP_i$ into island $ISLND_i$.
5: **for** $i = 1$ to $max\_generations/m$ **do**
6: **(Evolution Step)** For each $ISLND_i$, sub-population $SubP_i$ should be made to evolve through m generations by using the function **PMEO**($SubP_i, m$).
7: **(Migration Step)** For each $ISLND_i$, migrate some individuals of a sub-population in the island to another island.
8: **if F**(**BEST**($SubP_1 \cap \cdots \cap SubP_p$)) > **F**($I_{best}$) **then**
9: $I_{best} \leftarrow$ **BEST**($SubP_1 \cap \cdots \cap SubP_p$)
10: **end if**
11: **end for**

5.5 Evolution Step

In the Evolution Step, each sub-population is in an island is made to evolve through m generations by using the function **PMEO** (Algorithm 4). First, for each individual, the state of the individuals in P is changed by using MEO. Second, the function **CSS** copies a good sub-structure of an individual to another individual.

In the MEO steps, for each individual, the following steps are executed. Initially, the function evaluates the fitness value λ_i (step 3). Next, the following three steps are repeated while n is less than $num\_of\_candidates$. First, component $O_{selected}$ in I is selected by using the roulette selection (step 7). Second, the function generates an neighbor individual from I with the function **GNI**. The function **GNI** generates a neighbor individual by changing the state of component $O_{selected}$. Third, the neighbor individual is stored in C (step 8). Finally, the best individual in C is selected and I is replaced by it (step 11).

The state of $O_{selected}$ is changed by changing the order of child nodes in an intermediate node, which is an ancestor node of $O_{selected}$. The processing steps of **GNI** are as follows. First, T_1 or T_2 is selected randomly and $\mathscr{AS}(T_1, \mathscr{L}(O_{selected}))$

Algorithm 4 PMEO(P, m)

1: **for** $i = 1$ to m **do**
2: **for all** $I \in P$ **do**
3: Evaluate fitness value λ_i of each component O_i of I.
4: $C \leftarrow \phi$
5: $n \leftarrow 0$
6: **while** $n < num\_of\_candidates$ **do**
7: Select $O_{selected}$ by roulette selection (selection rates are the reciprocal of fitness values with components).
8: $C \leftarrow C \cup \mathbf{GNI}(I, O_{selected})$
9: $n \leftarrow n + 1$
10: **end while**
11: $I \leftarrow \mathbf{BEST}(C)$
12: **end for**
13: **CSS**(P)
14: **end for**

Algorithm 5 CSS(P)

1: Select individual $SI \in P$ by roulette selection (selection rates are the fitness values of components).
2: **for all** $I \in P, I \neq SI$ **do**
3: **for** $i = 1$ to n **do**
4: Calculate the difference $diff_i$ between the fitness value of O_i in SI and the fitness value of O_j in I, where O_i and O_j have the same label name.
5: **end for**
6: Select $O_{selected}$ by roulette selection (selection rates are $diff_i$).
7: $A \leftarrow \mathscr{AS}(T_1, \mathscr{L}(O_{selected}))$ or $\mathscr{AS}(T_2, \mathscr{L}(O_{selected}))$
8: $C \leftarrow \phi$
9: **for all** $a \in A$ **do**
10: Generate a new individual I' from I by changing the order of child nodes in a.
11: $C \leftarrow C \cup I'$
12: **end for**
13: $I \leftarrow \mathbf{BEST}(C)$
14: **end for**

or $\mathscr{AS}(T_2, \mathscr{L}(O_{selected}))$ are stored in the set $Ancestors$. Then, node a is selected randomly from A. Finally, the order of the child nodes in a is changed.

Suppose that the selected component is O_2 in Fig. 2. The function $\mathscr{AS}(T_1, \mathscr{L}(O_2))$ returns $\{v_{12}, v_{11}\}$ and $\mathscr{AS}(T_2, \mathscr{L}(L_2))$ returns $\{v_{22}, v_{21}\}$. If $Ancestors = \{v_{12}, v_{11}\}$ and v_{12} is selected as a, the order of child nodes in v_{12} is changed. In this case, order of node v_{15} and v_{13} are changed. As a result, a new individual I' is obtained by the change state.

Algorithm 5 shows the function **CSS**. At the beginning, an individual SI in P is selected by roulette selection (step 1). Each individual of P copies a sub-structure of SI by the following steps. First, the function calculates the difference $diff_i$ between the fitness value of O_i of SI and the fitness value of O_j of I, where O_i and O_j have the same label name (steps 3, 4, and 5). Second, $O_{selected}$ is selected by roulette selection (step 6). Next, $\mathscr{AS}(T_1, \mathscr{L}(O_{selected}))$ or $\mathscr{AS}(T_2, \mathscr{L}(O_{selected}))$ is stored

in A (step 7). Then, for all $a \in A$, a new individual I' is generated from I by changing the order of child nodes in a, and I' is stored in C (steps 9, 10, 11, and 12). Finally, the function selects the best individual from C (step 13).

5.6 Migration Step

In the Migration Step, some individuals in each island are migrated to another island. The island model requires $number\ of\ sub-populations$, $migration\ rate$, $migration\ interval$, and $migration\ model$. The first three items are user-given parameters. The last item consists of two things: $selection\ method$ and $topology$. The method used for the selection of individuals for migration is referred as $selection\ method$. The structure of the migration of individuals between sub-populations is referred as $topology$. In this study, we use uniform random selection as the $selection\ method$. In the Migration step, some individual are selected from a sub-population in each island according to $migration\ rate$. Moreover, the proposed algorithm uses the random ring migration topology. In this topology, the ring includes all islands, and the order of the islands is determined randomly every Migration step. Each island transfers some individuals to the next inland based on the direction of the ring.

6 Experimental Results

We performed four experiments for evaluating the performance of IDMEO. This section shows the experimental results.

6.1 Setup

In the experiments, the two data sets listed in Table 1 are used. The $Housekeeping$ data set consists of a phylogenetic tree of the housekeeping gene and its taxonomic tree. The $Moss$ data set consists of a phylogenetic tree of the rps4 gene and its taxonomic tree. The number of species in the $Housekeeping$ data set is 40 and that in the $Moss$ data set is 207.

Experiment 1 measured the number of crossovers of the best individual at each generation to compare IDMEO and PMEO. Experiment 2 also measured the number of crossovers of the best individual at each elapsed time to compare IDMEO and PMEO. Experiment 3 measured frequency of the number of crossovers of best individuals in fixed generations. Experiment 4 compares IDMEO with MGG.

Table 1 Data sets

	Taxonomic tree		Phylogenetic tree	
	Number of nodes	Number of leaf nodes	Number of nodes	Number of leaf nodes
Housekeeping	241	40	79	40
Moss	290	207	394	207

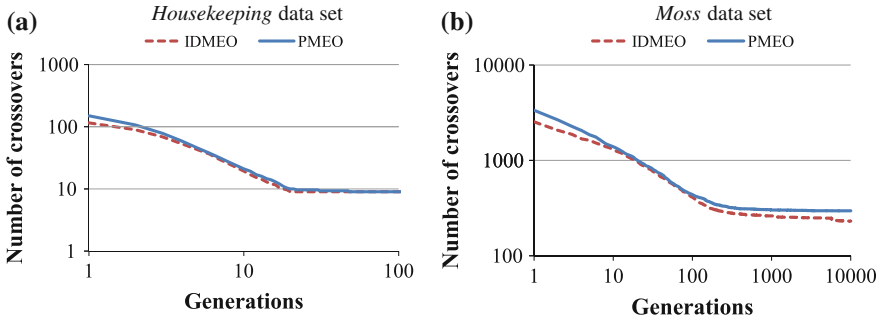


Fig. 3 Experiment 1 (**a** and **b** shows the number of crossovers of the best individual)

In PMEO and IDMEO, the number of individuals in the population was set to 100. The user parameter $num\_of\_candidates$ was set to 100 and m was set to 10,000 in PMEO. In IDMEO, the user parameter $num\_of\_candidates$, $migration\_interval(m)$, $number\ of\ sub-populations(p)$, and $migration\ rate$ were set to be 100, 10, 5 and 0.05, respectively. The number of individuals in a sub-population is 20. The number of crossovers was the average of three trials.

6.2 Experiment 1

In Experiment 1, we measured the number of crossovers of the best individual in each generation. Figure 3a, b show the number of crossovers (vertical axis: the number of crossovers, horizontal axis: generations). Figure 3a, b show that the number of crossovers of IDMEO in each generation was smaller than that in the case of PMEO. IDMEO showed better performance compared with PMEO. The number of crossovers in PMEO is converging into around 300 when we use *Moss* data set. On the other hand, in IDMEO, the number of crossovers is converging into around 250. The diversity of PMEO is small, because the number of sub-populations is one. Therefore, the fitness value of a individual will not be improved in the end of alternation of generations.

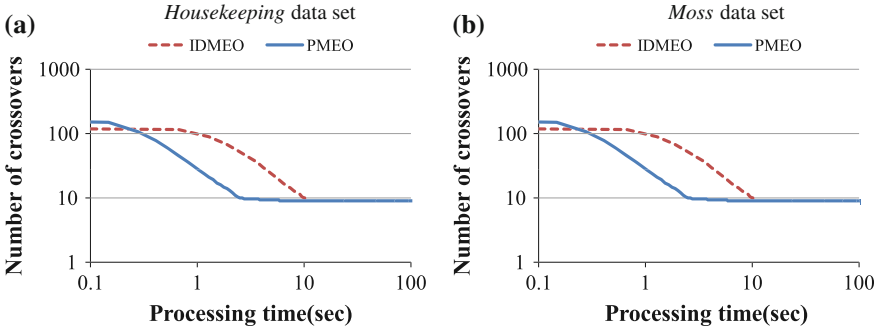


Fig. 4 Experiment 2 (**a** and **b** shows the number of crossovers of the best individual)

6.3 Experiment 2

In Experiment 2, we measured the number of crossovers of the best individual at different time instants. The computation time of IDMEO was longer than that in the case of PMEO because the former included the Migration Step. Therefore, it was necessary to compare the number of crossovers for the same computation time. Figure 4a, b show the number of crossovers at different time instants (vertical axis: the number of crossovers, horizontal axis: processing time). At the end of the processing, IDMEO have fewer crossovers than PMEO. This result indicates IDMEO performs better with fewer crossovers than PMEO.

6.4 Experiment 3

The number of crossovers of the best individual was measured 100 times for the 10,000th alternation generation. Figure 5a, b show frequency of the number of crossovers when *Housekeeping* data set is used. The number of crossovers of the

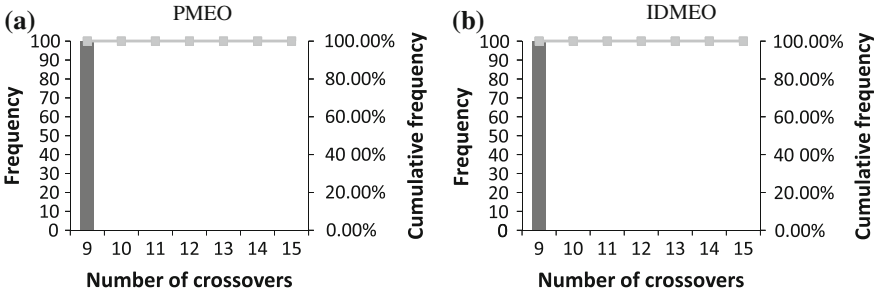


Fig. 5 Experiment 3 (*Housekeeping* data set)

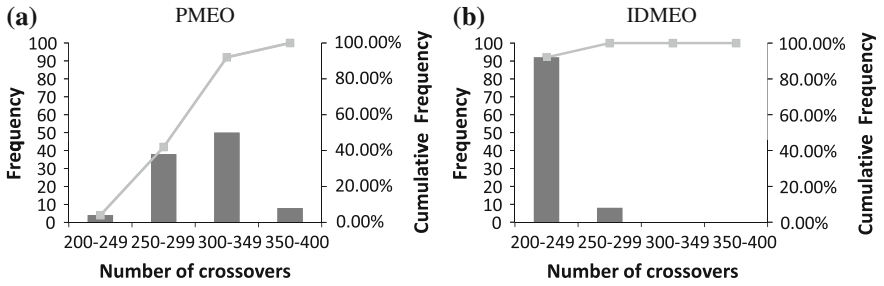


Fig. 6 Experiment 3 (*Moss* data set)

optimal solution of *Housekeeping* data set is 9. Both of them can obtain the best solution by 100 %. Figure 6a, b show the frequency of the number of crossovers for the *Moss* data set. In IDMEO, all the numbers of crossovers of optimal solutions were between 200 and 299. On the other hand, they were distributed between 200 and 400 for PMEO. Above all, although 90 % of optimal solutions were between 200 and 249 in the case of IDMEO, only a few optimal solutions were obtained between 200 and 249 in the case of PMEO.

6.5 Experiment 4

In Experiment 4, we compared IDMEO with MGG, which is the one of the best generation alternation models. In IDMEO, the number of individuals in the population was set to 100. The user parameter *num_of_candidates* was set to 100 and *m* was set to 30,000. In IDMEO, the user parameter *num_of_candidates*,

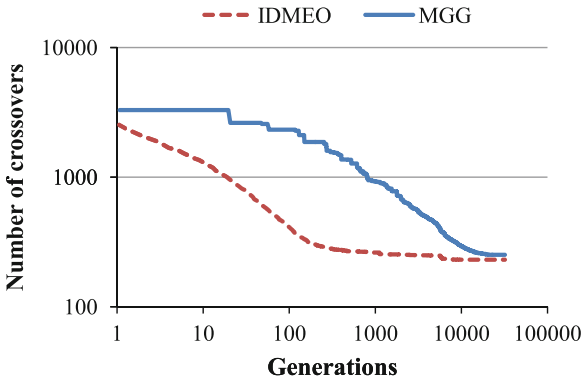


Fig. 7 In Experiment 4, we compared IDMEO with MGG using *Moss* data set (*vertical axis*: the number of crossovers, *horizontal axis*: generations)

$migration\_interval(m)$, $number\ of\ sub-populations(p)$, and $migration\ rate$ were set to be 100, 10, 5 and 0.05, respectively. The number of individuals in a sub-population is 20. In MGG, the number of individuals is 100, the number of children is 100, and the rate of mutation is 5 %.The number of crossovers was the average of three trials. Figure 7 show the number of crossovers (vertical axis: the number of crossovers, horizontal axis: generations). The performance of IDMEO also is better than that of MGG. In particular, the speed of convergence in IDMEO is faster than that in MGG.

7 Conclusion

In this chapter, the island-model-based distributed modified extremal optimization (IDMEO), which is used for reducing crossovers in a reconciliation graph, is proposed. IDMEO is a hybrid of population-based modified extremal optimization (PMEO) and the distributed genetic algorithm using the island model. In the island model, a population is divided into two or more sub-populations called islands and each island evolves individually. Each island can maintain different types of individuals at the end of alternation of generations. Therefore, IDMEO can maintain diversity at the end of alternation of generations. We have evaluated IDMEO by using actual data sets. Experimental results show that the performance of IDMEO is better than that of PMEO. Moreover, experimental results show that IDMEO can maintain diversity and performs better than PMEO. In the future work, we will develop extended IDMEO for making it applicable to other combination optimization problems.

References

1. Goodman M, Czelusniak J, Moore G, Romero-Herrera A, Matsuda G (1979) Fitting the gene lineage into its species lineage, a parsimony strategy illustrated bycladograms constructed from globin sequences. Syst Zoo 28:132–163
2. Page RDM (1998) Genetree: comparing gene and species phylogenies using reconciled trees. Bioinformatics 14(9):819–820
3. Page RDM, Charleston MA (1997) Reconciled trees and incongruent gene and species trees. Discr Math Theor Comput Sci 37:57–70
4. Goldberg DE (1989) Genetic algorithms in search, optimization, and machine learning, Addison-Wesley Professional
5. Kitakami H, Mori Y (2002) Reducing crossovers in reconciliation graphs using the coupling cluster exchange method with a genetic algorithm. Active mining. IOS press 79:163–174
6. Kitakami H, Nishimoto M (2000) Constraint satisfaction for reconciling heterogeneous tree databases. In: Proceedings of DEXA 2000, pp 624–633
7. Boettcher S (2000) Extremal optimization: heuristics via coevolutionary avalanches. Comput Sci Eng 2(6):75–82
8. Boettcher S, Percus A (2000) Nature's way of optimizing. Artif Intell 119(1–2):275–286
9. Boettcher S, Percus AG (1999) Extremal optimization: methods derived from co-evolution. In: Proceedings of GECCO 1999, pp 825–832

10. Tamura K, Mori Y, Kitakami H (2008) Reducing crossovers in reconciliation graphs with extremal optimization (in japanese). Trans Inf Process Soc Japan 49(4(TOM 20)):105–116
11. Hara N, Tamura K, Kitakami H (2010) Modified eo-based evolutionary algorithm for reducing crossovers of reconciliation graph. In: Proceedings of NaBIC 2010, pp 169–176
12. Bak P, Tang C, Wiesenfeld K (1987) Self-organized criticality. an explanation of 1/f noise. Phys Rev Lett 59:381–384
13. Tamura K, Kitakami H, Nakada A (2013) Distributed modified extremal optimization for reducing crossovers in reconciliation graph. In: Tecture Notes in engineering and computer science. Proceedings of the international multiConference of engineers and computer scientists 2013, 13–15 Mar 2013, Hong Kong, pp 1–6
14. Belding TC (1995) The distributed genetic algorithm revisited. In: Proceedings of the 6th international conference on genetic algorithms, pp 114–121
15. Whitley WD, Rana SB, Heckendorn RB (1997) Island model genetic algorithms and linearly separable problems. In: Selected papers from AISB workshop on evolutionary, computing, pp 109–125
16. Boettcher S, Percus AG (2004) Extremal optimization at the phase transition of the 3-coloring problem. Phys Rev E 69:066703
17. Duch J, Arenas A (2005) Community detection in complex networks using extremal optimization. Physi Rev E 72:027104
18. Meshoul S, Batouche M (2002) Robust point correspondence for image registration using optimization with extremal dynamics. In: Proceedings of DAGM-symposium 2002, pp 330–337
19. Zhou T, Bai WJ, Cheng LJ, Wang BH (2005) Continuous extremal optimization for lennard-jones clusters. Phys Rev E 72:016702
20. Chen MR, Li X, Zhang X, Lu YZ (2010) A novel particle swarm optimizer hybridized with extremal optimization. Appl Soft Comput 10(2):367–373
21. Chen MR, Lu YZ, Yang G (2008) Multiobjective optimization using population-based extremal optimization. Neural Comput Appl 17(2):101–109
22. Randall M, Lewis A (2006) An extended extremal optimisation model for parallel architectures. In: Proceedings of E-SCIENCE '06, p 114
23. Hiroshi S, Isao O, Shigenobu K (1997) A new generation alternation model of genetic algorithms and its assessment. J Jap Soc Artif Intell 12(5):734–744

Early-Warning System in Bridge Monitoring Based on Acceleration and Displacement Data Domain

Reni Suryanita and Azlan Adnan

Abstract Bridges should be monitored periodically in order to assess the bridge health at any given time. The sensors send the acceleration and displacement data of a bridge response under earthquakes loading to the system server. This study aims to conduct the early-warning intelligent system based upon the performance of the acceleration and displacement data. The damage detection in the system applied the Neural Networks for prediction of a bridge condition at the real time. The architecture of Neural Networks' model used one input layer, which consists of acceleration and displacement data domain, two hidden layers and an output layer with four neurons consist of safety level, Immediate Occupancy (IO), Life Safety (LS) and Collapse Prevention (CP). The IO, LS and CP are the bridge condition which indicates the extent of bridge health condition ranging from the light damage until high-risk level during and after subject to six earthquakes data. The training activation used the Gradient Descent Back-propagation and activation transfer function used Log Sigmoid function. The early-warning system is applied on 3 spans of box girder bridge model which is monitored in the local and remote server. The result showed that the evaluation of bridge condition using alert-warning in the bridge monitoring system can help the bridge authorities to repair and maintain the bridge in the future.

Keywords Acceleration · Bridge healthy · Displacement · Early-warning system · Gradient descent back-propagation · Log sigmoid function

R. Suryanita (✉)
Civil Engineering Departement, University of Riau, Jl. HR Subrantas KM 12.5, Pekanbaru-Riau, Indonesia
e-mail: renisuryanita@yahoo.co.id

A. Adnan
Engineering Seismological and Earthquake Engineering Research (E-Seer), Faculty of Civil Engineering, Universiti Teknologi Malaysia, 81310 Skudai-Johor Bahru, Malaysia
e-mail: azlanadnan@utm.my

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 275, DOI: 10.1007/978-94-007-7684-5_12,

1 Introduction

Bridge monitoring needs to be carried out regularly in order to maintain and evaluate bridge condition periodically. Currently, the information technology is capable to helping the bridge owner to supervise the bridge condition from remote area through the Internet connection. The installed sensors will sent the acceleration and displacement data to the acquisition tools. The prediction of a bridge damage uses the Neural Network for a bridge structure based on the acceleration and displacement data which has been conducted in the previous study [1].

Many researchers have discussed the application of Neural Network in the bridge engineering field such as [2, 3] and [4]. Other researchers have conducted study about the best performance of Neural Network for prediction of sensors' data for the axial bearing capacity by [5], and the strain of FBG sensors which are based on the time domain by [6]. However, there is limited discussion pertaining to the performance of acceleration and displacement data from sensors using the Neural Networks method, especially for early-warning system on the bridges monitoring.

This study aims to develop and apply the early-warning system on bridge management system based on the acceleration and displacement data domain for damage prediction due to earthquakes load. The system can detect even minor to major damage on the bridge structure. Thereby, the bridge authorities can provide appropriate assessment for maintenance, reparation and improvement the bridge function.

2 Bridge Monitoring Under the Earthquake Load

Regular monitoring of bridge can immensely help the bridge authorities to know and detect the bridge condition early through the sensors data reading. The sensors will sent the acceleration and displacement data to the server through the data acquisition. In structural dynamic, the response of the bridge structure due to earthquakes commonly is derived from (1)

$$[M]\{\ddot{u}\} + [C]\{\dot{u}\} + [K]\{u\} = -[M]\{\ddot{u}_g\} \tag{1}$$

where [M], [C] and [K] are matrix of mass, damping and stiffness respectively. Meanwhile $\ddot{u}$, $\dot{u}$, and u are each the vector of acceleration, velocity, and displacement of a bridge response. Vector $\ddot{u}_g$ is acceleration of earthquake excitation. By using the uncoupling procedure, the modal equation of n^{th} mode can be written as (2).

$$\ddot{u} + 2\xi_n \omega_n \dot{u}_n + \omega^2 u_n = -1/_{\varphi n} \ddot{u}_g \tag{2}$$

Displacement for each mode shown as (3)

$$u(t) = \sum \varphi_n u_n(t) \tag{3}$$

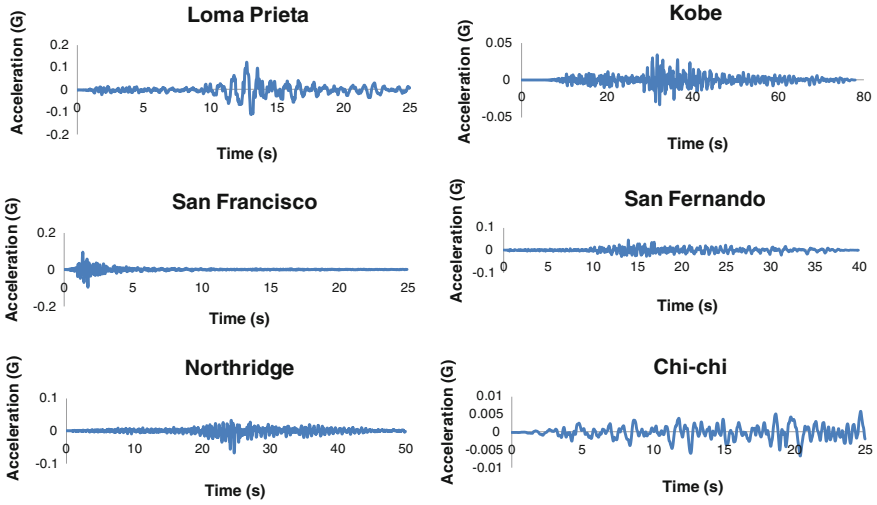


Fig. 1 Time history of six earthquakes data from PEER [9]

where ξ_n, ω, and φ_n are damping ratio, frequency and n number of mode shape respectively. The acceleration is generated by second time derivative of displacement function. The displacement values of a bridge response describe the performance of the bridge under an earthquake loading. In bridge monitoring, both of acceleration and displacement values can be obtained from measurement by sensors which are installed on the bridge. The acceleration and displacement values can be produced from finite-element analysis using a computer program [1].

According to [7], normally, damage of bridge structure is defined as the intentional or unintentional changes in material and geometric properties of the bridge, including changes in boundary or supporting conditions and structural connectivity, which adversely affect the current or future serviceability of the bridge. Damage can occur under large transient loads such as strong motion earthquakes and can also be accumulated incrementally over long periods of time due to factors such as fatigue and corrosion damage.

Time history analysis shall be performed with at least three time-histories data sets of ground motion. Since three time history data sets are used in the analysis of structure, the maximum value of each response parameter shall be used to determine design acceptability [8]. Time history data in this study is adopted from [9] as shown in Fig. 1. The Peak Ground Acceleration (PGA) of the earthquakes are 0.4731G (4.64 m/s^2) for Loma Prieta earthquake, 0.3051G (2.99 m/s^2) for San Francisco earthquake, 0.2363G (2.32 m/s^2) for Northridge earthquake, 0.122G (1.197 m/s^2) for Kobe earthquake, 0.1539G (1.51 m/s^2) for San Fernando earthquake, and 0.062G (0.61 m/s^2) for Chi-chi earthquake.

The acceptance criteria of piers damage are based on structural performance levels in Federal Emergency Management Agency (FEMA) 356. The damage criteria are

divided into 3 categories, Immediate Occupancy (IO), Life Safety (LS) and Collapse Prevention (CP). The IO category describes the structure as still safe to be occupied after an earthquake has occurred. In the LS category, some structural elements and components are severely damaged but the risk of life-threatening injury is low. The CP category describes that the structure is on the verge of partial or total collapse and there is significant risk of injury.

3 Application of Neural Networks in Early-Warning System

Reference [3] has applied the Neural Networks in the study of a bridge under dynamic load, especially general traffic load. The objective of the research is to estimate the bridge displacement which corresponds to the strain of the bridge. The other researchers [10] studied the acceleration-based approach using Neural Networks to predict the displacement of building response under earthquake excitation. The inputs data are the acceleration, velocity and displacement at ground and several stories of building.

Early-warning system in this study adopted the Neural Network Back Propagation (BPNN) algorithm to predict the criteria of damage during and after earthquakes. The best performances of BPNN depend on the selection of suitable initial weight, learning rate, momentum, networks architecture model and activation function. The architecture model for this system has n number of input neurons, two hidden layers with n neurons and an output layer consists of damage levels IO, LS and CP. The input networks consist of time-acceleration domain and time-displacement domain of the bridge seismic response analysis. The numbers of input correspond to the numbers of sensor which are installed on the bridge monitored. Meanwhile the output layer is the level of a bridge health condition due to an earthquake, which is resulted by finite-element analysis software. The architecture model of Neural Networks for this study is illustrated in Fig. 2.

The study used Gradient Descent Back-propagation as training function to minimize the sum squared error (E) between the output value of Neural Network and the given target values. The total error is defined as (4).

$$E = \frac{1}{2} \sum_{j \in J} (t_j - a_j)^2 \tag{4}$$

where t_j defines target value, a_j denotes activation value of output layer, and J is set of training examples. The steps are repeated until the mean-squared error (MSE) of the output is sufficiently small [11].

The final output is generated by a non linear filter Φ caller activation function or transfer function. The transfer function for this model used Log Sigmoid function, which has a range of [0, 1] to obtain the output. This function is differentiable function and suitable to be used in BPNN multilayer as shown in (5).

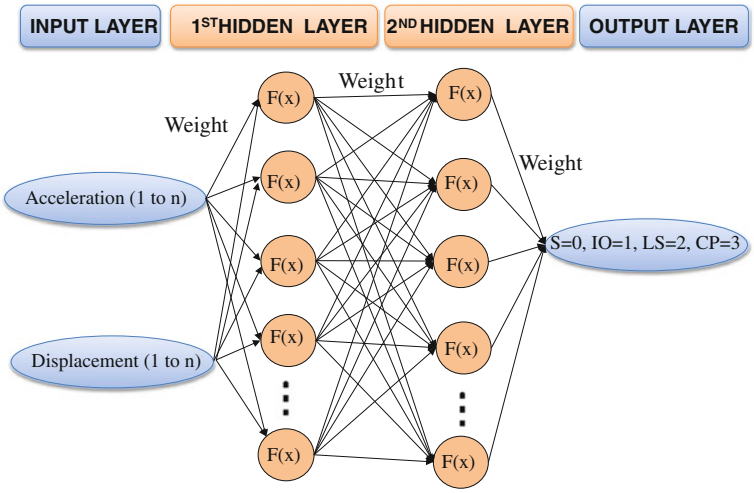


Fig. 2 The architecture model of neural networks with 2 hidden layers in the early-warning system

$$a_j = \frac{1}{(1 + e^{-a_{net,j}})} \tag{5}$$

where $a_{net,j} = [\sum_{i=1}^{l} w_{ij} a_i] + \theta_j$.

Each i represents one of the units of layer l connected to unit j and θ_j which represents the bias.

The weight, w_{ij} of networks has been adjusted to reduce the overall error. The updated weight on the link connected to the i^{th} and j^{th} neuron of two adjacent layers is defined as,

$$\Delta W_{ij} = \eta(\partial E/\partial W_{ij}) \tag{6}$$

where, η is the learning rate parameter with range 0–1 and $\partial E/\partial W_{ij}$ is the error gradient with reference to the weight.

The input data has been normalized by a linear normalization equation as follows:

$$z_i' = (z_i - z_{min})/(z_{max} - z_{min}) \tag{7}$$

where z_i' is the normalized input values, z_i the original data, z_{max} and z_{min} are the maximum and minimum values.

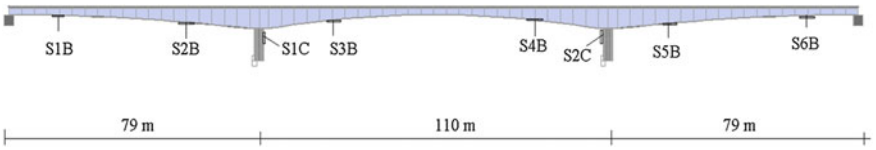


Fig. 3 Sensors location on the 3 spans of box girder bridge model

4 A Case Study

The previous study [1] has conducted the performance of displacement and acceleration data domain for a 3 spans box girder concrete bridge using 2 sensors on the piers. In this study the 8 sensors were assumed to be installed along the bridge as shown in Fig. 3. The sensors measured the acceleration and displacement values of the bridge response. The lengths of the bridge spans are 79, 110, and 79 m respectively.

The bridge model in Fig. 3 has been analyzed using the finite-element analysis software. The non linear time history analysis has been applied in the model so that the behavior and condition of the model due to earthquake can be known as a detail at the given time. The bridge model in this study has been simulated to receive six excitations of earthquake as shown in Fig. 1. Thereby, responses of bridge structure due to some earthquakes have been applied as input in the training process.

The damage of structural elements from finite-element analysis are described in Fig. 4. The criteria of bridge damage are based on standard of Federal Emergency Management Agency (FEMA)356 [8]. The operation level is described as B, which states transition from safe level to IO level. The level before damage is described as S (safe level). Figure 4 illustrates the point of high risk damage due to New Zealand earthquake occurred at bottom of piers (CP level).

Figures 5 and 6 show the response of the bridge model due to New Zealand earthquake. The acceleration and displacement responses of the bridge are measured during the 8 s at the point where S1C and S2C sensors are located. The damage level occurred after 4.70 s. This level consists of IO level (1st index), LS level (2nd index) and CP level (3rd index) at 4.70, 6.20, and 7.10 s respectively. The time before 4.70 s is categorized a safe level (zero index). The maximum acceleration values of bridge response are 1.57 m/s^2 at S1C sensor and 3.21 m/s^2 at S2C sensor as shown in Fig. 5.

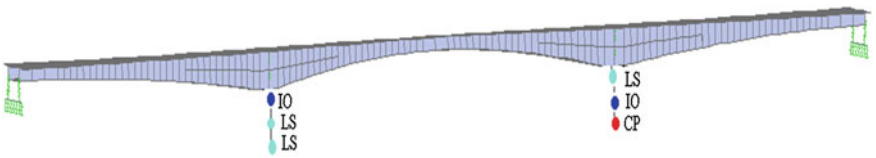


Fig. 4 Damage location of bridge model due to the New Zealand excitation earthquake

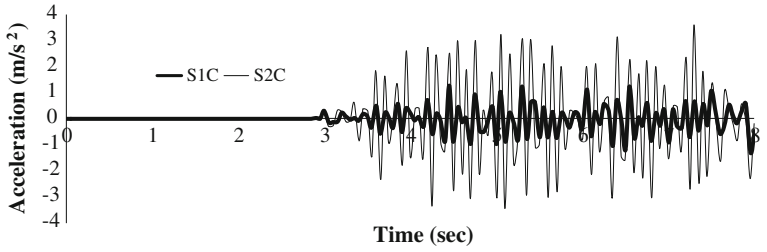


Fig. 5 The acceleration response of bridge model due to the excitation of New Zealand earthquake

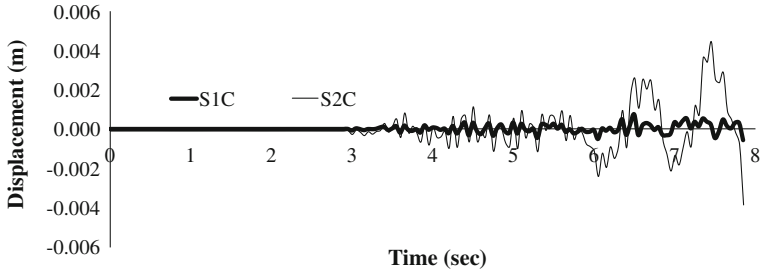


Fig. 6 The displacement response of bridge model due to the excitation of New Zealand earthquake

The maximum displacement value at S1C sensor is 0.0014 m whereas at S2C is 0.00457 m as shown in Fig. 6.

The study used two hidden layers to find the best result for prediction of bridge condition. The architecture model for 2 hidden layers has 17 neurons for input layer, 17 neurons for 1st hidden and 17 neurons for 2nd hidden layer and 4 neurons for output layer. The topology of neurons can be written as 17-17-17-4. The 17 neurons of input layer consist of 1 neuron for time domain, 8 neurons each for acceleration and displacement data domain. At the same time 4 neurons of output layer consist of the bridge damage levels which are categorized into 4 indexes. The indexes are 0 (zero) for safety level (S), 1 (one) for IO level, 2 (two) for LS level and 3 (three) for CP level.

The example of the input data Neural Networks from two sensors due to New Zealand earthquake is shown in Table 1. This data comes from Figs. 5 and 6. The ACC1 and ACC2 denote acceleration data domain for S1C and S2C sensors, whereas DISPL1 and DISPL2 denote displacement data domain for S1C and S2C sensor. The total data in Table 1 is 170 consist of the safety level has 159 data for time of occurrence 7.90 s, the IO level has 5 data for time of occurrence 2.0 s, the LS level has 2 data for time occurrence 0.05 s, and the CP level has 3 data for time of occurrence 0.15 s. The total numbers of input and output data are 5,891, which are obtained from six earthquakes excitation.

Table 1 The example of input data S1C and S2C sensors due to New Zealand earthquake

Data	Input					Output
	Time	S1C		S2C		
		ACC1	DISPL1	ACC2	DISPL2	
1	0	0	2.11E−02	0.00E+00	−9.77E−02	S = 0
2	0.05	1.23E−04	2.13E−02	−4.95E−04	−9.71E−02	S = 0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
160	7.95	−1.29514	−1.72	−2.29747	−10.31	IO = 1
161	8	−4.73E−01	−3.39E+00	−1.75E−01	−1.70E+01	IO = 1
162	8.05	1.77E+00	−4.75E+00	3.01E+00	−2.35E+01	IO = 1
163	8.1	3.81E−01	−2.98E+00	−3.32E+01	−2.65E+01	IO = 1
164	8.15	−8.37E−01	−2.12E+00	2.67E+01	−4.57E+01	IO = 1
165	8.2	4.46E−01	−2.78E+00	−1.99E+01	−3.22E+01	LS = 2
166	8.25	−2.82E−01	−2.66E+00	1.43E+01	−4.05E+01	LS = 2
167	8.3	−1.70E−01	−2.67E+00	−9.72E+00	−3.31E+01	CP = 3
168	8.35	7.63E−01	−3.34E+00	6.85E+00	−3.68E+01	CP = 3
169	8.4	−4.05E+00	−2.62E+00	−2.80E+00	−3.14E+01	CP = 3
170	8.45	−6.48E+05	2.55E+01	−8.78E+01	−2.73E+01	CP = 3

The Neural Networks in the study used 70 % data for training, 15 % data for testing and 15 % data for validation process. The parameters to indicate the end of training are the mean square error (MSE), maximum of epochs and learning rate (Lr). The MSE with 0.001 performance goal has been used in the networks, whereas the maximum number of epoch used is 50,000, and learning rate used is 0.1. The networks have been examined by the computer with specification Intel Core i5-2410M, the power of processor is 2.30 GHz with turbo boost up to 2.90 GHz and memory 4 GB.

The MSE of Neural Network models based on acceleration data domain with two hidden layers is as in Fig. 7. The figure illustrates that all MSE models have the same trend after 20,000 iterations. The MSE values of testing process are higher than other MSE values. However, overall the error on all processes decreases along the iterations. The error due the testing process is not used during the training process, but it is used to compare with the different models.

Similar to the MSE of acceleration data domain, the MSE of displacement is also shown the same trend after 20,000 iterations. Figure 8 shows the MSE of the model based on displacement data domain for two hidden layers. The MSE of validation has the fluctuation along the iterations before 15,000 epochs. The fluctuation describes the networks have not been convergent yet if the runtime is less than 20,000 epochs.

The result indicates the architectures model for 2 hidden layers with more than 20,000 epochs can be accepted and used for predict the damage level based on the acceleration and displacement data domain.

The best performance of MSE value is the smallest of MSE, because it means the smallest of the error occurred in the calculation. However the best regression value is the highest value which is closes to 1. The regression with value close to 1 defines the

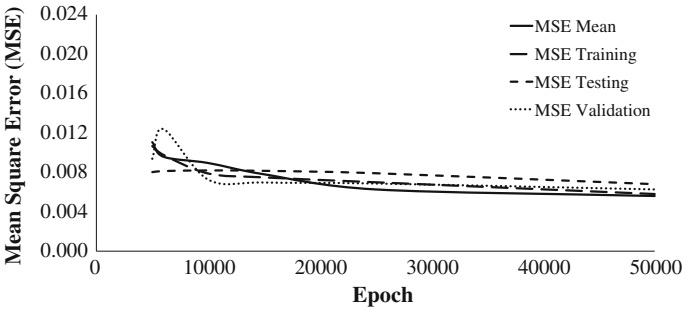


Fig. 7 The means square error of neural network model for 2 hidden layer of acceleration domain

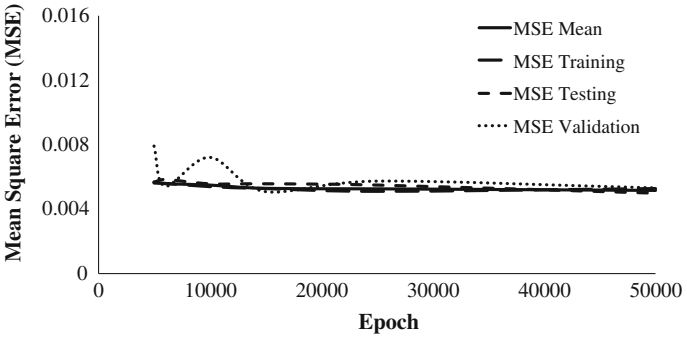


Fig. 8 The means square error of neural network model for 2 hidden layer of displacement domain

prediction value almost 100 % close to the actual one. The best performance of CPU time is defined as the shortest time to process the calculation in central processing unit (CPU). The CPU time is measured in seconds. The CPU time is dependent on the CPU's computational power and specification of the computer.

Table 2 shows the comparison of the acceleration and displacement data domain. The average of regressions (R-mean) for acceleration data domain is above 0.85 % whereas the Mean Square Error (MSE) is lower than 1 %. At the same time, the best of MSE and R-mean value of acceleration data domain are 0.0056 and 0.88689 at 50,000 epochs, whereas the best of MSE and R-mean value for displacement data domain are 0.0512 and 0.83001 at 50,000 epochs. The results shows acceleration data domain can produce higher R-mean values and smaller MSE rather than the displacement data domain for the bridge model which has 8 installed sensors.

The result shows that the Neural Networks model with 2 hidden layers is suitable for the prediction of damage level in bridges seismic monitoring system. Therefore the method can be applied to warn the bridge owner to evaluate the bridge condition early.

Table 2 Comparison of acceleration and displacement domain

Epochs	Acceleration			Displacement		
	MSE mean	R Mean	CPU time	MSE mean	R Mean	CPU time
5,000	0.0111	0.8532	357.4525	0.0565	0.81091	378.0502
6,000	0.00961	0.8543	387.2341	0.0555	0.81619	421.5607
10,000	0.00897	0.8566	405.1233	0.0546	0.81876	561.2712
15,000	0.00782	0.8589	587.8143	0.0525	0.8278	871.2349
25,000	0.00627	0.8678	778.4599	0.0522	0.82921	983.8231
50,000	0.0056	0.88689	1,298.3425	0.0512	0.83001	1,330.8237

5 Early-Warning System

The bridge monitoring system in the study has several components to support the main function which includes data acquisition module, intelligent engine module, alert system module, and monitoring module. The modules use the VB.NET which is provided in two versions involving local and remote monitoring from server. The local monitoring is located in the bridges whereas the remote monitoring accesses the data from any places via internet. HTTP server is utilized to provide the remote data that has a script converting acceleration data to HTML format. The testing using dummy data indicates that the developed intelligent monitoring could perform its

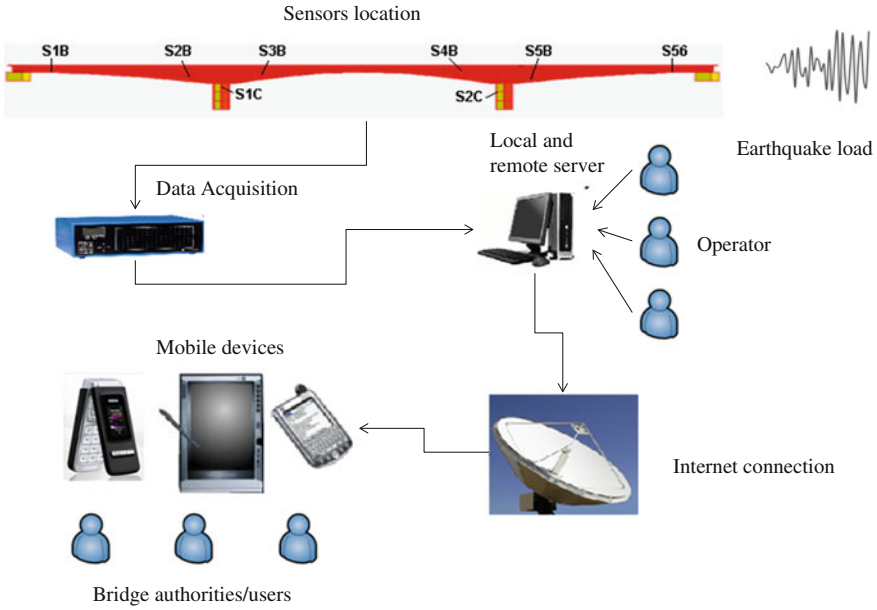


Fig. 9 The bridge monitoring system is developed in this study

functions including monitoring, predicting, and alerting. The monitoring system in the study is illustrated in Fig. 9.

The three steps of monitoring system are adopted from previous research [12]. First step is designing Neural Networks architecture including simulating the bridge damage level due to the earthquakes, training and testing neural, and obtaining the initial weights. The second step is designing and developing the intelligent monitoring software using VB.NET, namely SEER Monalisa. The last step is designing and developing the alert system. The early warning system is embedded in the SEER Monalisa software which is developed by Structural Earthquake Engineering Research at Universiti Teknologi Malaysia [13].

The software scopes are the data inputs from sensors such as accelerometers and strain gauges, feeding forward the inputs Neural Networks, predicting the output as bridges damage level and providing the alert warning as shown in Fig. 10.

The alerts are divided into four format namely the alert bars which are shown in different color (S: Green, IO: Yellow, LS: Orange, and CP: Red), alert sound/alarm, and alert-mail sent to the user. The software has a main function prediction of damage level when an earth quake occurred. After the prediction output indicated either IO, LS, or CP, the alert system will then notify the user that the condition of the bridge is not secure.

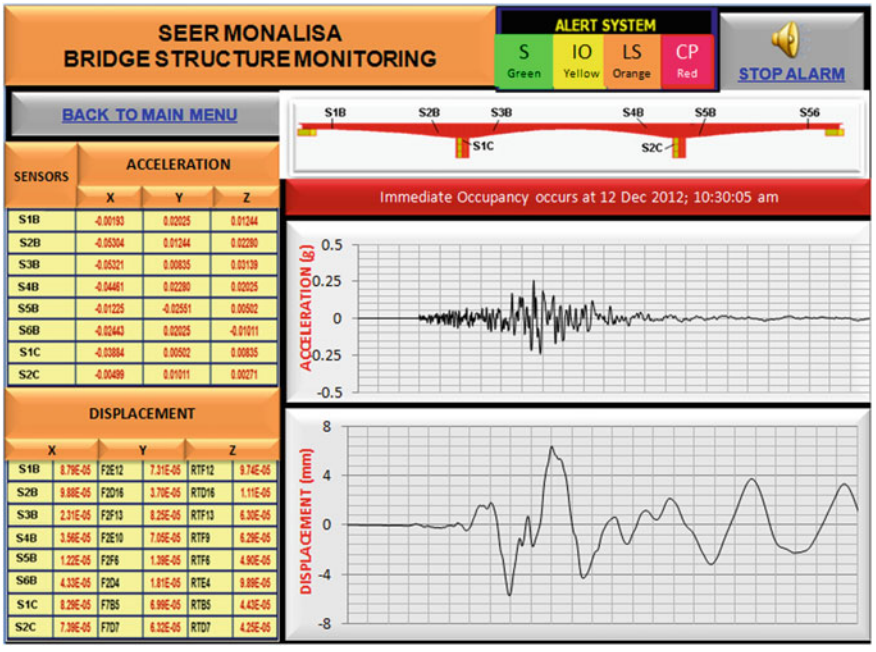


Fig. 10 Early warning system in the SEER-Monalisa bridge monitoring software (colour figure online)

6 Conclusion

The bridge health system used several sensors to detect the behavior of a bridge such as bridge deformation and damage. The sensors connected to the data logger and subsequently sent the information data such as displacement and acceleration to the server. The data is used as input by Neural Networks within the server system. The architecture of neural network method in this study is comprised of two hidden layers.

The Neural Network model which is based on acceleration and displacement data domain with two hidden layers, illustrated that all MSE models have the same trend after 20,000 iterations. The comparison of acceleration and displacement data domain for two hidden layers' model has been concluded based on MSE mean value, regression mean value and CPU time of the network model. Both comparisons showed that the MSE mean value decreased as the epoch increased.

Most bridge monitoring systems use the accelerometer sensors to measure the acceleration of bridge response, because the accelerometer sensor is simpler to install in the field. Furthermore, the acceleration from accelerometer sensors can be modified directly to conduct the displacement value before being entered into the Neural Networks system server. Consequently, the monitoring system is recommended to be used in the Neural Networks with two hidden layers based on displacement domain.

The implementation of an early-warning system in the intelligent Neural Network method for the bridge seismic monitoring system can help the bridge authorities to predict the stability and health condition of the bridge structure at any given time. The software is needed in order to disseminate the bridge health information to the public because it has a main function prediction of damage level when an earthquake occurred.

References

1. Suryanita R, Adnan A (2013) Application of neural networks in bridge health prediction based on acceleration and displacement data domain. In: Lecture notes in engineering and computer science: proceedings of the international multiconference of engineers and computer scientists 2013, vol 2202(1). Hong Kong, pp 42–47, 13–15 March 2013
2. Kerh T, Huang C, Gunaratnam D (2011) Neural network approach for analyzing seismic data to identify potentially hazardous bridges. Math Prob Eng 1–15
3. Ok S, Son W, Lim YM (2012) A study of the use of artificial neural networks to estimate dynamic displacements due to dynamic loads in bridges. J Phys: Conf Ser 382(1)
4. Cheng J, Li QS (2012) Artificial neural network-based response surface methods for reliability analysis of pre-stressed concrete bridges. Struct Infrastruct Eng 8(2):171–184
5. Maizir H, Kassim KA (2013) Neural network application in prediction of axial bearing capacity of driven piles. In: Proceedings of the international multiconference of engineers and computer scientists 2013, vol 2202(1). Hong Kong, pp 51–55, 13–15 March 2013
6. Kahandawa GC et al (2013) Use of fixed wavelength fibre-bragg grating (FBG) filters to capture time domain data from the distorted spectrum of an embedded FBG sensor to estimate strain with an artificial neural network. Sens Actuators, A 194:1–7

7. Wong K-Y (2007) Design of a structural health monitoring system for long-span bridges. Struct Infrastruct Eng 3(2):169–185
8. FEMA356 (2000) Prestandard and commentary for the seismic rehabilitation of buildings. Federal Emergency Management Agency
9. PEER (2012) Pacific earthquake engineering research ground motion database. 15 March 2011. Available from: http://www.peer.berkeley.edu/
10. Qian Y, Mita A (2008) Acceleration-based damage indicators for building structures using neural network emulators. Struct Control Health Monit 15(6):901–920
11. Jones MT (2005) In: Pallai D (ed) AI application programming. Boston, Charles River Media
12. Mardiyono M, Suryanita R, Adnan A (2012) Intelligent monitoring system on prediction of building damage index using artificial neural network. TELKOMNIKA Indonesian J Electr Eng 10(1):155–164
13. SEER-UTM (2007) Intelligent bridge health monitoring system software

An Intelligent Train Marshaling Plan of Freight Cars Considering I/O Arrangements

Yoichi Hirashima

Abstract This paper proposes a new marshaling method for assembling an outgoing train considering the layout of incoming freight cars as well as the outgoing ones. In the addressed problem, freight cars are initially located in a incomming train by the random layout, and they are rearranged and lined into a main track in a certain desired order for an outgoing train. In the proposed method, each set of freight cars that have the same destination make a group, and the desired group layout constitutes the best outgoing train. Then, the incoming freight cars are classified into several "sub-tracks" searching better assignment in order to reduce the total processing time. Classifications and marshaling plans based on the processing time are obtained by a reinforcement learning system. In order to evaluate the processing time, the total transfer distance of a locomotive and the total movement counts of freight cars are simultaneously considered. The total processing time is reduced by obtaining simultaneously the classification of incoming cars, the order of movements of freight cars, the position for each removed car, the layout of groups in a train, the arrangement of cars in a group and the number of cars to be moved.

Keywords Container transfer problem · Freight train · Marshaling · Q-learning · Reinforcement learning · Scheduling

1 Introduction

At railway stations, train marshaling operation plays an important role to connect different modes of transportation. Since freight trains can transport goods only between railway stations, modal shifts are required for area that has no railway. Goods on logistics are carried in containers, each of which is loaded on a freight car. A freight

Y. Hirashima (✉)
Osaka Institute of Technology, 1-79-1 Kitayama, Hirakata, Osaka 573-0196, Japan
e-mail: hirash-y@is.oit.ac.jp

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 275, DOI: 10.1007/978-94-007-7684-5_13,

train is consists of several freight cars, and each car has its own destination. Thus, the train driven by a locomotive travels several destinations decoupling corresponding freight cars at each freight station. In intermodal transports including rail, containers carried into the station are located at the freight yard in the arriving order. The initial layout of freight cars in the yard is determined by considering both arrangement of incoming train and the arriving order of the containers. For efficient shift in assembling outgoing train, freight cars must be rearranged before coupling to the freight train. In general, the rearrangement process is conducted in a freight yard that consists of a main-track and several sub-tracks. Freight cars are initially placed on sub-tracks, rearranged, and lined into the main track. This series of operation is called marshaling, and several methods to solve the marshaling problem have been proposed [1, 8]. Also, similar problems are treated by mathematical programming and genetic algorithm [2, 4, 9, 11], and some analyses are conducted for computational complexities [2, 3]. However, models assumed in these methods are different from the one in the addressed problem, and thus, these methods do not consider the processing time for each transfer movement of a locomotive. Recently, 3 reinforcement learning methods have been proposed in order to solve marshaling problems that have randomly defined initial layout in the fixed number of sub tracks [5–7]. The first one is derived based on the number of movements of freight cars, the second one is based on the transfer distance of locomotive, and the last one evaluates the processing time of marshaling.

In this paper, an unified method to evaluate "time", "distance" or "movement counts" is proposed for generating marshaling plan of freight cars in a train considrering initial classification in sub-tracks. In the proposed method that evaluates the processing time, the incoming freight cars are classified into several sub-tracks searching better assignment in order to reduce the total processing time. Then, classifications and marshaling plans based on the processing time are obtained by a reinforcement learning system. A movement of a freight car consists of 4 elements: (1). moving a locomotive to the car to be transferred, (2). coupling cars with the locomotive, (3). transferring cars to their new position by the locomotive, and (4). decoupling the cars from the locomotive. The processing times for elements 1 and 3 are determined by the transfer distance of the locomotive, the weight of the train, and the performance of the locomotive. The total processing time for elements 1 and 3 is determined by the number of movements of freight cars. Thus, the transfer distance of the locomotive and the number of movements of freight cars are simultaneously considered, and used to evaluate and minimize the processing time of marshaling for obtaining the desired layout of freight cars for an outgoing train. The total processing time of marshaling is considered by using a weighted cost of a transfer distance of the locomotive and the number of movements of freight cars. Then, the order of movements of freight cars, the position for each removed car, the arrangement of cars in a train and the number of cars to be moved are simultaneously optimized to achieve minimization of the total processing time. The *original* desired arrangement of freight cars in the main track is derived based on the destination of freight cars. In the proposed method, by grouping freight cars that have the same destination, several desirable positions for each freight car in a group are generated from the original

one, and the optimal group-layout that can achieve the smallest processing time of marshaling is obtained by autonomous learning. Simultaneously, the desirable classification of incoming cars as well as, the optimal sequence of car-movements, the number of freight cars that can achieve the desired layout of outgoing train is obtained by autonomous learning. Also, the feature is considered in the learning algorithm, so that, at each arrangement on sub-track, the corresponding evaluation value reflects the best movement of locomotive to achieve the desirable layout on the main track. The learning algorithm is derived based on the Q-Learning [12], which is known as one of the well established realization algorithm of the reinforcement learning.

In the learning algorithm, the state is defined by using a layout of freight cars, the car to be moved, the number of cars to be moved, and the destination of the removed car. An evaluation value called Q-value is assigned to each state, and the evaluation value is calculated by several update rules based on the Q-Learning algorithm. In the learning process, a Q-value in a certain update rule is referred from another update rule, in accordance with the state transition. Then, the Q-value is discounted by discount factor calculated according to the target of evaluation: the movement counts of freight cars, the transfer distance of the locomotive, or the processing time of marshaling. Consequently, Q-values at each state represent the total evaluatoin of marshaling to achieve the best layout from the state. In order to show effectiveness of the proposed method, computer simulations are conducted for several methods.

2 Problem Description

A freight yard is assumed to have 1 main track and m sub-tracks. Define k as the number of freight cars carried in and placed on the sub-tracks. Then, they are moved to the main track by the desirable order based on their destination. In the yard, a locomotive moves freight cars from sub-track to sub-track or from sub-track to main track. The movement of freight cars from sub-track to sub-track is called removal, and the car-movement from sub-track to main track is called rearrangement. For simplicity, the maximum number of freight cars that each sub-track can have is assumed to be n, the ith car is recognized by an unique symbol c_i ($i = 1, \ldots, k$). Figure 1 shows the outline of freight yard in the case $k = 30, m = n = 6$.

In the figure, track $\mathrm{T_m}$ denotes the main track, and other tracks [1], [2], [3], [4], [5], [6] are sub-tracks. The main track is linked with sub-tracks by a joint track, which is used for moving cars between sub-tracks, or for moving them from a sub-track to the main track. Figure 1a depicts an example of classification and Fig. 1b is an example of rearrangement. In Fig. 1a, after cars c_1 through c_{12} and c_{20} through c_{25} are classified into sub-tracks [1] [2] [3], c_{19} is placed on the sub-track [6]. Then, c_{26} through c_{30} carried by trucks are placed on sub-track [6] by the arriving order. In Fig. 1b, freight cars are moved from sub-tracks, and lined in the main track by the descending order, that is, rearrangement starts with c_{30} and finishes with c_1. When the locomotive L moves a certain car, other cars locating between the locomotive and the car to be moved must be removed to other sub-tracks. This operation is called

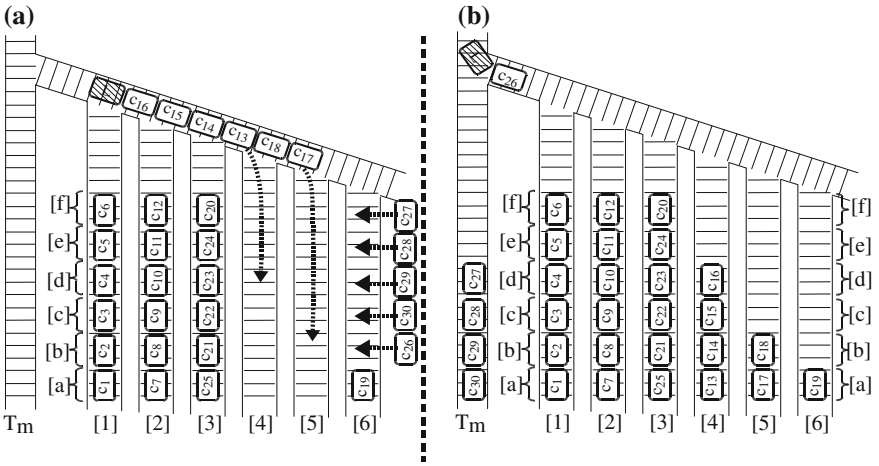


Fig. 1 Freight yard, **a** Classification stage, **b** Rearrangement stage

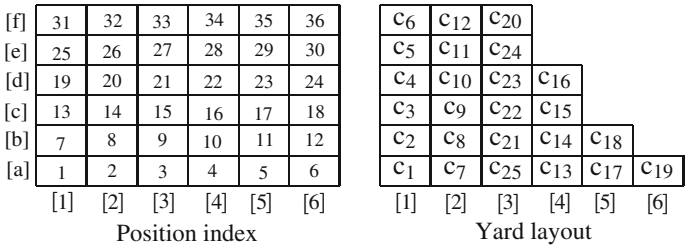


Fig. 2 Example of position index and yard state

removal. Then, if $k \leq n \cdot m - (n-1)$ is satisfied for keeping adequate space to conduct removal process, every car can be rearranged to the main track.

In each sub-track, positions of cars are defined by n rows. Every position has unique position number represented by $m \cdot n$ integers, and the position number for cars at the main track is 0. Figure 2 shows an example of position index for $k = 30, m = n = 6$ and the layout of cars for Fig. 1b.

In Fig. 2, the position "[a][1]" that is located at row "[a]" in the sub-track "[1]" has the position number 1, and the position "[f][6]" has the position number 36. For unified representation of layout of cars in sub-tracks, the first car is placed at the row "[a]" in every track, and a newly placed car is coupled with the adjacent freight car. In the figure, in order to rearrange c_{25}, cars $c_{24}, c_{23}, c_{22}, c_{21}$ and c_{20} have to be removed to other sub-tracks. Then, since $k \leq n \cdot m - (n-1)$ is satisfied, c_{25} can be moved even when all the other cars are placed in sub-tracks.

In the freight yard, define $x_i (1 \leq x_i \leq n \cdot m, i = 1, \cdots, k)$ as the position number of the car c_i, and $s = [x_1, \cdots, x_k]$ as the state vector of the sub-tracks. For example, in Fig. 2, the state is represented by $s = [1, 7, 13, 19, 25, 31,$

2, 8, 14, 20, 26, 32, 4, 10, 16, 22, 5, 11, 6, 33, 9, 15, 21, 27, 3, 0, 0, 0, 0, 0]. A trial of the rearrange process starts with the initial layout, rearranging freight cars according to the desirable layout in the main track, and finishs when all the cars are rearranged to the main track.

3 Desired Layout in the Main Track

In the main track, freight cars that have the same destination are placed at the neighboring positions. In this case, removal operations of these cars are not required at the destination regardless of layouts of these cars. In order to consider this feature in the desired layout in the main track, a group is organized by cars that have the same destination, and these cars can be placed at any positions in the group. Then, each destination makes a corresponding group, and the order of groups lined in the main track is predetermined by destinations. This feature yields several desirable layouts in the main track.

Figure 3 depicts examples of desirable layouts of cars for a group and the desired layout of groups in the main track. In the figure, freight cars $c_1, \ldots, c_6$ to the destination$_1$ make group$_1$, $c_7, \ldots, c_{18}$ to the destination$_2$ make group$_2$, and $c_{19} \ldots, c_{25}$ to the destination$_3$ make group$_3$. Groups$_{1,2,3}$ are lined by ascending order in the main track, which make a desirable layout. Also, in the figure, examples of layout in group$_1$ are in the dashed square.

The layout of groups lined by the reverse order do not yield additional removal actions at the destination of each group. Thus, in the proposed method, the layout lined groups by the reverse order and the layout lined by ascending order from both ends of the train are regarded as desired layouts. Figure 4 depicts examples of material handling operation for extended layout of groups at the destination of group$_1$. In the figure, step 1 shows the layout of the incoming train. In case (a), cars in group$_1$ are separated at the main track, and moved to a sub-track by the locamotive L at step 2. In cases (b), (c), cars in group$_1$ are carried in a sub-track, and group$_1$ is separated at the sub-track. In the cases, group$_1$ can be located without any removal actions for cars in each group. Thus, these layouts of groups are regarded as candidate for desired one in the learning process of the proposed method.

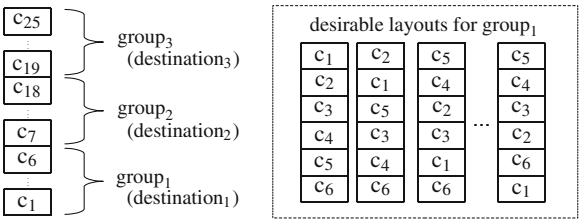


Fig. 3 Example of groups

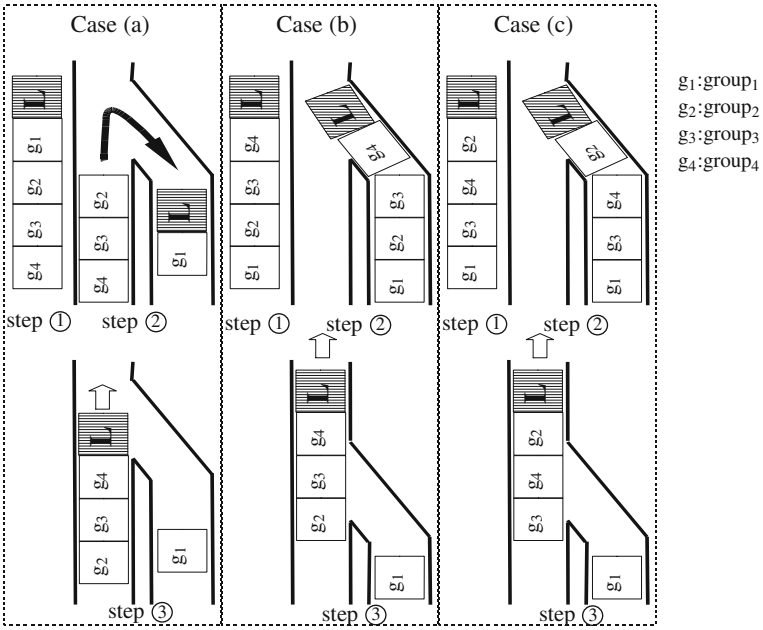


Fig. 4 Group layouts

4 Direct Rearrangement

When there exists a rearranging car that has no car to be removed in front of it, its rearrangement precedes any removals. In the case that several cars can be rearranged without a removal, rearrangements are repeated until all the candidates for rearrangement requires at least one removal. If several candidates for rearrangement require no removal, the order of selection is random, because any orders satisfy the desirable layout of groups in the main track. In this case, the arrangement of cars in sub-tracks obtained after rearrangements is unique, so that the movement count of cars has no correlation with rearrangement order of cars that require no removal. This operation is called direct rearrangement. When a car in a certain sub-track can be rearrange directly to the main track and when several cars located adjacent positions in the same sub-track satisfy the layout of group in the main track, they are coupled and applied direct rearrangement.

Figure 5 shows 2 examples of arrangement in sub-tracks existing candidates for rearranging cars that require no removal. At the top of figure, from the left side, a desired layout of cars and groups, the initial layout of cars in sub-tracks after the assignment process of incoming cars, and the position index in sub-tracks are depicted for $m = n = 4, k = 9$. c_1, c_2, c_3, c_4 are in $group_1$, c_5, c_6, c_7, c_8 are in $group_2$, and $group_1$ must be rearranged first to the main track. In each group, any layouts of cars can be acceptable. In both cases, c_2 in step 1 and c_3 in step 3 are applied the direct

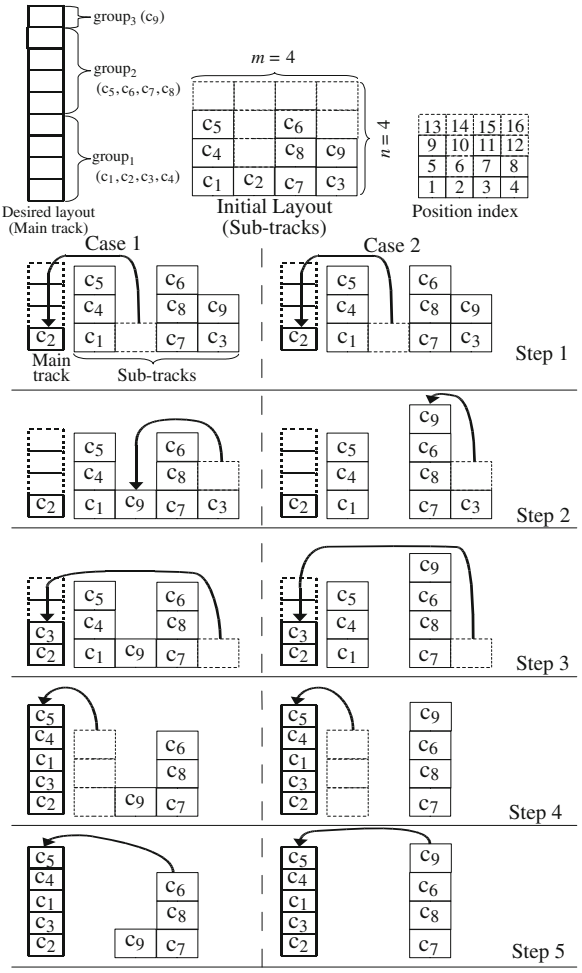


Fig. 5 Direct rearrangements

rearrangement. Also, in step 4, 3 cars c_1, c_4, c_5 located adjacent positions are coupled with each other and moved to the main track by a direct rearrangement operation. In addition, at step 5 in case 2, cars in $group_2$ and $group_3$ are moved by a direct rearrangement, since the positions of c_7, c_8, c_6, c_9 are satisfied the desired layout of groups in the main track. Whereas, at step 5, case 1 includes 2 direct rearrangements separately for $group_2$ and $group_3$.

5 Marshaling Process

A marshaling process consists of following 7 operations:

(I) selection of a layout of groups in the main track,
(II) classification of the incoming freight cars into sub-tracks,
(III) direct rearrangement,
(IV) selection of a freight car to be rearranged into the main track,
(V) selection of a removal destinations of the cars in front of the car selected in (IV),
(VI) selection of the number of cars to be moved,
(VII) removal of the cars to the selected sub-track.

Operations (I), (II) are conducted once in each marshaling process, (III)–(VII) are repeated until one of desirable layouts is achieved in the main track. A series of operations from the selection in (I) to the final movement of a freight car for achieving the desirable layout is defined as a trial.

Now, define h as the number of candidates of the desired layout of groups. Each candidate in operation (I) is represented by u_{j_1} $(1 \leq j_1 \leq h)$.

In the operation (II), as a result of classification, a sub-track for each car is determined from the tail of the train. The sub-track is defined as T_C, and candidates of T_C are defined as $u_{j_2}(h+1 \leq j_2 \leq h+m)$. u_{j_2} are sub-tracks each of which have the car belonging to the group moved to the main track before the car of the tail of the incoming train. When there is no such sub-track, T_C is selected from m sub-tracks. Then, the number of groups classified to T_C is determined. Candidates are groups that satisfy the movement-order selected in (I), and are defined as u_{j_3} $(h+m+1 \leq j_3 \leq h+m+v)$, where v is the number of the candidates.

In the operation (IV), each group has the predetermined position in the main track. The car to be rearranged is defined as c_T, and candidates of c_T can be determined by excluding freight cars that have already rearranged to the main track. These candidates must belong to the same group.

Also, define r as the number of groups, g_l as the number of freight cars in $\mathrm{group}_l (1 \leq l \leq r)$, and $u_{j_4}(h+m+v+1 \leq j_4 \leq h+m+v+g_l)$ as candidates of c_T.

In the operation (V), the removal destination of cars located in front of the car to be rearranged is defined as c_M. Then, defining $u_{j_5}(h+m+v+g_l+1 \leq j_5 \leq h+m+v+g_l+m-1)$ as candidates of c_M, excluding the sub-track that has the car to be removed, and the number of candidates is $m-1$.

In the operation (VI), defining p as the number of removal cars required to rearrange c_T, and defining q as the number of removal cars that can be located the sub-track selected in the operation (V), the candidate numbers of cars to be moved are determined by u_{j_6} $(1 \leq u_{j_6} \leq \min\{p,q\}, h+2m+v+g_l \leq j_6 \leq h+2m+v+g_l+\min\{p,q\})$.

In both cases of Fig. 5, the direct rearrangement is conducted for c_2 at step 1, and the selection of c_T conducted at step 2, candidates are $u_{h+m+v+1} =$

[1], $u_{h+m+v+2}$ = [4], that is, sub-tracks where cars in group$_1$ are located at the top. $u_{h+m+v+3}$, $u_{h+m+v+4}$ are excluded from candidates. Then, $u_{h+m+v+2}$ = [4] is selected as c_T. Candidates for the location of c_9 are $u_{h+m+v+5}$ = [1], $u_{h+m+v+6}$ = [2], $u_{h+m+v+7}$ = [3], sub-tracks [1], [2], and [3]. In case 1, $u_{h+m+v+6}$ = [2] is selected as c_M, and in case 2, $u_{h+m+v+7}$ = [3] is selected. After direct rearrangements of c_3 at step 3 and c_1, c_4, c_5 at step 4, the marshaling process is finished at step 5 in case 2. Then, total step counts of marshaling process for case 2 is 5, whereas 6 for case 1.

6 Processing Time for a Movement of Locomotive

6.1 Transfer Distance of Locomotive

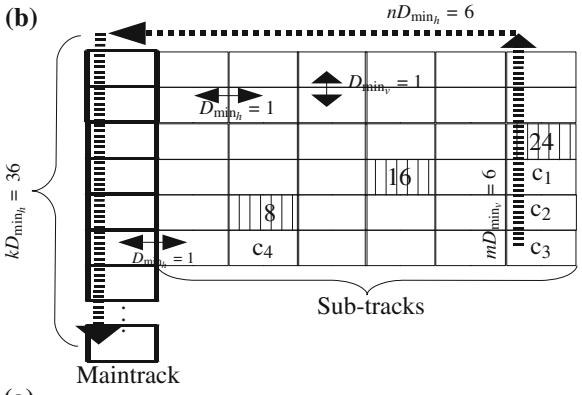
When a locomotive transfers freight cars, the process of the unit transition is as follows: (E1) starts without freight cars, and reaches to the joint track, (E2) restart in reverse direction to the target cars to be moved, (E3) joints them, (E4) pull out them to the joint track, (E5) restart in reverse direction, and transfers them to the indicated location, and (E6) disjoints them from the locomotive. Then, the transfer distance of locomotive in (E1), (E2), (E4) and (E5) is defined as D_1, D_2, D_3 and D_4, respectively. Also, define the unit distance of a movement for cars in each sub-track as $D_{\min_v}$, the length of joint track between adjacent sub-tracks, or, sub-track and main track as $D_{\min_h}$. The location of the locomotive at the end of above process is the start location of the next movement process of the selected car. Also, the initial position of the locomotive is located on the joint track nearest to the main track.

Figure 6 shows an example of transfer distance. In the figure, $m = n = 6$, $D_{\min_v} = D_{\min_h} = 1$, $k = 18$, (a) is position index, and (b) depicts movements of locomotive and freight car. Also, the locomotive starts from position 8, the target is located on the position 18, the destination of the target is 4, and the number of cars to be moved is 2. Since the locomotive moves without freight cars from 8 to 24, the transfer distance is $D_1 + D_2 = 12$ ($D_1 = 5$, $D_2 = 7$), whereas it moves from 24 to 16 with 2 freight cars, and the transfer distance is $D_3 + D_4 = 13$ ($D_3 = 7$, $D_4 = 6$).

6.2 Processing Time for the Unit Transition

In the process of the unit transition, the each time for (E3) and (E6) is assumed to be the constant t_E.

The processing times for elements (E1), (E2), (E4) and (E5) are determined by the transfer distance of the locomotive $D_i (i = 1, 2, 3, 4)$, the weight of the freight cars W moved in the process, and the performance of the locomotive. Then, the time each for (E1), (E2), (E4) and (E5) is assumed to be obtained by the function



31	32	33	34	35	36
25	26	27	28	29	30
19	20	21	22	23	24
13	14	15	16	17	18
7	8	9	10	11	12
1	2	3	4	5	6

Fig. 6 Calculation of transfer distance

$f(D_i, W)$ derived considering dynamics of the locomotive, limitation of the velocity, and control rules. Thus, the processing time for the unit transition t_{U} is calculated by $t_{\mathrm{U}} = \mathrm{t_E} + \sum_{i=1}^{2} f(D_i, 0) + \sum_{j=4}^{5} f(D_j, W)$. The maximum value of t_{U} is define as $t_{\max}$ and is calculated by

$$t_{\max} = \mathrm{t_E} + f(kD_{\min_v}, 0) + f(mD_{\min_h}, 0) + f(mD_{\min_h} + n, W) + f(kD_{\min_v}, W). \tag{1}$$

7 Learning Algorithm

Defining G_{o} as the desired layout selected among u_{j_1}, $Q_1(G_{\mathrm{o}})$ is updated by the following rule when one of desired layout is achieved in the main track:

$$Q_1(G_{\mathrm{o}}) \leftarrow \max \left\{ \begin{array}{c} Q_1(G_{\mathrm{o}}), \\ (1-\alpha) Q_1(G_{\mathrm{o}}) + \alpha \mathrm{R} \prod_{i=1}^{l_{\mathrm{a}}} \gamma_i \end{array} \right\} \tag{2}$$

where l_{a} denotes the total movement counts required to achieve the desired layout, α is learning rate, γ is discount factor calculated for each movement, R is reward that is given only when one of desired layout is achieved in the main track.

Define $s(t)$ as the state at time t, T_{c} as the sub-track selected as the destination for the removed car, p_{C} as the number of classified groups, q_{M} as the movement counts

of freight cars by direct rearrangement, and s' as the state that follows s. In the classification stage, Q_2, Q_3 are defined as evaluation values for (s_1, u_{j_2}), (s_2, u_{j_3}) respectively, where $s_1 = [s, G_o]$, $s_2 = [s_1, T_c]$. $Q_2(s_1, T_c)$ and $Q_3(s_2, p_C)$ are updated by following rules:

$$\begin{aligned} Q_2(s_1, T_c) &\leftarrow \max_{u_{j_3}} Q_3(s_1, u_{j_3}), \\ Q_3(s_2, p_C) &\leftarrow (1-\alpha) Q_3(s_2, p_C) + \alpha V_1, \end{aligned} \tag{3}$$

$$V_1 = \begin{cases} R \prod_{i=1}^{l_a} \gamma_i \\ \text{(all cars assigned)} \\ \gamma \max_{u_{j_2}} Q_2(s_1, u_{j_2}). \\ \text{(otherwise)} \end{cases} \tag{4}$$

Define p_M as the number of cars selected in operation (VI), and l_b as the number of movements of cars conducted sequentially in the direct rearrangement. Q_4, Q_5 and Q_6 are defined as evaluation values for (s_1, u_{j_4}), (s_3, u_{j_5}), (s_4, u_{j_6}) respectively, where $s_3 = s_1$, $s_4 = [s_3, c_T]$, $s_5 = [s_4, c_M]$. $Q_4(s_3, c_T)$, $Q_5(s_4, c_M)$ and $Q_6(s_5, p_M)$ are updated by following rules:

$$Q_4(s_3, c_T) \leftarrow \max_{u_{j_5}} Q_5(s_4, u_{j_5}), \tag{5}$$

$$Q_5(s_4, c_M) \leftarrow \max_{u_{j_6}} Q_6(s_5, u_{j_6}), \tag{6}$$

$$Q_6(s_5, p_M) \leftarrow \tag{7}$$

$$\begin{cases} (1-\alpha) Q_6(s_5, p_M) + \alpha \left[R + V_2 \prod_{i=1}^{l_b+1} \gamma_i \right], \\ (u \text{ is a rearrangement}) \\ (1-\alpha) Q_6(s_5, p_M) + \alpha [R + \gamma V_3], \\ (u \text{ is a removal}) \end{cases}$$

$$V_2 = \max_{u_{j_4}} Q_4(s_3', u_{j_4}),\ V_3 = \max_{u_{j_5}} Q_5(s_4', u_{j_5}).$$

7.1 Calculation of γ

A discount is conducted in each movement of freight cars. When the movement counts of freight cars are evaluated, γ is set as constant.

When the transfer distance is evaluated, γ is used to reflect the transfer distance of the locomotive and calculated by the following equation:

$$\gamma = \delta \frac{D_{\max} - \beta D}{D_{\max}},\ 0 < \beta < 1, 0 < \delta < 1. \tag{8}$$

When the processing time is evaluated, γ is used to reflect the total processing time of marshaling and calculated by the following equation:

$$\gamma = \delta \frac{t_{\max} - \beta t_{\mathrm{U}}}{t_{\max}},\ 0 < \beta < 1, 0 < \delta < 1. \tag{9}$$

Propagating Q-values by using Eqs. (5)–(9), Q-values are discounted according to the number of movements of freight cars. Then, the magunitude of γ reflects the target of evaluation, that is, "time", "distance" or "movement counts". By selecting the movement that has the largest Q-value after adequate learning, the best marshaling plan can be obtained [10].

In the learning stages, each u_j $(1 \leq j \leq h + 2m + v + g_l + \min\{p, q\})$ is selected by the soft-max action selection method [10]. In order to normalize the probability of the selection through a marshaling process, probability P for selection of each candidate is calculated by the following manner [5]:

$$\tilde{Q}_i(s_{i-1}, u_{j_i}) = \frac{Q_i(s_{i-1}, u_{j_i}) - \min_u Q_i(s_{i-1}, u_{j_i})}{\max_u Q_i(s_{i-1}, u_{j_i}) - \min_u Q_i(s_{i-1}, u_{j_i})}, \tag{10}$$

$$P(s_{i-1}, u_{j_i}) = \frac{\exp(\tilde{Q}_i(s_{i-1}, u_{j_i})/\xi)}{\sum_{u \in u_{j_i}} \exp(\tilde{Q}_i(s_{i-1}, u)/\xi)}, \tag{11}$$

$$(i = 2, 3, 4, 5, 6),$$

$$P(u_{j_1}) = \frac{\exp(Q_1(u_{j_1})/\xi)}{\sum_{u \in u_{j_1}} \exp(Q_1(u)/\xi)},$$

where ξ is the thermo constant that determine the range of probability.

8 Computer Simulations

Computer simulations are conducted for $m = 12, n = 6, k = 36$ and learning performances of following 5 methods are compared:

(A) proposed method that evaluates the processing time of the marshaling operation, considering the layout of groups and classification of incoming cars,
(B) a method that evaluates the processing time considering the layout of groups, and the classification is fixed,
(C) a method that evaluates the processing time, has the fixed layout of groups, and has the fixed classification,
(D) the method same as (A) except for evaluating the movement counts of freight cars,

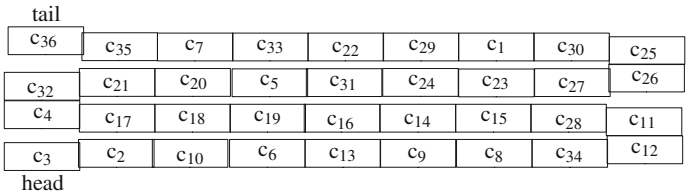


Fig. 7 Arrangement of incoming train

c_{10}	c_9	c_{12}	c_{15}	c_{19}	c_4	c_{20}	c_{24}	c_{26}	c_1	c_{33}	c_{36}
c_2	c_{13}	c_{34}	c_{28}	c_{16}	c_{17}	c_{21}	c_{31}	c_{27}	c_{30}	c_{22}	c_{35}
c_3	c_6	c_8	c_{11}	c_{14}	c_{18}	c_{32}	c_5	c_{23}	c_{25}	c_{29}	c_7

Fig. 8 Classification for methods (B), (C)

(E) the method same as (A) except for evaluating the transfer distance of the locomotive.

The initial arrangement of incoming train is described in Fig. 7. The original rearrangement order of groups is $\text{group}_1, \text{group}_2, \text{group}_3, \text{group}_4$. Cars $c_1, \ldots, c_9$ are in group_1, $c_{10}, \ldots, c_{18}$ are in group_2, $c_{19}, \ldots, c_{27}$ are in group_3, and $c_{28}, \ldots, c_{36}$ are in group_4. Other parameters are set as $\alpha = 0.9$, $\beta = 0.2$, $\delta = 0.9$, $\text{R} = 1.0$, $\xi = 0.1$. Method (C) accepts only the original rearrangement order of groups, whereas other methods consider extended layout of groups. In methods (B), (C), the classification generates the fixed layout depicted in Fig. 8.

The locomotive assumed to accelerate and decelerate the train with the constant force 100×10^3 N, and to be 100×10^3 kg in weight. Also, all the freight cars have the same weight, 10×10^3 kg. The locomotive and freight cars assumed to have the same length, and $D_{\min_v} = D_{\min_h} = 20$ m. The velocity of the locomotive is limited to no more than 10 m/s. Then, the locomotive accelerates the train until the velocity reaches 10 m/s, keeps the velocity, and decelerates until the train stops within the indicated distance. When the velocity does not reach 10 m/s at the half way point, the locomotive starts to decelerate immediately.

The results are shown in Fig. 9. In the figure, horizontal axis expresses the number of trials and the vertical axis expresses the minimum processing time to achieve a desirable layout found in the finished trials. Each result is averaged over 20 independent simulations. In Fig. 9, the learningF performance of method (A) is better than that of method (B), because solutions derived by method (A) considers the classification of groups effectively for reducing the total processing time. Since the group layout for the outbound train and classification are fixed, method (C) is not effective to reduce the total processing time as compared to methods (A), (B). Moreover, method (A) generates better marshaling plan as compared to methods (D) and (E). Since the total processing time dependes on both the movement counts of freight cars and the transfer distance of the locomotive, the learning performance can be spoiled in methods (D), (E) that evaluate only the movement counts of freight cars

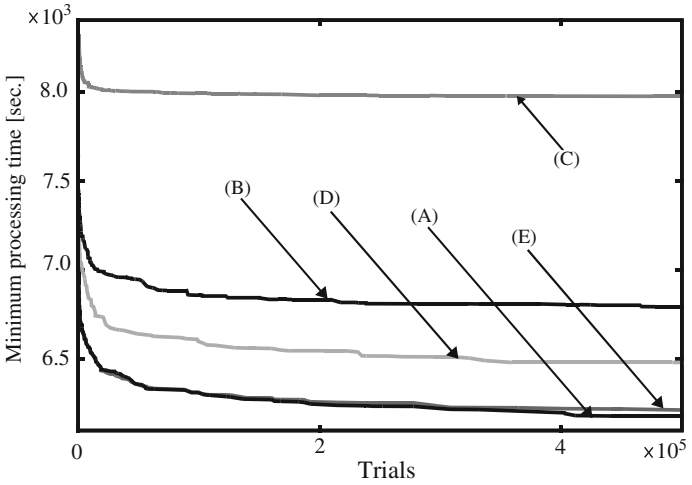


Fig. 9 Comparison of learning performances

Table 1 Total processing time

Methods	Processing time (s)		
	Best	Average	Worst
Method (A)	5876.332	6157.908	6294.066
Method (B)	6437.393	6785.930	6919.313
Method (C)	7857.711	7973.113	8003.003
Method (D)	6179.987	6424.613	6623.311
Method (E)	5876.322	6168.090	6342.982

or the transfer distance of locomotive. Total transfer distances of the locomotive at 1.5×10^6th trial are described in Table 1 for each method.

9 Conclusions

A new scheduling method has been proposed in order to rearrange and line cars in the desirable order onto the main track considering classifications for incoming freight cars. The learning algorithm of the proposed method is derived based on the reinforcement learning, considering the total processing time of marshaling, the transfer distance of the locomotive, or the movement counts of freight cars. In order to improve the evaluation of marshaling, the proposed method learns the classification of incoming cars and the layout of groups, as well as the arrangement of freight cars in each group, the rearrangement order of cars, the number of cars to be moved and the removal destination of cars for assembling an outgoing train, simultaneously.

In computer simulations, the marshaling plan and the learning performance of the proposed method has been improved by learning assignment of incoming cars and arrangement of outgoing cars with evaluation of processing time.

References

1. Blasum U, Bussieck MR, Hochstättler W, Moll C, Scheel HH, Winter T (2000) Scheduling trams in the morning. Mathl Methods Oper Res 49(1):137–148
2. Dahlhaus E, Manne F, Miller M, Ryan J (2000) Algorithms for combinatorial problems related to train marshalling. In: Proceedings of the 11th Australasian workshop on combinatorial algorithms, pp 7–16
3. Eggermont C, Hurkens CAJ, Modelski M, Woeginger GJ (2009) The hardness of train rearrangements. Oper Res Lett 37:80–82
4. He S, Song R, Chaudhry S (2000) Fuzzy dispatching model and genetic algorithms for railyards operations. Euro J Oper Res 124(2):307–331
5. Hirashima Y (2011) A new rearrangement plan for freight cars in a train: Q-learning for minimizing the movement counts of freight cars. Lecture Notes in Electrical Engineering, 70 LNEE, pp 107–118
6. Hirashima Y (2011) A reinforcement learning method for train marshaling based on movements of locomotive. IAENG Int J Comput Sci 38(3):242–248
7. Hirashima Y (2013) An intelligent train marshaling plan of freight cars based on the processing time considering group layout. In: Lecture notes in engineering and computer science: proceedings of the international multiConference of engineers and computer scientists 2013, IMECS 2013, Hong Kong, pp 30–35
8. Jacob R, Marton P, Maue J, Nunkesser M (2007) Multistage methods for freight train classification. In: Proceedings of 7th workshop on algorithmic approaches for transportation modeling, optimization, and systems, pp 158–174
9. Kroon L, Lentink R, Schrijver A (2008) Shunting of passenger train units: an integrated approach. Transport Sci 42:436–449
10. Sutton R, Barto A (1999) Reinforcement learning. MIT Press, Cambridge, MA
11. TOMII N, Jian ZL (2000) Depot shunting scheduling with combining genetic algorithm and pert. In: Proceedings of 7th international conference on computer aided design, manufacture and operation in the railway and other advanced mass transit systems, pp 437–446
12. Watkins CJCH, Dayan P (1992) Q-learning. Mach Learn 8:279–292

A Neural-Network-Based Hand Posture Recognition Method

Yea Shuan Huang and Yun Jiun Wang

Abstract In various pattern recognition applications, angle variation is always a main challenging factor for producing reliable recognition. To increase the endurance ability on angle variation, this paper adopts a Hierarchical Temporal Memory (HTM) algorithm which applies temporal information to organize time-sequence change of image features, and constructs invariant features so that the influence of angle variation can be effectively learnt and overcome. The proposed multi-angle HTM-based posture recognition method consists of two main modules of Hand Posture Image Pre-processing (HPIP) and Hand Posture Recognition (HPR). In HPIP, each input image is first processed individually by skin color detection, foreground segmentation and edge detection. Then, the three processed results are further combined linearly to locate a hand posture region. In HPR, the normalized image is forwarded to a HTM model for learning and recognizing of different kinds of hand postures. Experiment results show that when using the same continuous unconstrained hand posture database, the proposed method can achieve an 89.1 % high recognition rate for discriminating three kinds of hand postures, which are scissors, stone and paper, and outperforms both Adaboost (78.1 %) and SVM (79.9 %).

Keywords Foreground segmentation · Hierarchical temporal memory · Multi-angle hand posture recognition · Skin-color detection · Spatial pooler clustering · Temporal pooler grouping

Y. S. Huang (✉) · Y. J. Wang
Department of Computer Science and Information Engineering, Chung-Hua University, No. 707, Sec. 2, WuFu. Rd., Hsinchu, Taiwan
e-mail: yeashuan@chu.edu.tw

Y. J. Wang
e-mail: cheisea@hotmail.com

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 275, DOI: 10.1007/978-94-007-7684-5_14,

1 Introduction

In recent years, the interaction of Human Computer Interface (HCI) has become a popular research domain. Many latest types of HCI has come out gradually such as LED-based multi-touch sensor for LCD screens [1], audio modeling system [2], and moving object sensory detection system [3]. Comparing to other methods, moving object sensory detection system requires only a commonly-used camera to interact with users, which is the most nature and convenient way for HCI and therefore has gained much research attention in recent years. In general, a posture recognition system mainly contains two processing modules: posture detection and posture recognition. For posture detection, it could be divided into two types of processing methods. The first type is skin color detection which locates the posture area by either using one set or multiple sets of threshold values to determine the range of skin color in a certain color space or applying the method of probability learning model to compute the probability that a certain color belongs to skin color. The second type is the moving object detection which in general considers only a single moving hand in the image so that the location of the hand area can be known easily when a moving object was detected. There are three common ways in detecting moving objects, which are background subtraction [4], temporal differencing [5], and optical flow [6]. The first two apply subtraction of different images and determine the changed part of image as foreground and the non-changed part of image as background. While optical flow utilizes the displacement of moving objects from multiple images to count the variation of gray-scale value of images at a time for each pixel point, and construct the speed field formed by speed vector. As for posture recognition, Adaboost, Support Vector Machine (SVM), and Neural Network are the most commonly used recognition algorithms. Three methods have each pros and cons on recognition accuracy and tolerance limits. For example, the recognition rate of SVM is better than those of the other two. However, when it comes to the context of image preprocessing of training the data base, Adaboost has a higher correct rate than others. Furthermore, if only the execution speed of recognition is considered, Adaboost is the fastest in general.

While developing a hand posture recognition system, image noise, light, and angle may influence the recognition accuracy considerably. Therefore, the purpose of this study is to figure out how to derive stable features from posture images under the environment of varying light and angles, and to construct a robust hand posture recognizer. To this end, we propose a multi-angle HTM-based posture recognition method which includes a useful hand posture region locating step and a forearm excluding step to derive the proper hand posture region, and adopts a HTM algorithm to induce the continuous changing images and form invariant features so that the constructed classifier is insensitive to various noise such as angle variation.

2 Modular Architecture

This section presents the overall framework of the proposed method which mainly contains two modules of Hand Posture Image Pre-processing (HPIP) and Hand Posture Recognition (HPR) as shown in Fig. 1. At the HPIP module, each input image is first processed by skin color detection, foreground segmentation, and edge detection individually. Then, the three processed results are combined with a linear weighting method to acquire the hand posture region. Then, the image of the processed hand posture region will be normalized to a fixed size. At the HPR module, the normalized hand posture image is forwarded to a HTM model for learning and recognizing of different kinds of hand postures.

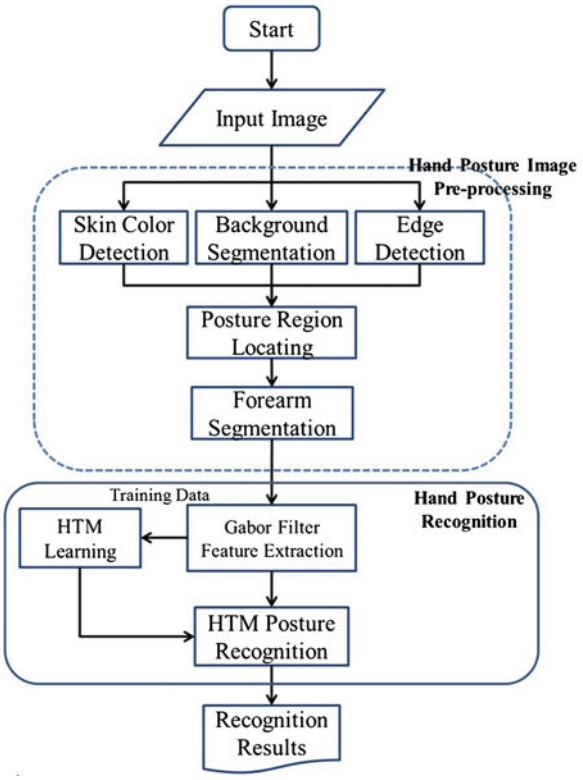


Fig. 1 Flow chart of hand posture recognition algorithm

3 Hand Posture Image Pre-processing

The purpose of this module is to locate the valid hand posture region from the input image. To this end, skin color detection, foreground segmentation, and edge detection are performed on the input image, and their processed results are further combined with a linear weighting method. Then a forearm segmentation process will check and derive the valid hand posture region. Here, a Codebook algorithm [7] is used for foreground segmentation, and a 3×3 Sobel operator is used to detect the edge pixels.

3.1 Skin Color Detection

Since most pixels of hands belong to skin color, skin color detection becomes essential in locating hands from images. We adopt a Bayesian Classifier [8] with YC_bC_r color space to detect the skin color pixels. By manually identifying skin and non-skin color pixels on images, a Bayesian skin-color detector based on C_bC_r information can be trained. Let H_S and H_n be individually the histograms of skin and non-skin color pixels, (C_b, C_r) denote a color vector, $s\left[(C_b, C_r)\right]$ and $n\left[(C_b, C_r)\right]$ be the pixel numbers of color (C_b, C_r) in H_S and H_n respectively, C_s and C_n represent the total pixel numbers of H_S and H_n, $P\left((C_b, C_r)|skin\right)$ and $P\left((C_b, C_r)|nonskin\right)$ be the probability that a skin pixel appears a specific (C_b, C_r) value, $P\left((C_b, C_r)|nonskin\right)$ be the probability that a non-skin pixel appears a specific color vector (C_b, C_r), $P\,(skin)$ and $P\,(nonskin)$ be the proportion of skin and non-skin pixels, and $P\left(skin|(C_b, C_r)\right)$ be the probability that apixel with color(C_b, C_r) is indeed a skin pixel. When a pixel with color vector (C_b, C_r), then the probability that it is a skin pixel becomes

$$P\left(skin|(C_b, C_r)\right) = \frac{P\left((C_b, C_r)|skin\right) P(skin)}{P\left((C_b, C_r)|skin\right) P\,(skin) + P\left((C_b, C_r)|nonskin\right) P(nonskin)}$$

$$P\left((C_b, C_r)|skin\right) = \frac{s[(C_b, C_r)]}{C_s}$$

$$P\left((C_b, C_r)|nonskin\right) = \frac{n\left[(C_b, C_r)\right]}{C_n}$$

$$P\,(skin) = \frac{C_s}{C_s + C_n},\ P\,(nonskin) = \frac{C_n}{C_s + C_n}$$

We defined a threshold θ, when $P\left(skin|(C_b, C_r)\right) > \theta$, then a pixel with color vector (C_b, C_r) is determined to be a skin pixel; otherwise, it is determined to a non-skin pixel. In this paper, θ is set to 0.06.

3.2 Foreground Segmentation

In the foreground segmentation step, we adopt the Codebook background model which quantizes and clusters the variation patterns of background of each pixel. To construct the Codebook background model [7], according to a sequence of images each pixel will create its own codebook which contains a number of codeword. Assume that $C = \{c_1, c_2, \ldots, c_L\}$ represents the codebook of one pixel which has L codeword, each codeword c_i consists of a *RGB* color vector $v_i = \left\{\overline{R_i}, \overline{G_i}, \overline{B_i}\right\}$ and six parameters $aux_i = \left(\breve{I}_I \hat{I}_i, f_i, \lambda_i, p_i, q_i\right)$, where $\breve{I}$ and $\hat{I}$ represent the maximum and minimum luminance of this codeword, f is the number of success match, λ is the value of Maximum Negative Run-Length (MNRL) which is the longest time that this codeword has not been updated, p is the first time of this codeword to appear, and q is the last time of this codeword to appear. In the training step, let $X = \{x_1, x_2, \ldots, x_N\}$ be the N RGB color vectors at a certain pixel position in N continuous images. For each x_t, if it is similar to an existed codeword c_m, then c_m will be updated; however if there isn't any existed codeword similar to x_t, a new codeword should be constructed. The closeness of two colors is examined by two measurements: color distance and luminance distance. Only the two measurements are small enough, the two colors are considered to be close.

3.3 Hand Region Locating and Forearm Segmentation

After performing foreground segmentation, edge detection, and skin color detection on each input image, three binary images (0 and 255) are generated. With a linear weighting method, the three binary images are combined together by

$$Result\,(x) = \begin{cases} 255, \text{if} \left[\begin{array}{l} \alpha \times Res_{Skin}\,(x) \\ +\beta \times Res_{Edge}\,(x) \\ +\gamma \times Res_{Codebook}(x) \end{array}\right] > 128 \\ 0, \quad \text{otherwise} \end{cases}$$

where $Result\,(x)$ is the integrated binary value of x, $Res_{Skin}\,(x)$, $Res_{Edge}(x)$ and $Res_{Codebook}(x)$ are the individual binary result of skin color detection, edge detection and foreground segmentation, and α, $\beta$and γ are three weight values representing the importance of the three kinds of responses. By applying morphology operation and connected component analysis on the integrated binary image, the valid hand region can be easily obtained.

For hand posture recognition, both palm and fingers are important, but the forearm part is irrelative and may even degrade the recognition rate. Therefore, it is necessary to remove the forearm information effectively. For this purpose, a forearm removal mechanism is designed. The main idea of this mechanism is deduced from the facts

that palm and fingers in general contains much more edge information than the forearm part, and the gravity center (GC) of the detected edge points in a valid hand region generally is located at the upper part of palm. Therefore, if a GC is estimated, then the proper hand region could be approximately defined. With this realization, we compute the gravity center from the detected edge points of the extracted valid hand region first, then the four distances (x_1, x_2, y_1, and y_2) between this GC and the left, right, top, and bottom boundaries of the valid hand region are calculated individually. Afterwards, the left, right, and bottom boundaries are updated independently if some of the following situations are matched.

$$\begin{cases} \text{right boundary} = GC.x + 1.1 \times x_1 & \text{, if } x_2 > 1.1 \times x_1; \\ \text{left boundary} = GC.x - 1.1 \times x_2 & \text{, if } x_1 > 1.1 \times x_2; \\ \text{bottom boundary} = GC.y + 1.2 \times y_1 & \text{, if } y_2 > 1.2 \times y_1; \end{cases}$$

After that, each resultant hand image is normalized into 128×128 pixels, and this normalized image becomes the input data to the successive HTM model for training and recognizing different kinds of hand postures.

4 Hand Posture Recognition

Hierarchical Temporal Memory (HTM) [9] is a biometric engine model based on the memory-prediction theory of brain function described by Jeff Hawkins in his book On Intelligence [10]. HTM is able to discover and infer the high-level causes of observed input patterns and sequences, so that it can build an increasingly complex model of the world. By learning the sequence of continuous images, HTM can get invariant features of various kinds of objects so that it is able to get robust recognition on unseen patterns. Each HTM region learns by identifying and memorizing spatial patterns - combinations of input bits that often occur at the same time. It then identifies temporal sequences of spatial patterns that are likely to occur one after another. After an HTM has learned the patterns in its world, it can perform inference robustly on novel inputs [11, 12].

4.1 Structure of HTM Network

A typical HTM network is a tree-shaped hierarchy of levels that are composed of smaller elements called nodes. A single level in the hierarchy is also called a region. Higher hierarchy levels can reuse patterns learned at the lower levels by combining them to memorize more complex patterns. HTM uses the time clue to perform an unsupervised learning, summarize the training data, and produce the invariant feature. Figure 2 shows the basic computation structure of HTM which mainly contains four kinds of processing nodes:

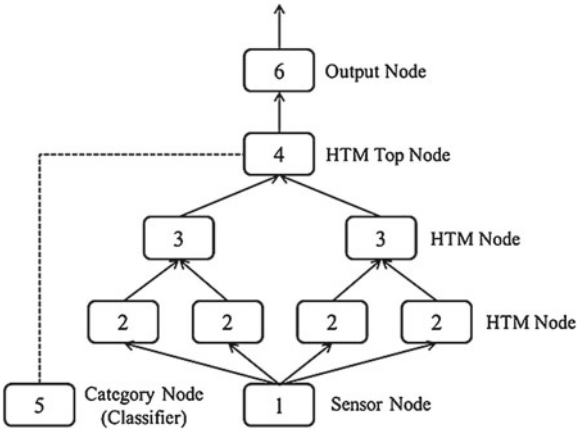


Fig. 2 Structure of HTM network

1. Sensor Node: it locates at the bottom layer of a HTM network, receives the sensed data (such as images, sound and pressure measurements, etc.) and transports the data to the one-layer above HTM nodes.
2. HTM node: it is the main computation node of HTM and can be formed into several levels according to the complexity of the dealing problem. Each HTM node has the same basic functionality which performs unsupervised learning to infer the input data and produce time and space similar categories. In the learning stage, nodes in lower level can derive smaller concepts which are more restricted in time and space, and nodes of upper level can derive more general concepts. Input data comes in from the lower level nodes and the HTM nodes output the generated concepts. The top level usually has a single node that stores the most genera concepts. When in inference mode, a node in each level interprets information coming in from its child nodes in the lower level as probabilities of the concepts it has memorized.
3. Category Node: during the training stage, it performs a supervised learning algorithm (such as SVM) to construct a robust classifier based on the invariant feature inferred by the HTM top node. Then, during the recognition stage, it produces the recognition result for each input pattern.

Briefly speaking, the sensor node at the bottom of network gets the data to learn and train the model; it will produce the inference belief and transport to the HTM node in next level. In Fig. 2, the inference belief is transported from sensor node2 to HTM node3, and then further transported from HTM node3 to HTM top node4. The HTM top node4 will then transport the summarized invariant features to the category node5, and the category node5 uses a classifier to recognize and output the final recognition result.

4.2 HTM Algorithm

HTM adopts a Bayesian network, spatial and temporal clustering algorithms with a tree-shaped hierarchy of neural nodes. Each HTM node contains a Spatial Pooler (SP) and a Temporal Pooler (TP) which performs two stages of processes (learning stage and inference stage). During learning, a node receives a temporal sequence of spatial patterns as its input. SP identifies frequently observed patterns and memorizes them as coincidences. Patterns that are considerably similar to each other are treated as the same coincidence, so a large amount of possible input patterns are reduced to a few number of known coincidences. TP partitions coincidences that are likely to follow each other in the training sequence into temporal groups. During inference (recognition), the node calculates the set probabilities that a pattern belongs to each known coincidence. Then it calculates the probabilities that the input represents each temporal group. The set of probabilities assigned to the groups is called a node's "belief" about the input pattern. This belief is the result of the inference that is passed to one or more "parent" nodes in the next higher level of the hierarchy. Figure 3 shows a schematic example of the learning process results in SP and TP. Figure 3a displays that there are in total 6 coincidences (#1–#6) stored in a SP. Figure 3b shows that the 6 spatial coincidences are clustered into 3 temporal groups (G1, G2 and G3) because coincidence #1 and #2, coincidence #3 and #4, and coincidence #5 and #6 have large correlations and often appear successively in time sequence. Basically, each temporal group takes one bin to represent, so there is a 3-bin output of TP. Figure 3c shows that after training SP and TP, a specific input pattern is taken into the HTM and it matches coincidence #3 in SP and group 2 in TP. Through

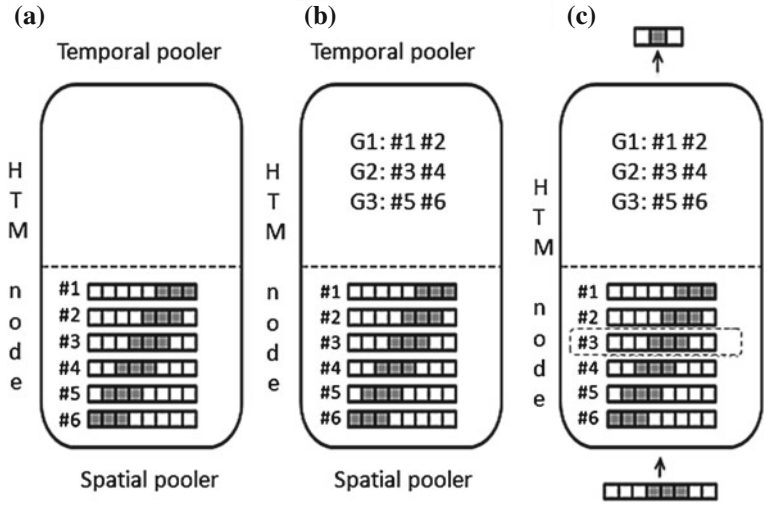


Fig. 3 HTM learning process, **a** SP learns the training data, **b** TP groups the data with time inference, **c** invariant feature inference on unknown data

both spatial and temporal clustering mechanisms, resolution in space and time is reduced considerably. Therefore, the derived information from an input pattern is rather insensitive to both spatial and temporal variations, and can be regarded as an invariant feature for representing this input pattern to further recognition.

(B.1) *Spatial Pooler Clustering*

The most important task of SP is to quantify the infinite input data and transform it into a limited number of coincidences. Each input pattern is first transformed into a fixed-size vector, and then this vector is compared with every coincidence recorded in SP. If there is no similar coincidence to this vector, a new coincidence should be constructed by taking this vector as its value and setting its occurrence number to be 1; otherwise, the occurrence number of the similar coincidence should be increased by one. For measuring the similarity between an input pattern x and a coincidence w, the Euclidean distance is calculated

$$d(x, w) = \sqrt{\sum_{j=1}^{N} (x_j - w_j)^2}$$

where N is the vector dimension of x and w. Let D be a threshold. If $d(x, w)$. is smaller than D, w and x are regarded to be similar to each other, otherwise they are not similar. If D is set too high, the number of coincidences will be small, and this will cause the inference ability of a HTM node degenerated. However, if D is set too low, the number of coincidences will be rather large, and this will not only take longer computation time in the inference stage but also decrease the generalization ability of this HTM node.

(B.2) *Temporal Pooler Grouping*

After SP has constructed all coincidences, TP then proceeds to perform the temporal clustering and construct the frequently occurred temporal groups. Each temporal group may contain several incidences which are highly related to occur in time sequence with at least one other coincidence in the group. To do this, the HTM algorithm builds a time-adjacency matrix to record the occurrence numbers of every two

Table 1 Temporal adjacency matrix

	#1	#2	#3	#4	#5	#6	#7	#8
#1	4	12	13	1	8	2	3	1
#2	0	1	6	9	2	3	14	2
#3	8	7	2	2	5	0	4	11
#4	0	8	4	3	1	3	2	9
#5	8	1	4	1	3	0	2	1
#6	1	4	1	2	1	5	0	12
#7	1	9	4	0	1	3	4	2
#8	3	1	5	11	0	8	3	1

coincidences by sequentially sending all the training data into the trained SP. Table 1 shows one example of time-adjacency matrix, where (i, j). represents the situation that coincidence i appears right after coincidence j, and the value of (i, j). is the occurrence number of situation (i, j). For example, the number in (6,8) is 12, it means that there are in total 12 times that coincidence 6 appears after coincidence 12 in the training sequence. After sending all training data to a SP, a $N \times N$ time-adjacency matrix is obtained and N is the total number of coincidences in this SP. Then, the time-adjacency matrix is further transformed into a Markova graph, and this graph records the transformation probabilities of all coincidence pair as shown in Fig. 4. These transformation probabilities indeed become the fundamental information for the following TP clustering process.

According to the Markov graph transformed from the time-adjacency matrix, the temporal pooler grouping by Agglomerative Hierarchical Clustering (AHC) [19] is performed. AHC sequentially clusters the coincidences having high correlation into the the same group and then removes all members of this group from the Markov graph. The clustering process is iterated until every coincidence has belonged to one group. Taking Fig. 5 as an illustration example, the first group contains coincidences

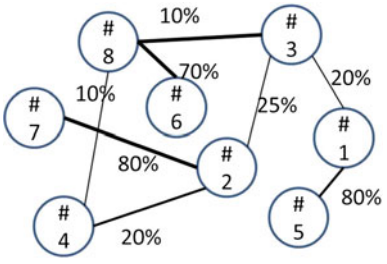


Fig. 4 Markov chart of Table 1

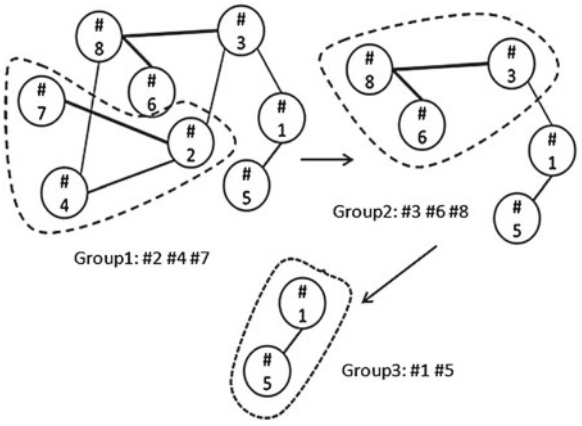


Fig. 5 Example of agglomerative hierarchical clustering

#2, #4 and #7, the second group contains coincidences #3, #6 and #8, and finally the third group contains coincidences #1 and #5.

5 Result and Experiment

In order to evaluate the performance of the proposed method, a database with various angles of three kinds of hand postures (rock, paper and scissor) is constructed. This database involves four persons and eight different environments including one simple background scene and seven complex background scenes (as shown in Fig. 6). The image resolution is 640 × 480 pixels, and the range of angle variation is ±60° from left to right [as shown in Fig. 7(a)], ±45° from top to bottom [as shown in Fig. 7(b)],

Fig. 6 Various hand postures in different backgrounds

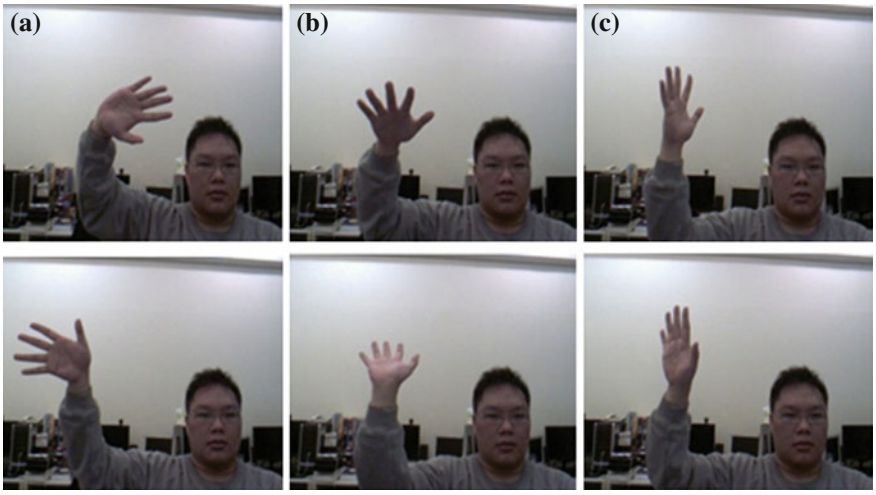


Fig. 7 Examples of orientation variation, **a** ±60° from *left to right*, **b** ±45° from *top to bottom*, and **c** ±30° in a horizontal self-spin

Table 2 Results of recognizing multi-angle hand postures

Types of background	Hand postures	Recognition rate (%)	Average (%)
Singles background	Scissor	95.5	95
	Stone	96	
	Paper	93.5	
Complex background (1)	Scissor	87.5	87.3
	Stone	93	
	Paper	81.5	
Complex background (2)	Scissor	91.5	91.8
	Stone	95.5	
	Paper	89.5	
Complex background (3)	Scissor	92.5	91
	Stone	94.5	
	Paper	86	
Complex background (4)	Scissor	80	81.7
	Stone	88.5	
	Paper	76.5	
Complex background (5)	Scissor	90.5	89.2
	Stone	92.5	
	Paper	84.5	
Complex background (6)	Scissor	82	84.8
	Stone	88.5	
	Paper	84	
Complex background (7)	Scissor	92	91.5
	Stone	95	
	Paper	87.5	
Average of recognition rates	Scissor	88.9	89.1
	Stone	92.9	
	Paper	85.3	

and $\pm 30^{\circ}$ with a horizontal self-spin [as shown in Fig. 7(c)]. Hand gesture images of this database were taken in a quite free and unconstrained way. In total, there are 3,841 rock samples, 4,097 paper samples, and 4,242 scissor samples. Among them, only 150 images of each kind of hand gestures are used for training HTM and the remaining images are used for testing the trained HTM. Because this database possesses considerably large variations in both illumination and orientations, it makes the derived performance more representative to the practical applications.

Then, for each background scene 200 images of each kind of hand postures were randomly selected as testing samples, and therefore in total there are 1,600 testing images for each kind of hand postures. In Table 2, the corresponding recognition results are listed with respect to different hand postures and background scenes. Also, as shown in Fig. 8, the red block contains an entire arm which obviously is not suitable for recognition, the green block is the hand posture region after forearm being removed which is much easy to be recognized, and the recognition results are

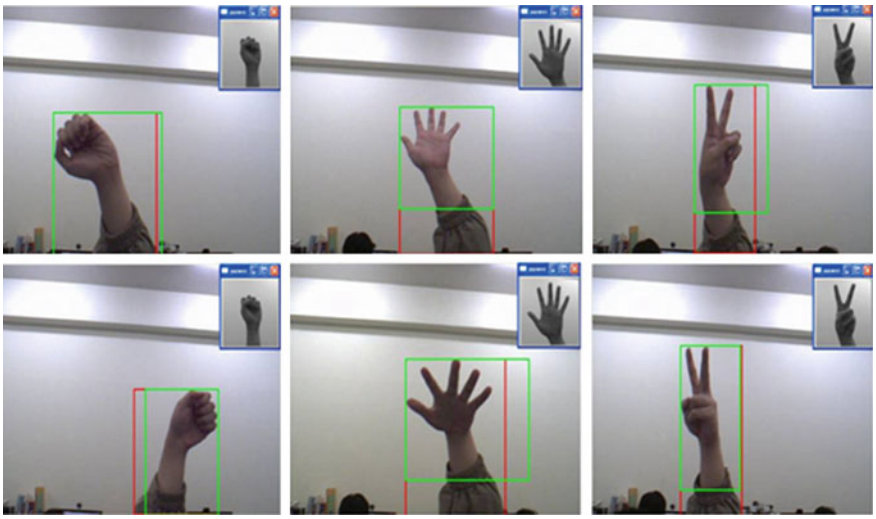


Fig. 8 Examples of different hand posture recognition (colour figure online)

Table 3 Confusion matrix of recognizing multi-angle hand postures

Input Recognition	Scissor	Stone	Paper
Scissor	88.9	3.3	7.8
Stone	5.5	92.9	1.6
Paper	11	3.7	85.3

shown in the upper-right part of each image. According to Table 2, the recognition rate reaches to 95 % for the plain-background taken images and an average 89.1 % for all the eight complex-background taken images. Table 3 is the corresponding confusion matrix which shows the correlation between input classes and recognized classes, and from this table it reveals that scissor and paper are much easier to be misrecognized compared with other pairs of hand gestures. There are two possible reasons for this phenomenon. First, when the angle of a hand gesture varies too large, its image would result in considerable fingers overlapping, and thus lead to a recognition failure. Second, because of the uneven illumination distribution, a binary hand image could become fractured and consequently leads to misrecognition.

We also performed an experiment by using the same 450 images (150 for each kind of hand gestures) to train a SVM-based hand gesture classifier and the same 4,800 testing images to test the trained classifier. The goal of this experiment is simply to compare the recognition abilities of both HTM and SVM, therefore their hand gesture preprocessing procedure on the training and testing images is the same. The detected hand gesture block images then are inputted to train and test the classifier. Table 4 lists the recognition result with an average recognition rate of 79.9 % which is

Table 4 Result of recognition with algorithm of SVM

	Scissor (%)	Stone (%)	Paper (%)
Training images	150	150	150
Testing images	1,600	1,600	1,600
Recognition rate (%)	78.3	86.1	75.3
Average (%)	79.9		

worse than that of HTM with 89.1 %. This experiment reveals clearly that HTM has much superior recognition ability than SVM. This superiority mostly comes from HTM has the ability to extract the spatially and temporally invariant feature from images.

6 Conclusion

This paper proposes an effective multi-angle hand posture recognition method. A useful hand posture region locating method is designed which involves skin color detection, foreground segmentation, edge detection, and forearm removal. As a result, the image of hand posture region is derived correctly and stably. Next, use HTM algorithm to process and generalize the continuous changing images to form invariant feature and to overcome the impact of angle changing. Finally, construct a high-efficient hand posture recognizer. Under the same test of multi-angle images, the proposed method performs better than both Adaboost and SVM algorithm.

Acknowledgments This work was supported by National Science Institute, Republic of China, under grants NSC NSC 101-2221-E-216-037-MY2.

References

1. Echtler F et al (2010) An LED-based multitouch sensor for LCD screens. In: Proceedings of the ACM tangible and embedded interaction conference, pp 227–230
2. Xie L, Liu Z (Apr. 2007) Realistic mouth-synching for speech-driven talking face using articulatory modelling. Proc IEEE Trans Multimedia 9(3):500–510
3. Cucchiara R et al (2003) Detecting moving objects. Ghosts, and shadows in video streams. Proc IEEE Trans Pattern Anal Mach Intell 25(10):1337–1342
4. McIvor AM (2000) Background subtraction techniques. In: Proceedings of the image and vision computing. Auckland, New Zealand
5. Murali S, Girisha R (2009) Segmentation of motion objects from surveillance video sequences using temporal differencing combined with multiple correlation. In: Proceedings of the IEEE international conference on advanced video and signal based surveillance, pp 472–477
6. Sun S et al (2000) Motion estimation based on optical flow with adaptive gradients. In: Proceedings of the IEEE international conference on image processing, vol 1, pp 852–855

7. Kim K et al (2005) Real-time foreground-foreground segmentation using codebook model. In: Proceedings of the real-time imaging, vol 11(3), pp 172–185
8. Chai D, Bouzerdoum A (2000) A Bayesian approach to skin color classification inYCbCr color space. In: Proceedings of the IEEE region ten conference, vol 2. Kuala Lumpur, Malaysia, pp 421–424
9. Hawkins J, George D (2006) Hierarchical temporal memory concepts, theory, and terminology, numenta, http://www.numenta.com/htm-overview/education/Numenta_HTM_Concepts.pdf
10. Hawkins J, Balkeslee S (2004) On Intelligence. Owl Books, New York
11. Vutsinas CN et al (2008) A neocortex model implementation on reconfigurable logic with streaming memory. In: Proceedings of the IEEE international symposium parallel and distributed processing, pp 1–8
12. Johnson SC (1967) Hierarchical clustering schemes. Proc Springer J Pyschometrika 32(3):241–254
13. Huang Y-S, Wang Y-J (2013) Multi-angle hand posture recognition based on hierarchical temporal memory.docx. In: Proceedings of the international multiconference of engineers and computer scientists 2013, Hong Kong, pp 70–75, 13–15 March 2013

Efficient Approach One-Versus-All Binary Tree for Multiclass SVM

Boutkhil Sidaoui and Kaddour Sadouni

Abstract In this paper we propose and examine the performance of a framework for solving multiclass problems with Support Vector Machine (SVM). Our methods based on the principle binary tree, leading to much faster convergence and compare it with very popular methods proposals in the literature, both in terms of computational needs for the feedforward phase and of classification accuracy. The proposed paradigm builds a binary tree for multiclass SVM, using the technical of portioning by criteria of natural classification: Separation and Homogeneity, with the aim of obtaining optimal tree. The main result, however, is the mapping of the multiclass problem to a several bi-classes sub-problem, in order to easing the resolution of the real and complex problems. Our approach is more accurate in the construction of the tree. Further, in the test phase OVA Tree Multiclass, due to its Log complexity, it is much faster than other methods in problems that have big class number. In this context, two corpus are used to evaluate our framework; TIMIT datasets for vowels classification and MNIST for recognition of handwritten digits. A recognition rate of 57 %, on the 20 vowels of TIMIT corpus and 97.73 % on MNIST datasets for 10 digits, was achieved. These results are comparable with the state of the arts. In addition, training time and number of support vectors, which determine the duration of the tests, are also reduced compared to other methods.

Keywords Binary tree · Classification · Homogeneity · Machine learning · Multiclass · Separation · Support vector machine.

B. Sidaoui (✉)
Mathematics and Computer Science Department, University of Tahar Moulay Saida,
BP 138 ENNASR Saida, 20000 Saida, Algeria
e-mail: b.sidaoui@gmail.com

K. Sadouni
Computer Science Department, University of Sciences and Technology USTO-MB,
BP 1505 Elmanouar Oran, 31000 Oran, Algeria
e-mail: kaddour_sadouni@hotmail.com

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 275, DOI: 10.1007/978-94-007-7684-5_15,

1 Introduction

Kernel Methods and particularly Support Vector Machines (SVMs) (SVM) introduced during the last decade in the context of statistical learning, received lately a lot of attention of the part of the community of research in algorithms of machine learning, for his solid theoretical foundations and their practical successes on concrete problems [1, 2, 10, 38, 39]. SVMs are arguably the single most important development in supervised classification of recent years. SVMs often achieve superior classification performance compared to other learning algorithms across most domains and tasks; they are fairly insensitive to the curse of dimensionality and are efficient enough to handle very large-scale classification in both sample and variables. SVMs [37, 39, 40], have been successfully used for the solution of a large class of machine learning tasks [15, 24] such as categorization, prediction, novelty detection, ranking and clustering. The success of Support Vector Machines on a number of high dimensional tasks has prompted a renewed interest in kernel methods. The time it takes to classify a new example using a trained support vector machine is proportional to the number of support vectors. While SVMs are a very robust and powerful technique for supervised classification, the large size and slow query time of a trained SVM is one hindrance to their practical application especially for Multiclass problems.

Where SVM was originally developed for binary problems and its extension to multi-class problems is not straightforward. How to effectively extend it for solving multiclass classification problem is still an on-going research issue. Several multiclass methods have been proposed which successfully help to alleviate this problem. The most popular and widely successful methods for applying SVMs to multiclass classification problems usually decompose the multi-class problems into several two-class problems. Many real–world applications consist of multiclass classification problems as: recognition of handwritten digits and automatic Speech Recognition (ASR). ASR has been the focus of both the machine learning and speech communities for the past few decades due to its importance in any conceivable man machine interface (MMI) [11]. ASR is considered as the most difficult and challenging problem to be solved for many years to come. The large variability and the richness of voice data represent a fertile field to evaluate the performance of recognition systems. Many approaches have been proposed towards the goal of improving the performance of ASR. Hidden Markov Models (HMM) are the most widely used recognizers for ASR, due to their ability of efficiently modeling the sequential nature of the speech frames [17, 32, 33]. Class of such methods [5, 7, 8, 16, 21, 23] focuses on new techniques for feature extraction. Others, keep the MFCC features along with derivative information, but use novel powerful discriminative methods [6, 9, 22], and [30]. Our contributions use the standard MFCC with powerful discriminative methods (SVM). In this paper, the task of vowels classification is used to evaluate our approach.

The remainder of the paper is organized as follows. We begin this paper with a brief introduction about multiclass problems and methods of machine learning. In Sect. 2, kernel methods SVM will be discussed for problems classifica-

tion. In the following section, we provide a brief state of art of the multi-class approaches, the most popular, and published in the literatures. In the next section, we introduce our efficient framework for multiclass SVM and we discuss the implementation of our architecture OVA Tree Multiclass. In Sect. 5, we give results of experiments on the vowels sets of TIMIT data base and MNIST corpus. Finally, the last section is devoted to conclusions and some remarks pertaining to future work.

2 Support Vector Machines

The main idea of binary SVMs is to implicitly map data to a higher dimensional space via a kernel function and then solve an optimization problem to identify the maximum-margin hyper-plane that separates training instances [39]. The separator is based on a set of boundary training instances (training examples). Kernels can be interpreted as dissimilarity measures of pairs of objects in the training set X. In standard SVM formulations, the optimal hypothesis sought is of the form (1).

$$\Phi(\xi) = \sum \alpha_i k(x, x_i) \tag{1}$$

where α_i are the components of the unique solution of a linearly constrained quadratic programming problem, whose size is equal to the number of training patterns. The solution vector obtained is generally sparse and the non zero α_i's are called support vectors (SV's). Clearly, the number of SV's determines the query time which is the time it takes to predict novel observations and subsequently, is critical for some real time applications such as speech recognition.

The training process is implicitly performed in a Reproducing Kernel Hilbert Space in which k(x;y) is the inner product of the images of two example x, y. Moreover, the optimal hypothesis can be expressed in terms of the kernel that can be defined for non Euclidean data such biological sequences, speech utterances etc. Popular positive kernels include the Linear (2), Polynomial (3), and Gaussian (4), kernels:

$$k(x_i, x_j) = x_i^T x_j \tag{2}$$

$$k(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0 \tag{3}$$

$$k(x_i, x_j) = \exp(-\gamma \left\| x_i - x_j \right\|^2), \gamma > 0 \tag{4}$$

2.1 SVM Formulation

Given training vectors $x_i \in \Re^n, i = 1, ..., m$, in two classes, and a vector $y \in \Re^m$ such that $y_i \in \{1, -1\}$, Support Vector Classifiers [10, 37, 39, 40] solve the following linearly constrained convex quadratic programming problem:

$$\begin{gathered} \text{maximize } W(\alpha) = \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j k\left(x_i, x_j\right) - \sum_{i=1}^{m} \alpha_i \\ \text{under the constraints} : \forall i, 0 \leq \alpha_{\mathrm{i}} \leq C \\ \sum_{i=1}^{m} \alpha_i y_i = 0 \end{gathered} \tag{5}$$

The optimal hypothesis is:

$$f(x) = \sum_{i=1}^{m} \alpha_i y_i k(x, x_i) + b \tag{6}$$

where the bias term b can be computed separately [24]. Clearly, the hypothesis f depends only on the non null coefficients α_i whose corresponding patterns are called Support vectors (SV). The quadratic programming (QP) objective function involves the problem Gram matrix K whose entries are the similarities $k(x_i, x_j)$ between the patterns x_i and x_j. It is important to note, on one hand, that the pattern input dimension d, in the above formulation, is implicit and does not affect to some extent the complexity of training, provided that the Gram matrix K can be efficiently computed for the learning task at hand. On the other hand, the patterns representation is not needed and only pair wise similarities between objects must be specified.

This feature makes SVM very attractive for high input dimensional recognition problems and for the ones where patterns can't be represented as fixed dimensional real vectors such as text, strings, DNA etc. For large scale corpora however, the quadratic programming problem becomes quickly computationally expensive, in terms of storage and CPU time. It is well known that general-purpose QP solvers scale with the cube of the problem dimension which is, in our case, the number of training patterns m. Specialized algorithms, typically based on gradient descent methods, achieve impressive gains in efficiency, but still become impractically slow for problems whose size exceeds 100,000 examples. Several attempts have been made to overcome this shortcoming by using heuristically based decomposition techniques such as Sequential Minimal Optimization [24] implemented in LibSVM package [14].

2.2 Extension Multiclass and Related Works

Many real–world applications consist of multiclass classification problems. Unfortunately, the Support Vector Machine (SVM), which is now considered one of the state–of–the–art algorithms in machine learning research, is intrinsically bi-class and its efficient extension to multiclass problems is still an ongoing research issue [12, 26, 29]. Several frameworks have been introduced to extend SVM to multiclass contexts and a detailed account of the literature is out of the scope of this paper.

The most common way to build a multiclass SVM is to combine several subproblems that involve only binary classification. This idea is used by various approaches as One-Versus-All (OVA), One-Versus-One (OVO) [29, 39] and Directed Acyclic Graph (DAGSVM) [26]: in the first case, K binary classifiers are constructed, where K is the number of classes. The kth classifier is trained by labeling all the examples in the kth class as positive and the remainder as negative. The final hypothesis is given by the formula:

$$f_{ova}(x) = \arg\max_{i=1,\ldots,k} (f_i(x)) \tag{7}$$

In the second case, OVO proceeds by training k(k-1)/2 binary classifiers corresponding to all the pairs of classes. The hypothesis consists of choosing either the class with most votes (voting). In the third case, each node of graph represents a binary classifier (DAGSVM) [26, 31].

There was debate on the efficiency of multiclass methods from statistical point of view clearly, voting and DAGSVM are cheaper to train in terms of memory and computer speed than OVA. Hsu and Lin [12] Investigated the performance of several SVM multi-class paradigms and found that the one-against-one achieved slightly better results on some small to medium size benchmark data sets. Furthermore, other interesting implementations of SVM multi-classes have been developed in recent years. Such as, two architectures proposed by [3] and [20], in the first Cha and Tappert built the decision tree by genetic algorithms where they are considered each tree as chromosome. While in the second architecture Madzarov and All propose a clustering method to construct the binary tree. Another approach in this direction has been presented by Lei and Venu in [18], they use a recursive method in to build the binary tree.

In term of complexity, OVA approach needs to create k binary classifiers; the training time is estimated empirically by a power law [25] stating that $T \approx \alpha M^2$ where M is the number of training samples and α is proportionality constant. According to this law, the estimated training time for OVA is:

$$T\_Time_{OVA} \approx k\alpha M^2 \tag{8}$$

Without loss of generality, let's suppose that each of the k classes has the same number of training samples. Therefore, each binary SVM-(OVO) approach requires $2M/k$ samples. Hence, the training time for OVO is:

$$T\_Time_{OVO} \approx \alpha \frac{k(k-1)}{2} \left(\frac{2M}{k}\right)^2 \approx 2\alpha M^2 \tag{9}$$

The training time for DAG is same as OVO. In the training phase, our framework One-Versus-All Tree multiclass has (k−1) binary classifiers (k is the number of classes). The random structure of the optimal tree complicates the calculation of the training time. However, an approximation is defined as: Let's assume that each of the k classes has the same number of training samples. The training time is summed over all k−1 nodes in the different $\lceil \log_2(k) \rceil$ levels of tree. In the *i*th level, there are at the most 2^{i-1} nodes and each node uses $M/2^{i-1}$ training samples. Hence, the total training time:

$$T\_Time_{SVMAG} \approx \sum_{i=1}^{\lceil \log_2(k) \rceil} \alpha 2^{i-1} \left(\frac{M}{2^{i-1}}\right)^2 \approx \alpha M^2 \tag{10}$$

It must be noted that the training time of our approach does not include the time to build the hierarchy structure (binary tree) of the k classes. In the testing phase, OVA require k binary SVM evaluations and OVO necessitate $\frac{k(k-1)}{2}$ binary SVM evaluations, while DAGSVM performs faster than OVO and OVA, since it requires only k−1 binary SVM evaluations. The two architectures proposed by [3, 20] and [34] and One-Versus-All Tree multiclass proposed in this paper are even fasters than DAGSVM because the depth of the binary tree is $\lceil \log_2(k) \rceil$. In addition, the total number of supports (SVs) vectors in all models will be smaller than the total number of SVs in the others (OVA, OVO and DAGSVM). Therefore it allows converge rapidly in the test phase.

The advantage of the approaches presented in [3, 18, 20] and the approach shown in this paper lie mainly in the test phase, because it uses only the models necessary for recognition. Which make the testing phase faster. However, in [3, 20] and [18] a problem of local minima is clearly. To avoid this problem, the work presented in [34] and this approach proposed in this work is to find the binary tree, using the similarity idea to find the right partitioning into two disjoint groups (one against rest), the partial optimization avoids falling into a local optimum, the details of the algorithm is discussed in the next section.

3 Binary Tree Multiclass SVM

This approach uses multiple SVMs set in a binary tree structure [6]. In each node of the tree, a binary SVM is trained using two classes. All samples in the node are assigned to the two subnodes derived from the current node. This step repeats at every node until each node contains only samples from one class. That said, until the leaves of the tree. The main problem that should be considered seriously here is how to construct the optimal tree? With the aim of partitioning correctly the training

samples in two groups, in each node of the tree. In this sense, many works have been proposed in Literatures as [3, 18, 20] and [34]. In this paper, we introduce a new multiclass method, which we call One-versus-All Tree (OVA Tree Multiclass) for solving multiclass problems with a Support Vector Machine [36].

3.1 One-Versus-All Binary Tree

In this paper, we exploit the main idea of the OVA approach and criteria's of similarity. Binary tree constructing is based on the research of the best hyperplan separating a class against the rest by using two criteria's of similarity. The aim is to construct a binary tree where each node represents a partition of two classes (one class against the remains). While the leaves represent the class's labels. But for a fixed number of classes, the optimal partition is not necessarily unique.

In this paper, we study several criteria specific to each family by systematically applying the same optimization strategy: a number of classes is fixed and, at each step, we construct a binary partition in each node of binary tree. At each level of the tree, we try to build a partition fixed number of classes (equal to 2) that optimizes a certain criterion. In this work, we begin with a partition has one group containing all classes. So we start with root (up) of tree, and we looks for the class of larger diameter (farthest compared to the other). We use the same procedure until no partitioning is possible. There are three families of natural classification criteria; it is separation, homogeneity and dispersion.

According to the first, a good score this well-separated classes, we seek to maximize the differences between classes, which are functions of distances between classes, i.e the smallest distances inter-classes and greater distances possible intra-class (in our case: two classes in each partition). This top-down approach consists to separate the class remains the most distant classes, we calculate a distance between all pair's (x_i, x_j) classes' prototypes and we deduce the class most distant from the remainder.

We obtain the first partition, root of the tree, contains all classes. This process is repeated only for the remainder; until there is only one class per group (no partitioning is possible). It is well known that for this type of optimization problem, the optimal partition is not necessarily unique, particularly if there are values equal distance, there may be several optimal trees and different partitions with the same optimal value. To avoid this bias and the problem of local minima, we calculate the total inertia of each tree (this is practically possible for a value not big of k,) defined by:

$$f_{\text{overallinertia}} = \sum_{i=1}^{L} inertie(i) \tag{11}$$

where L is the length of the tree. So, the global optimum is:

$$optimum_{Tree} = \arg\max\{f_{\text{overallinertia}}, i = 1..T\} \tag{12}$$

where: T is the number of all trees. According to the second, the classes are as concise as possible; it is desired to minimize the diameter, that is to say the maximum intra-distances classes. In this case, we generate partitions by a clustering technique bottom upwards. We are constructing binary partitions, which one of the two classes containing only one element, by grouping the closest classes. This bottom-up approach is similar to a hierarchical classification method. Initially, we compute a similarity matrix of all pair's (x_i, x_j) prototypes of classes and selecting the smallest similarity, which gives us the partition of the last level of tree (the smallest partition). This process is repeated, updating the center of gravity at each level, to the root of tree. To prevent the problem of local minima, we calculate the total inertia defined by the formula (12) of each tree.

According to the third, we minimize a function of inertia, the sum of squared deviations in a center, whether real or virtual.

The OVA Tree Multiclass method that we propose is based on recursively partitioning the classes in two disjoint groups (one contains one class only) in every node of the binary tree, and training a SVM that will decide in which of the groups the incoming unknown sample should be assigned. In the general case, the number of partitions into two parts (groups) of a set of k elements is given by the following formula [13]:

$$N\_partions_{k,2} = \sum_{i=0}^{2} (-1)^i \frac{(2-i)^k}{i!(2-i)!} \tag{13}$$

Corresponding construction of the binary tree, two cases can be expected: the number k is small in this case; we calculate all possible partitions and then deduce the optimal partition. Where the number k is greater than 6 (k > 6), we determine the optimal partition by Methods of optimization, because it is impossible to cover all possible partitions.

In this manuscript, we are interested at the first case. The OVA Tree two Algorithms are begun by a set X of training samples labeled by $y_i \in \{c_1, c_2, ..., c_k\}$, each OVA Tree Multiclass is summarized in two phases: Calculate k gravity centers and find the right partition of k gravity centers into two groups, by using separation, homogeneity criteria's. For that, a function of overall inertia is calculated for each binary tree. Indeed, this function of overall inertia is sum of inertia's of various partitions of tree. The Fig. 1 illustrates an example of binary tree for 7 classes [36].

3.2 Implementation of Tree SVM

The two optimal trees are implemented to obtain a multiclass approaches. For each one, we take advantage of both the efficient computation of the tree architecture and the high classification accuracy of SVMs Utilizing this architecture, k − 1 SVMs

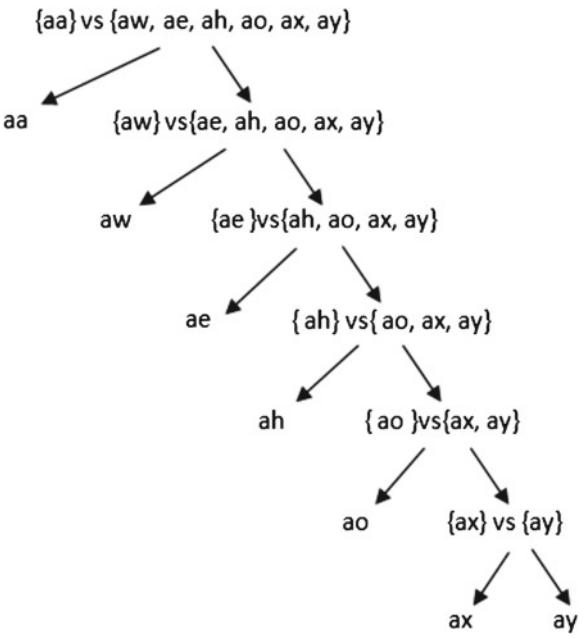


Fig. 1 Binary tree for 7 vowels

needed to be trained for k class problem. This binary tree is used to train a SVM classifier in the root node of the tree, using the samples of the first group as positive examples and the samples of the second group as negative examples.

The classes from the first clustering group are being assigned to the left subtree, while the classes of the second clustering group are being assigned to the right subtree. The process continues recursively (dividing each of the groups into two subgroups applying the procedure explained above), until there is only one class per group which defines a leaf in the decision tree.

The recognition of each test sample starts at the root of the tree. At each node of the binary tree a decision is being made about the assignment of the input pattern into one of the two possible groups represented by transferring the pattern to the left or to the right sub-tree. Each of these groups may contain multiple classes. This is repeated recursively downward the tree until the sample reaches a leaf node that represents the class it has been assigned to.

4 Experimental and Results

In this paper, we performed our experiments on two corpuses: TIMIT corpus [35] and MNIST corpus [27]. For TIMIT datasets, 20 vowels used in [28] are selected to evaluate our approach. The 20 vowels set are: {aa, aw, ae, ah, ao, ax, ay, axr, ax-h, uw,

ux, uh, oy, ow, ix, ih, iy, eh, ey, er}. The 10 classes (handwritten digits) of MNIST corpus are: {0, 1, 2, 3, 4, 5, 6, 7, 8, and 9}.

In all the experiments reported below, we performed cross validation for tuning SVM hyper parameters. The GNU SVM light [14] implementation is used for our OVA Tree Multiclass Machine used in this paper, and LibSVM [4] for SVC [19] software's to compare our results. All the experiments were run on standard Intel (R) core ^{TM} 2 Duo CPU 2.00 GHZ with 2.99 Go memory running the Windows XP operating system.

The following tables, Tables 1 and 2, summarize the preliminary results, for the classification supervised of 20 vowels, and Table 3 present results for 10 handwritten digits of MNIST listed above.

Table 1 Results obtained by SVM (OVO) for 20 vowels

	T	C	g	Test (%)	CPU time (s)
	2	2000	0.0005	59.64	510.89
	2	5000	0.0005	59.83	670.15
SVM (OVO)	2	10000	0.0005	60.14	504.00
	2	1000	0.005	58.11	945.65
	2	200	0.005	59.82	938.34

Table 2 Results obtained by OVA tree multiclass for 20 vowels of TIMIT data sets

	T	C	g	Test (%)	CPU time (s)
Separation	0	0	–	40.80	120
	2	100	0.005	56.93	504
Homogeneity	2	0.0075	10	54.95	537
	2	0.0075	0	48.73	609
	2	0.005	10	53.62	572
	2	0.005	100	56.92	543

Table 3 Results obtained by OVA tree multiclass for 10 digits of MNIST data sets

	T	C	d	Test (%)	CPU time (s)
Separation	0	0	–	91.69	300
	1	0	2	97.08	480
	1	10	2	97.73	420
	1	100	2	97.73	421
Homogeneity	0	0	–	91.75	264
	1	0	2	97.09	467
	1	100	2	97.72	418
	1	1000	2	97.72	420

5 Conclusion

We introduced and implemented an efficient framework for SVM multiclass using separation and homogeneity criteria's. The Binary Tree of support vector machine (SVM) multiclass paradigm was shown through extensive experiments to achieve state of the art results in terms of accuracy. The preliminary experiments of our binary architecture named OVA Tree Multiclass [36] indicate that the results are comparable to those of other methods, particularly [29] and [34] who used the same corpus. Nevertheless, parameters SVM optimization can improve the performance of our contribution.

However, we believe that in order to improve automatic pattern recognition technology, more research efforts must take advantage of the solid mathematical basis and the power of SVM binary, and should be invested in developing general purpose efficient machine learning paradigms capable of handling large scale multi-class problems.

References

1. Boser B, Guyon IM, Vapnik V (1992) A training algorithm for optimal margin classifiers. 5th annual workshop on computational learning theory. ACM Press, Pittsburgh
2. Boser B, Guyon I, Vapnik V (1992) A training algorithm for optimal margin classifiers. In 5th annual workshop on computational learning theory. Pittsburgh, pp 144–152
3. Cha SH, Tappert C (2009) A genetic algorithm for constructing compact binary decision trees. J Pattern Recogn Res 4(1):1–13
4. Chang C-C, Lin C-J (2013) LIBSVM toolkit: a library for support vector machines. Software available at: http://www.csie.ntu.edu.tw/cjlin/libsvm/
5. Erdogan H (2005) Regularizing linear discriminant analysis for speech recognition. Interspeech'2005, Lisbon, Portugal (4–8 Sep 2005)
6. Fei B, Liu J (2006) Binary tree of SVM: a new fast multiclass training and classification algorithm. IEEE Trans Neural Netw 17(3):696–704
7. Furui S (1986) Speaker-independent isolated word recognition using dynamic features of speech spectrum. IEEE Trans Acoust Speech Signal Process 34:52–59
8. GF Choueiter, JR Glass (2005) A wavelet and filter bank framework for phonetic classification. ICASSP
9. Grave A, Schmidhuber J (2005) Framewise phoneme classification with bidirectional LSTM networks. In: Proceedings of IJCNN, vol 4, pp 2047–2052
10. Guyon I, Boser B, Vapnik V (1993) Automatic capacity tuning of very large VC-dimension classifiers. Adv Neural Inf Process Sys 5:147
11. Hong D (2007) Speech recognition technology: moving beyond customer service. Comput Bus (Online 1st Mar 2007)
12. Hsu CW, Lin CJ (2002) A comparison of methods for multiclass support vector machines. IEEE Trans Neural Netw 13(2):415–425
13. Jain AK, Dubes R (1988) Algorithms for clustering data. Prentice Hall, NJ
14. Joachims T (2001) Making large-scale SVM learning practical. Software available at: http://svmlight.joachims.org/
15. Joachims T (1998) Making large-scale support vector machine learning practical. In: Schölkopf B, Burges C, Smola A (eds) Advances in kernel methods. MIT Press, Cambridge

16. Kamal M, Mark H-J (2003) Non-linear independent component analysis for speech recognition. International conference on computer communication and control technologies, Orlando
17. Lee K-F, Hon H-W (1989) Speaker-independent phone recognition using Hidden Markov Models. IEEE Trans Acoust Speech Signal Process 37(11):1641–1648
18. Lei H, Govindaraju V (2005) Half-against-half multi-class support vector machines. In: Oza NC, Polikar R, Kittler J, Roli F (eds) Multiple classifier system, vol 3541. Springer, Berlin, pp 156–164
19. LibCVM toolkit of the improved core vector machine (CVM), which are fast support vector machine (SVM) training algorithms using core-set approximation on very large scale data sets available at: http://c2inet.sce.ntu.edu.sg/ivor/cvm.html
20. Madzarov G, Gjorgjevikj D, Chorbev I (2009) A multi-class SVM classifier utilizing binary decision tree. Informatica 33:233–241
21. Moreno P (1999) On the use of support vector machines for phonetic classification. In: Proceedings of ICCASP, vol 2. Phoenix, AZ, pp 585–588
22. Morris J, Fosler-Lussier E (2006) Discriminative phonetic recognition with conditional random fields. HLTNAACL
23. Naomi H, Saeed V, Paul MC (1998) A novel model for phoneme recognition using phonetically derived features. In: Proceeding EUSIPCO
24. Osuna E et al (1997) Training support vector machines, an application to face detection. Proceedings IEEE computer society conference on computer vision and pattern recognition, In, pp 130–136
25. Platt J (1999) Fast training of support vector machines using sequential minimal optimization. In: Schölkopf B, Burges CJC, Smola AJ (eds) Advances in kernel methods: support vector learning
26. Platt JC, Cristianini N, Shawe-Taylor J (2000) Large margin DAGs for multiclass classification. Adv Neural Inf Process Sys 12:547–443
27. Platt JC, Cristianini N, Shawe-Taylor J (2000) Large margin DAGs for multiclass classification. Adv Neural Inf Process Sys 12:547–553
28. Rifkin R et al (2007) Noise robust phonetic classification with linear regularized least squares and second order featues. ICASSP
29. Ryan R, Klautau A (2004) In defense of one-vs-all classification. J Mach Learn Res 5:101–141
30. Salomon J, King S, Osborne M (2002) Framewise phone classification using support vector machines. ICSLP
31. Schölkopf B, Smola AJ (2002) Learning with kernels. MIT Press, Cambridge
32. Sha F et al. (2007) Comparaison of large margin training to other discriminative methods for phonetic recognition by hidden markov models. IEEE international conference on acoustics, speech and signal processing, vol 4. Honolulu, USA
33. Sha F, Saul LK (2006) Large margin hidden markov models for automatic speech recognition. NIPS
34. Sidaoui B, Sadouni K (2013) Approach multiclass SVM utilizing genetic algorithms. IMECS 2013 conference, Hong Kong
35. Slaney M (1998) Auditory toolbox version 2. Tech. Report#010. Internal Research Corporation
36. The MNIST database of handwritten digits has a training set of 60,000 examples, and a test set of 10,000 examples. The digits have been size-normalized and centred in a fixed-size image available at: http://yann.lecun.com/exdb/mnist/
37. Vapnik V (1982) Estimation of dependences based on empirical data. Nauka, Moscow. (English translation, Springer, 1979, NY)
38. Vapnik V (2000) The nature of statistical learning theory. Springer, NY
39. Vapnik V (1998) Statistical learning theory. Wiley, NY
40. Vapnik V, Chervonenkis AJ (1974) Theory of pattern recognition. Nauka, Moscow. (Theorie der Zeichenerkennung, 1979, Akademie-Verlag, Berlin)

Parametric Control of National Economy's Growth Based on Regional Computable General Equilibrium Model

Abdykappar Ashimov, Yuriy Borovskiy, Bahyt Sultanov, Nikolay Borovskiy, Rakhman Alshanov and Bakytzhan Aisakova

Abstract Application efficiency of the proposed method of parametric identification for large mathematical models in case of regional non-autonomous computable general equilibrium model (CGE model) is illustrated in the paper. The problem of control of discrete non-autonomous dynamic system is formulated. Theorems about sufficient conditions for its solution and continuous dependence of corresponding optimal values of criterion on uncontrollable functions are presented. Based on the model we conduct analysis of economic growth sources and application efficiency of parametric control theory for conducting state economic policy, directed to economic growth and reduction of disproportion in regional economic development.

Keywords CGE model · Discrete non-autonomous dynamic system · Economic growth · Economic growth sources · Parametric control · Parametric identification

A. Ashimov (✉) · Y. Borovskiy · B. Sultanov · N. Borovskiy · R. Alshanov · B. Aisakova
Kazakh National Technical University, 22 Satpayev Street, Almaty 050013, Kazakhstan
e-mail: ashimov37@mail.ru

Y. Borovskiy
e-mail: yuborovskiy@gmail.com

B. Sultanov
e-mail: sultanov_bt@pochta.ru

N. Borovskiy
e-mail: nborowski86@gmail.com

R. Alshanov
e-mail: alshanovra@yandex.ru

B. Aisakova
e-mail: aisakova_b@mail.ru

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 275, DOI: 10.1007/978-94-007-7684-5_16,

1 Introduction

As it is known, there is a broad consensus among macroeconomists on the application of mathematical models for macroeconomic analysis [1, 2] and solution of problems of economic growth [3, 4]. One of the topical problems of regulation of economic growth is the problem of ensuring sustainable growth of the national economy, taking into account the requirements of smoothing the levels of socio-economic development of certain regions of the country.

Reference [5] describes the CGE model, "Russia: Center-Federal Districts", based on which the scenario approach is considered for assessing the feasibility of smoothing regional disparities in the Russian Federation after solving the calibration problem. However issues of analysis of economic growth sources for each region and problems of finding the optimal rules for state economic policy are not addressed in [5].

A parametric control theory for macroeconomic analysis and evaluating optimal values of economic state policy tools based on the set classes of macroeconomic models are proposed in [6, 7]. In this paper some new elements of this theory are given: theorem about sufficient conditions for existence of solution to variational calculus problem on synthesis of optimal law of parametric control for discrete non-autonomous dynamic system and theorem about sufficient conditions for continuous dependence of corresponding optimal criterion values on uncontrollable functions.

The present work contains illustrations of some statements of the parametric control theory on example of a discrete non-autonomous large-scale CGE model "Center-Regions":

- Parametric identification of the model based on statistical data of the Republic of Kazakhstan economy,
- Analysis of sources of regional economic growth based on production functions of estimated model and
- Synthesis of parametric control optimal laws for problems in the sphere of economic growth and reduction in disproportion of regional socio-economic development.

A version of this paper was presented at the International MultiConference of Engineers and Computer Scientists 2013 and has been published in conference proceedings [8].

2 Some Elements of the Parametric Control Theory

We consider the discrete controllable system of the following type:

$$x(t+1) = f(x(t), u(t), a(t)), \quad t = 0, 1, \ldots, n-1; \tag{1}$$

$$x(0) = x_0. \tag{2}$$

Here t—time, which takes nonnegative integer values; $x = x(t) = (x^1(t), \ldots, x^m(t))$—the vector-function of the system status of a discrete argument; $u = u(t) = (u^1(t), \ldots, u^q(t))$—the regulation, the vector-function of a discrete argument; $a = a(t) = (a^1(t), \ldots, a^s(t))$—known vector-function of a discrete argument; $x_0 = (x_0^1, \ldots, x_0^m)$—initial system status, the known vector; f—known vector-function of its arguments.

The method to choose optimal values of economic instruments is related to the following model, represented by

- the optimality criterion (where F—known function)

$$K = \sum_{t=1}^{n} F[t, x(t)] \to \max(\min); \tag{3}$$

- phase constraints imposed on the system's solution of the following type (where $X(t)$ is a given set):

$$x(t) \in X(t), \quad t = 1, \ldots, n; \tag{4}$$

- by explicit constraints imposed on regulation (where $U(t)$—given set):

$$u(t) \in U(t), \quad t = 0, \ldots, n-1. \tag{5}$$

Based on the relations Eqs. (1)–(5) we obtain the following variational problem, called the problem of variational calculus for the synthesis of optimal laws of parametrical regulation for a discrete system.

Problem 1 *Given the known function a, find a regulation u, which satisfies the condition* Eq. (5)*, so that corresponding to it solution of a dynamic system* Eqs. (1), (2) *satisfies the condition* Eq. (4) *and provides maximum (minimum) for the functional* Eq. (3).

Let V_a be the set of allowed pairs "state-regulation" of the considered system given the known function a, i.e. such pairs of vector-functions (x, u), that satisfy the relations Eqs. (1), (2), (4), (5); $X = \bigcup_{t=1}^{n} X(t)$, $U = \bigcup_{t=0}^{n-1} U(t)$.

The proofs of the next two theorems are based on application of continuous functions' properties, and particularly, on application of the properties of the functions that are continuous on the compact.

Theorem 1 *Assume that given the known function a the set V_a is non-empty, the sets $X(t)$ and $U(t-1)$ are closed and limited for all $t = 1, \ldots, n$, the function f is continuous with respect to the first two arguments on the set $X \times U$, and the function F is continuous with respect to the second argument on the set X. Then the Problem 1 has a solution.*

Next we consider uncontrollable functions a in Eq. (1) as the elements of the Euclidian space R^{sm}.

Theorem 2 *Assume that the conditions of the Theorem 1 are hold for any values of* $a \in A$ *(where A—some open set in Euclidian space* R^s*), the function* f *is continuous with respect to the third argument in A and satisfies the Lipschits condition with respect to the first argument in X uniformly with respect to the second and the third argument in* $U \times A$*. Then the optimal value of the criterion for the Problem1 continuously depends on the uncontrollable function a which takes it values in A.*

Efficiency of the developed theory of parametrical regulation is illustrated below on the subclasses of CGE models.

3 Representation of CGE Models

Considered CGE model "Center-Regions" is presented in general view with the help of the following system of relations [7].

1. Subsystem of recurrent relations, connecting the values of endogenous variables for the two consecutive years:

$$x_1(t+1) = f_1(x_1(t), x_2(t), x_3(t), u(t), a(t)). \tag{6}$$

Here $t = 0, 1, \ldots, n-1$—number of a year, discrete time; $x(t) = (x_1(t), x_2(t), x_3(t)) \in R^m$—vector of endogenous variables of the system;

$$x_i(t) \in X_i(t) \subset R^{m_i}, \quad i = 1, 2, 3. \tag{7}$$

Here $x_1(t)$ include the values of capital stocks of a regions' economic agents, budgets of economic agents and other; $x_2(t)$ include demand and supply values of economic agents of regions in different markets and other; $x_3(t)$—different types of market prices and budget shares in markets with fixed prices for different economic agents; $m_1 + m_2 + m_3 = m$; $u(t) \in U(t) \subset R^q$—vector-function of controllable parameters. Values of the coordinates of this vector correspond to different tools of state economic policy, for example, such as shares of state budget and shares of states budgets, different tax rates and other; $a(t) \in A \subset R^s$—vector-function of uncontrollable parameters (factors). Values of the coordinates of this vector characterize different dependent on time external and internal socio-economic factors: prices for imported and exported goods, population size of the county, parameters of production functions and other; $X_1(t), X_2(t), X_3(t), U(t)$—compact sets with non-empty interiors; $X_i = \bigcup_{t=1}^{n} X_i(t), i = 1, 2, 3$; $X = \bigcup_{i=1}^{3} X_i$; $U = \bigcup_{t=0}^{n-1} U(t)$, A—open connected set; f_1: $X \times U \times A \to R^{m_1}$—continuous mapping.
2. A subsystem of algebraic equations characterizing behavior and interaction of agents in different markets during the selected year, these equations allow

expression of variables $x_2(t)$ by exogenous parameters and other endogenous variables:

$$x_2(t) = f_2(x_1(t), x_3(t), u(t), a(t)). \quad (8)$$

Here $f_2 : X_1 \times X_3 \times U \times A \to R^{m_2}$—continuous mapping.

3. Subsystem of recurrent relations for iterative calculation of market prices equilibrium values in different markets and budget shares in markets with state prices for different economic agents:

$$x_3(t)[Q+1] = f_3(x_2(t)[Q], x_3(t)[Q], L, u(t), a(t)). \quad (9)$$

Here $Q = 0, 1, \ldots$ —number of iteration; L—set of positive numbers (adjusted coefficients of iteration, when their values decrease economic system reaches equilibrium faster, but the risk that prices go to negative domain increases; $f_3 : X_2 \times X_3 \times (0, +\infty)^{m_3} \times U \times A \to R^{m_2}$—continuous mapping (contracting at fixed t; $x_1(t) \in X_1(t)$; $u(t) \in U(t)$; $a(t) \in A$ and some fixed L. In this case f_3 mapping has the unique fixed point, where the iteration process Eqs. (8), (9) converges.

CGE model Eqs. (6), (8), (9) at fixed values of the functions $u(t)$ and $a(t)$ at each point of time t defines values of endogenous variables $x(t)$, corresponding to equilibrium of demand and supply prices in markets of goods and services of agents within the framework of the following algorithm.

1. We assume that $t = 0$ and initial values of the variables $x_1(0)$ are set.
2. For the current t we set initial values for variables $x_3(0)[0]$ in different markets and for different agents; with the help of (8) we compute values $x_2(t)[0] = f_2(x_1(t), x_3(t)[0], u(t), a(t))$, (initial demand and supply values of agents in markets of goods and services).
3. For the current t the process of iteration Eqs. (3), (4) is run. Meanwhile for each value of Q current demand and supply values are found with the help of (8): $x_2(t)[Q] = f_2(x_1(t), x_3(t)[Q], u(t), a(t))$, through refinement of market prices and budget shares of economic agents.

   A condition for iteration process to stop is the equality of supply and demand in different markets (accurate within 0.01 %). As a result we obtain equilibrium values of market prices for each market and budget shares in markets with state prices for different economic agents. Q index is omitted for such equilibrium values of endogenous variables.
4. On the next step values of variables $x_1(t+1)$ are found with the help of obtained equilibrium solution for time t applying differential equations Eq. (6). A value of t increases by 1. Jump to step 2.

The number of steps 2, 3, 4 iterations is defined according to problems of parametric identification, forecasting and control for the time intervals selected in advance.

The considered CGE model can be presented as continuous mapping $f: X \times U \times A \to R^m$, giving transformation of values of the system's endogenous variables

for a null year to the corresponding values of the consecutive year according to the algorithm stated above. Here the compacts $X(t) = X_1(t) \times X_2(t) \times X_3(t)$, giving a compact X in the space of endogenous variables are determined by set of possible values of x_1 variable and corresponding equilibrium values of variables x_2 and x_3, estimated by the ratios Eqs. (8) and (9).

We assume that for selected point $x_1(0) \in \mathrm{Int}(X_1)$ and corresponding point $x(0) = (x_1(0), x_2(0), x_3(0))$, computed with the help of Eqs. (8) and (9), an inclusion $x(t) = f^t(x(0)) \in \mathrm{Int}(X(t))$ is true at some fixed $u(t) \in \mathrm{Int}(U(t))$, $a(t) \in A$ for $t = 0, \ldots, n$. (n—fixed non-negative integer number). This mapping f defines a discrete dynamic system in the set X, on the trajectory of which the following initial condition is imposed:

$$\left\{f^t, t = 0, 1, \ldots\right\}, \quad x|_{t=0} = x_0. \tag{10}$$

Based on this description below we consider a particular CGE model "Center-Regions".

4 Brief Description and Parametric Identification of the CGE Model "Center-Regions"

The considered model on statistical data for the Republic of Kazakhstan and its 16 regions is presented by the following 66 economic agents (sectors):

- 16 legal and 16 shadow sectors of economy of all regions;
- 16 aggregate consumers of all regions;
- 16 regional authorities;
- Government, represented by central government and also by non-budget funds;
- Banking sector, involving Central bank and commercial banks.

Here the first 32 economic sectors are producing agents.

The considered model is presented within the framework of general expressions of ratios Eqs. (6), (8), (9) respectively by $m_1 = 240$, $m_2 = 4{,}554$, $m_3 = 160$ expressions, with the help of which values of its 4,954 endogenous variables are calculated. This model also contains 39,122 estimated exogenous parameters.

The problem of parametric identification of the researched macroeconomic model is to find estimates of unknown values of its parameters at which a minimum value of the objective function is reached. This objective function characterizes deviations of values of the model's output variables from corresponding observed values (known statistical data for the time interval $t = t_1, t_1+1, \ldots, t_2$). This problem is to find minimum value of the function of several variables (parameters) at some closed set in the domain D of the Euclidian space with constraints of type Eq. (7), imposed on values of endogenous variables. Standard methods of finding the function's minimums are often inefficient due to existence of multiple local minimums of an objective function

in case of high dimensionality of the region of possible arguments' values. Below we present an algorithm, that considers peculiarities of the parametric identification problem of macroeconomic models and that allows to avoid the problem of "local extremums".

The domain of type $D = \prod_{i=1}^{(q+s)(t_2-t_1+1)+m_1}[a^i, b^i]$, where $[a^i, b^i]$—possible values interval of the parameter p^i; $i = 1, \ldots, (q+s)(t_2-t_1+1)+m_1$, is considered as a domain $D \subset \prod_{t=t_1}^{t_2}[U(t) \times A(t)] \times X_1(t_1)$ for estimating possible values of exogenous parameters (values of exogenous functions $u(t), a(t)$ and initial conditions of dynamic equations Eq. (6)). Meanwhile, parameter values, for which we have observed values, are searched at intervals $[a^i, b^i]$ with centers at corresponding observed values (in case if there is one such value) or at some intervals, covering observed values (in case if there are several such values). Other intervals $[a^i, b^i]$ for parameter search have been selected with the help of indirect estimations of their possible values. To find minimal values of continuous multivariable function $K\colon D \to R$ with additional constraints of type Eq. (7) at computational experiments the Nelder-Mead algorithm of directed search has been applied. Using this algorithm for the starting point $p_1 \in D$ can be interpreted as converging to the point (of local minimum) $p_0 = \arg\min_{D,(7)} K$ of sequence $\{p_1, p_2, p_3, \ldots\}$, where $K(p_{j+1}) \le K(p_j)$, $p_j \in D$, $j = 1, 2, \ldots$.

To solve the problem of parametric identification of the considered CGE model two criterions (auxiliary and main respectively) are proposed:

$$K_A(p) = \sqrt{\frac{1}{n_\alpha(t_2-t_1+1)}\sum_{t=t_1}^{t_2}\sum_{i=1}^{n_A}\alpha_i\left(\frac{y^i(t)-y^{i*}(t)}{y^{i*}(t)}\right)^2},$$

$$K_B(p) = \sqrt{\frac{1}{n_\beta(t_2-t_1+1)}\sum_{t=t_1}^{t_2}\sum_{i=1}^{n_B}\beta_i\left(\frac{y^i(t)-y^{i*}(t)}{y^{i*}(t)}\right)^2}. \quad (11)$$

Here $\{t_1, \ldots, t_2\}$—identification time interval; $y^i(t)$, $y^{i*}(t)$—estimated and observed values of output variables of the model respectively; $n_B > n_A$; $\alpha_i > 0$ and $\beta_i > 0$—some weight coefficients, their values are determined during the process of solving the parametric identification problem for the dynamic system; $\sum_{i=1}^{n_A}\alpha_i = n_\alpha$, $\sum_{i=1}^{n_B}\beta_i = n_\beta$.

Algorithm of solving the problem of parametric identification of the model is selected with the help of following steps.

1. Problems A and B are solved simultaneously for a vector of initial values of parameters $p_1 \in D$. As a result points p_{A0} and p_{B0} of minimums criteria K_A and K_B are found respectively.
2. If $K_B(p_{B0}) < \varepsilon$ is true for some sufficiently small value ε, then the problem of parametric identification of the model Eqs. (6), (8), (9) is solved.
3. Otherwise problem A is solved applying point p_{B0} as initial point p_1, problem B is solved applying point p_{A0} as p_1. Jump to step 2.

Table 1 Observed, calculated values of output variables of the model and corresponding deviations

Indicator	Year					
	2000	2001	2002	2003	2004	2005
$Y^*(t)$	5.30	6.26	6.33	6.87	7.84	8.44
$Y(t)$	5.16	6.42	6.44	6.98	7.81	8.23
$\Delta Y(t)$	−2.90	−1.00	−3.00	0.10	0.60	2.40
$Y_g^*(t)$	2.31	2.62	2.88	3.18	3.52	3.86
$Y_g(t)$	2.25	2.58	2.93	3.21	3.47	3.81
$\Delta Y_g(t)$	0.60	2.30	0.10	−1.50	0.50	2.30
	2006	2007	2008	2009	2010	2011
$Y^*(t)$	9.04	9.87	9.92	1.04	1.09	1.14
$Y(t)$	8.79	9.65	9.85	1.05	1.13	1.15
$\Delta Y(t)$	1.50	−2.80	0.40	0.30	−0.10	0.20
$Y_g^*(t)$	4.72	5.14	5.30	5.36	5.50	5.65
$Y_g(t)$	4.70	5.19	5.17	5.52	5.36	5.58
$\Delta Y_g(t)$	−2.80	1.90	2.50	−1.90	−2.80	−2.20

Quite large number of iterations of steps 1, 2, 3 provides an opportunity for searched values of parameters to exit from neighborhood points of nonglobal minimums of one criterion with the help of another criterion, thus solve the problem of parametric identification.

As a result of joint solution of problems A and B according to the specified algorithm applying statistical data on evolution of the Republic of Kazakhstan economy we have obtained values $K_A = 0.034$ and $K_B = 0.047$. Relative magnitude of deviations of parameter calculated values used in the main criterion from corresponding observed values is less than 4.7 %.

Further calculation of the estimated model on the parametric identification interval and outside the period of parametric identification (forecasted estimation) with the help of extrapolated values $u(t)$, $a(t)$ is called a basic calculation.

Results of calculation and of retrospective basic calculation of the model for 2011, partially presented in Table 1, demonstrate estimated, observed values and deviations of estimated values of main output variables of the model from corresponding observed values. Here the time interval 2000–2010 corresponds to the period of parametric identification of the model; 2011—is a period of retroforecasting; $Y(t)$—total gross output of a legal sector ($\times 10^{12}$ tenge, in prices of 2000; tenge—national currency of Kazakhstan); $Y_g(t)$—GDP of a state ($\times 10^{12}$ tenge, in prices of 2000); a sign "*" corresponds to observed values, a sign "Δ" corresponds to deviations (in percentage) of estimated values from corresponding observed values.

5 Analysis of Regional Economic Growth Sources

In this section we make analysis of economic growth sources of legal sectors of the Republic of Kazakhstan regions on the basis of the CGE model "Center-Regions", which exogenous functions and parameters have been evaluated as a result of solving the parametric identification problem of the model based on the statistical data of the socio-economic development of the Republic of Kazakhstan for 2000–2010.

The researched model uses the following expressions of multiplicative production functions of legal sectors of 16 regions:

$$Y_i(t+1) = A_i^{\mathrm{r}}(t) \times \left[\sum_{j=1}^{16}(D_i^{\mathrm{z}j1}(t) + D_i^{\mathrm{z}j2}(t))\right]^{A_i^{\mathrm{z}}} \times \exp[A_i^{\mathrm{zIm}} \times D_i^{\mathrm{zIm}}(t)]$$
$$\times \left[\frac{K_i(t) + K_i(t+1)}{2}\right]^{A_i^{\mathrm{k}}} \times \quad \exp[A_i^{\mathrm{l}} \times D_i^{\mathrm{l}}(t)]. \tag{12}$$

Here t—time in years; Y_i—real output of a legal sector in region i (i—number of a region, $i = 1, \ldots, 16$, see Table 2); $D_i^{\mathrm{z}j1}$—real demand of a i region's legal sector for intermediate goods, produced by a legal sector of a region j; $D_i^{\mathrm{z}j2}$—real demand of a i region's legal sector for intermediate goods, produced by a shadow sector of a region j; D_i^{zIm}—real demand of a i region's legal sector for imported intermediate goods; D_i^{l}—demand of a i region's legal sector for labor; K_i—real capital funds of a i region's legal sector; A_i^{r}, A_i^{z}, A_i^{zIm}, A_i^{k}, A_i^{l}—known exogenous functions.

Let us evaluate influence of growth rate of this function's arguments on growth rates of output $Y_i(t+1)$ of legal sector in a region in the assumption of constant exogenous functions A_i^{z}, A_i^{zIm}, A_i^{k}, A_i^{l}. Such assumption is used at extrapolation of these functions for the period of forecasting: 2012–2015.

Having taking the logarithms of both sides Eq. (12), then having found the total increment of the function and having dropped the high-order infinitesimals we obtain the following estimate for growth rate $y_i = \frac{\Delta Y_i}{Y_i}$ of real output of a legal sector in a region i depending on growth rates of endogenous arguments ($D_i^{\mathrm{z}} = \sum_{j=1}^{16}\left(D_i^{\mathrm{z}j1} + D_i^{\mathrm{z}j2}\right)$, D_i^{zIm}, $K_i^{\mathrm{m}}(t) = \frac{K_i(t)+K_i(t+1)}{2}$, D_i^{l}) of production functions and an exogenous coefficient of technical progress (A_i^{r}).

$$y_i = \frac{\Delta A_i^{\mathrm{r}}}{A_i^{\mathrm{r}}} + A_i^{\mathrm{z}}\frac{\Delta D_i^{\mathrm{z}}}{D_i^{\mathrm{z}}} + \left(A_i^{\mathrm{zIm}} D_i^{\mathrm{zIm}}\right)\frac{\Delta D_i^{\mathrm{zIm}}}{D_i^{\mathrm{zIm}}} + A_i^{\mathrm{k}}\frac{\Delta K_i^{\mathrm{m}}}{K_i^{\mathrm{m}}} + (A_i^{\mathrm{l}} D_i^{\mathrm{l}})\frac{\Delta D_i^{\mathrm{l}}}{D_i^{\mathrm{l}}}. \tag{13}$$

Here D_i^{z}—total demand of a i region's legal sector for intermediate goods, produced by legal as well as shadow sectors of all regions; K_i^{m}—annual average real capital stock of the legal sector in a region i.

Table 2 Coefficients characterizing the effects of factors of economic growth

i	Region	Value of coefficient			
		α_i	β_i	γ_i	δ_i
1	Akmola	0.764	0.584	0.524	0.951
2	Aktobe	2.088	1.630	2.947	1.528
3	Almaty	0.170	2.812	0.938	0.299
4	Atyrau	2.449	1.372	2.930	2.755
5	West Kazakhstan	0.903	2.911	2.835	0.302
6	Zhambyl	1.442	0.681	0.315	2.174
7	East Kazakhstan	1.087	0.512	2.471	0.334
8	Karaganda	1.644	1.035	2.394	2.446
9	Kostanay	1.672	2.584	2.324	0.616
10	Kyzylorda	0.251	1.489	1.258	1.173
11	Mangystau	2.516	0.260	0.559	2.508
12	Pavlodar	1.606	1.290	2.291	2.098
13	North Kazakhstan	2.527	1.378	2.554	1.192
14	South Kazakhstan	1.964	1.742	2.383	0.539
15	Astana city	2.097	2.802	0.382	1.962
16	Almaty city	1.851	1.954	1.980	2.978

Let $a_i = \frac{\Delta A_i^{\mathrm{r}}}{A_i^{\mathrm{r}}}$ denote the rate of technical progress of a i region's legal sector; $z_i = \frac{\Delta D_i^{z}}{D_i^{z}}$—intermediate goods consumption rate by a legal (or shadow) sector in a region j; $z_i^{\mathrm{Im}} = \frac{\Delta D_i^{z\mathrm{Im}}}{D_i^{z\mathrm{Im}}}$ – imported intermediate goods consumption rate by a legal sector in a region i, $k_i = \frac{\Delta K_i^{\mathrm{m}}}{K_i^{\mathrm{m}}}$ – capital accumulation rate in a i region's legal sector; $l_i = \frac{\Delta D_i^{\mathrm{l}}}{D_i^{\mathrm{l}}}$—labor costs growth rate in a region i, where the sign "Δ" indicates increment of a variable in one year; time in Eq. (13) is omitted for briefness.

Coefficients at the right-hand side of Eq. (13) at the rates indicated above characterize degree of influence of the considered factors on economic growth and allows to compare their influence with influence of technical progress growth rate, at which the coefficient is equal to 1. Having denoted these coefficients in terms of $\alpha_i = A_i^{\mathrm{z}}$, $\beta_i = A_i^{z\mathrm{Im}} D_i^{z\mathrm{Im}}$, $\gamma_i = A_i^{\mathrm{k}}$, $\delta_i = A_i^{\mathrm{l}} D_i^{\mathrm{l}}$, we get brief version of (13):

$$y_i = a_i + \alpha_i z_i + \beta_i z_i^{\mathrm{Im}} + \gamma_i k_i + \delta_i l_i. \tag{14}$$

Below we present the values of the coefficients that determine the contributions of sources of economic growth in legal sector in each region on the basis of the researched model for 2011 (See Table 2). The coefficients in Table 2 show by how many percent (approximately) rate of output growth in legal sector of a region will increase if growth factor (growth rates for corresponding intermediate goods, investment goods, labor) increases by 1% compared to the base case.

Analysis of the coefficients α_i, β_i, γ_i, δ_i in Table 2 shows that if we drop the rate of technical progress, which influence on legal sectors growth rate of all regions in this model is the same, then out of four rest rates of economic growth factors, the greatest impact on the rate of real output in regions 1, 6, 8, 16 has a rate of labor costs; in regions 2, 4, 7, 12, 13, 14—capital accumulation growth rate; in regions 3, 5, 9, 10, 15—consumption rate of imported intermediate goods; and for the rest region 11—consumption rate of imported intermediate goods produced by all regions.

The results of the analysis enable to select the following budget shares of legal sectors in 16 regions as tools for solving regional economic growth problem: $O_{ij}^{1}(t)$—budget share of a legal sector in a region i, assigned to pay for goods and services, purchased from legal sector in a region j; $O_{ij}^{2}(t)$—budget share of a legal sector in a region i, assigned to pay for goods and services purchased from a shadow sector in a region j; $O_{i}^{\mathrm{Im}}(t)$—budget share of a legal sector in a region i, assigned to purchase imported intermediate goods and services; $O_{i}^{\mathrm{l}}(t)$—budget share of a legal sector in a region i, assigned to labor costs; $O_{i}^{\mathrm{n}}(t)$—budget share of a legal sector in a region i, assigned to purchase investment goods. This approach is implemented in the following section.

6 Finding Optimal Values of Economic Instruments

A method of selecting optimal values of economic tools for considered problems of regional economic growth within the framework of parametric regulation theory is associated with the following model, presented by

- optimality criterion K_r, characterizing average growth rate of GRP (gross regional product) as well as relative deviations of per capita GRP in regions from per capita GRP in region 14 (Atyrau region—region, that has the highest value of the stated indicator among all regions of a country in 2000–2011) in 2012–2015:

$$K_r = \frac{1}{4}\sum_{t=2012}^{2015} tYg(t) - \frac{1}{4\sum_{i=1,i\neq 4}^{16}\varepsilon_i}\times \sum_{i=1,i\neq 4}^{16}\left(\varepsilon_i \sum_{t=2012}^{2015}\left|\frac{Yg_i^{1}(t) - Yg_4^{1}(t)}{Yg_4^{1}(t)}\right|\right). \tag{15}$$

  Here: $tYg(t)$—annual GDP rate of a country; $Yg_i^{1}(t)$—per capita GRP in a region i; ε_i—weight coefficient, its value is equal to $\varepsilon_i = 1$ for underdeveloped regions where per capita GRP is lower than national average (regions 1, 3, 5, 6, 9, 10, 13, 14) and $\varepsilon_i = 0.1$ for developed regions, where per capita GRP is higher than national average (regions 2, 7, 8, 11, 12, 15, 16).
- constraints on the state that include constraints imposed on consumer price levels and constraints imposed on per capita GRP in regions:

$$P_{ic}(t) \le \overline{P}_{ic}(t), Yg_i^l(t) \ge \overline{Y}g_i^l(t), i = 1, \ldots, 16, t = 2012, \ldots, 2015. \quad (16)$$

Here: P_{ic}—consumer price level in region i with parametric regulation; $Yg_i^1(t)$—per capita GRP in region i with parametric regulation; a sign "$^-$" indicates basic values of the corresponding indicator (without parametric regulation).

- explicit constraints on control ($O_{ij}^1(t)$, $O_{ij}^2(t)$, $O_i^{\mathrm{Im}}(t)$, $O_i^{\mathrm{l}}(t)$, $O_i^{\mathrm{n}}(t)$; $i, j = 1, \ldots, 16$; $t = 2012, \ldots, 2015$) of Problem 2 (see below):

$$\begin{aligned}
&O_{ij}^1(t) \ge 0;\ O_{ij}^2(t) \ge 0;\ O_i^{\mathrm{Im}}(t) \ge 0;\ O_i^{\mathrm{l}}(t) \ge 0;\\
&O_i^{\mathrm{n}}(t) \ge 0;\ \sum_{j=1}^{16}\left(O_{ij}^1(t) + O_{ij}^2(t)\right) + O_i^{\mathrm{Im}}(t) +\\
&O_i^{\mathrm{l}}(t) + O_i^{\mathrm{n}}(t) \le 1;\ 0.5 \le O_{ij}^1(t)/\bar{O}_{ij}^1 \le 2;\\
&0.5 \le O_{ij}^2(t)/\bar{O}_{ij}^2 \le 2;\ 0.5 \le O_i^{\mathrm{Im}}(t)/\bar{O}_i^{\mathrm{Im}} \le 2;\\
&0.5 \le O_i^{\mathrm{l}}(t)/\bar{O}_i^{\mathrm{l}} \le 2;\ 0.5 \le O_i^{\mathrm{n}}(t)/\bar{O}_i^{\mathrm{n}} \le 2. \quad (17)
\end{aligned}$$

Here $\overline{O}_{ij}^1$, $\bar{O}_{ij}^2$, $\bar{O}_i^{\mathrm{Im}}$, $\bar{O}_i^{\mathrm{l}}$, $\bar{O}_i^{\mathrm{n}}$—fixed basic values of the stated shares, obtained as a result of solving the problem of parametric identification of the model applying the data for 2000–2010.

- explicit constraints on control ($O_i^{\mathrm{k}}(t)$; $i = 1, \ldots, 16$; $t = 2012, \ldots, 2015$) of Problem 3 (see below):

$$O_i^{\mathrm{k}}(t) \ge 0;\quad \sum_{i=1}^{16}\sum_{t=2012}^{2015} O_i^{\mathrm{k}}(t) \le 9.2 \times 10^{12}. \quad (18)$$

Here $O_i^{\mathrm{k}}(t)$—additional investments for subsidizing the legal sector of the region i in year t; 9.2×10^{12}—constraints imposed on the total volume of the stated additional investments in tenge for the period of 2012–2015.

Using the relations Eqs. (15)–(17) and (15), (16), (18) we formulate correspondingly the following problems of parametric control of regional economic growth.

Problem 2 *Based on the CGE model "Center-Regions" find such values of budget shares* ($O_{ij}^1(t)$, $O_{ij}^2(t)$, $O_i^{Im}(t)$, $O_i^l(t)$, $O_i^n(t)$; $i, j = 1, \ldots, 16$; $t = 2012, \ldots, 2015$) *of legal sectors of regions' economy, satisfying the condition* Eq. (17)*, so that corresponding to it solution of the CGE model "Center-Regions" satisfies the conditions* Eq. (16) *and provides maximum of the criterion* Eq. (15).

Problem 3 *Based on the CGE model "Center-Regions" find values of additional investments* ($O_i^k(t)$; $i = 1, \ldots, 16$; $t = 2012, \ldots, 2015$)*, assigned for subsidizing legal sectors of economy in regions, satisfying the condition* Eq. (18)*, so that corresponding to it solution of the CGE model "Center-Regions" satisfies the conditions* Eq. (16) *and provides maximum of the criterion* Eq. (15).

Table 3 Several indicators values' changes in the result of solution of the Problems 2 and 3

	Problem 2	Problem 3
Indicator's change in comparison with basic variant		
Criterion (%) K_r	4.7	5.7
Average growth rate of GDP in 2012–2015, in percentage points	2.25	2.63
GDP per capita in 2015 (%)	10.01	10.5
Maximum GRP to minimum GRP ratio in 2015 (%)	−16.35	−17.07
Mean square deviation of GRP per capita from maximum GRP per capita in 2015 (%)	−13.8	−12.72
Indicator's change for 2015 in comparison with 2011		
Maximum GRP per capita to minimum GRP per capita ratio (%)	−17.3	−17.28
GRP per capita in underdeveloped region (%)	18.5–24.8	20.4–28.4
GRP per capita in developed region (%)	3.5–5.7	3.4–5.9

The results of numerical solution of Problems 2 and 3, which demonstrate an efficiency of application of proposed parametric control method, are given in the following Table 3.

Below we give some results of solution of the Problems 2 and 3 for two regions in the Republic of Kazakhstan, in which maximum changes of economic indicators were observed.

Figure 1 (left) presents the result of solving the Problem 2—graphics of per capita GRP for Akmola region (in thousand tenge in prices of 2000) without control and with parametric control. Per capita GRP growth is 21.2 % by 2015 compared to the base case. If we compare a value reached in 2015 with corresponding value for 2011, then we observe a growth by 69 % (the highest growth among all regions compared with data for 2011).

Figure 1 (right) presents results of solving the Problem 3 – graphics of per capita GRP (in thousand tenge in prices for 2000) for South Kazakhstan region without control and with parametric control. Per capita GRP in this region is 28.4 % by 2015 compared to the base case (the highest growth among all regions). The achieved in 2015 value of this indicator exceeds 1.5 times its value for 2011.

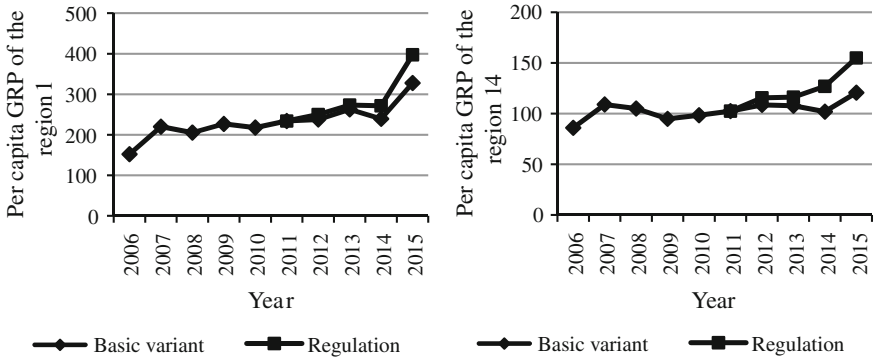


Fig. 1 Per capita GRP of the region 1—Akmola region (*left*) and Per capita GRP of the region 14—South Kazakhstan region (*right*)

7 Conclusion

The paper shows efficiency of the proposed method of parametric identification of large-scale CGE models.

Theorems about sufficient conditions for existence of solution to one parametrical control problem and for continuous dependence of corresponding optimal values of criterion on uncontrollable functions are presented.

The results of computational experiments on evaluation and analysis of sources of regional economic growth on the basis of CGE model "Center-Regions" are presented.

Efficiency of parametric control theory application at solving the problems of regional economic growth on the basis of CGE model "Center-Regions" is shown.

The obtained results can be applied during the development and implementation of an effective state policy in the sphere of economic growth and reduction of regional socio-economic growth disproportion.

References

1. Acemoglu D (2008) Introduction to modern economic growth. Princeton University Press, NJ
2. Turnovsky SJ (1997) Methods of macroeconomic dynamics. The MIT Press, Cambridge
3. André FJ, Cardenete MA, Romero C (2010) Designing public policies. An approach based on multi-criteria analysis and computable general equilibrium modeling. Lecture notes in economics and mathematical systems, 1st edn, vol 642. Springer, Berlin
4. Tapiero CS (1998) Applied stochastic models and control for finance and insurance. Kluwer Academic Publishers, The Netherland
5. Makarov VL, Bakhtizin AR, Sulashkin SS (2007) The use of computable models in public administration. Scientific Expert, Moscow (in Russian)
6. Ashimov AA, Sagadiyev KA, Borovskiy YuV, Iskakov NA (2008) On the market economy development parametrical regulation theory. Kybernetes 37(5):623–636

7. Ashimov AA, Sultanov BT, Adilov ZM, Borovskiy YuV, Novikov DA, Alshanov RA (2013) Macroeconomic analysis and parametrical control of a national economy. Springer, New York
8. Ashimov AA, Borovskiy YuV, Sultanov BT, Alshanov RA, Aisakova BA (2013) Parametric control of regional economic growth based on the one computable general equilibrium model. In: Lecture notes in engineering and computer science: proceedings of the international multiconference of engineers and computer scientists 2013, 13–15 Mar 2013, Hong Kong, pp 171–176

Integrating Dynamic Composition Estimation with Model Based Control for Ethyl Acetate Production

Weerawun Weerachaipichasgul and Paisan Kittisupakorn

Abstract To achieve a high purity of the ethyl acetate production in the batch reactive distillation, an optimal operating condition and an effective control strategy are needed to improve the product quality (maximum of the high purity product). An off-line dynamic optimization is prior determined by maximizing productivity for the batch reactive distillation. A dynamic composition estimation (EKF) based on simplified mathematical models and on-line temperature measurements, is incorporated to estimate the compositions in the reflux drum and the reboiler. The estimate performances of the EKF are investigated the influence of changing in the initial compositions. Model based control, model predictive control (MPC), has been implemented to provide tracking of the desired product compositions subject to rigorous model equations. Simulation results demonstrate that the EKF can still provide good estimates of compositions in the reflux drum and reboiler with respect to the initial compositions change. The MPC based on rigorous mathematical models with the dynamic composition estimator can control the distillation according to the optimal trajectory and then can achieve maximum product as determined. In the presence of the forward reaction rate constant mismatch (unknown/uncertain parameters case), the EKF is still able to provide good accuracy. The MPC integrating with dynamic composition estimation is robust and able to handle the mismatch.

Keywords Batch reactive distillation · Dynamic composition estimation · Dynamic optimization · Ethyl acetate production · Model based controller · Rigorous mathematical model

W. Weerachaipichasgul · P. Kittisupakorn (✉)
Department of Chemical Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok 10330, Thailand
e-mail: Weerawun@gmail.com

P. Kittisupakorn
e-mail: Paisan.K@chula.ac.th

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 275, DOI: 10.1007/978-94-007-7684-5_17,

1 Introduction

Ethyl acetate is an important organic solvent widely applied in the chemical industry that is most frequently produced by the esterification reaction between acetic acid and ethanol. However, the main problem of the production of the ethyl acetate is the equilibrium limitation from the reversible reaction. To overcome restrictions given by chemical reaction equilibrium, reactive distillation (RD) is considered to operate, firstly. RD combines with the distillation and chemical reaction into a single unit operation.

It is well known that the operation of the batch distillation is very attractive; the flexibility in purifying different mixtures under a variety of operational condition, and the separation multicomponent mixture in a single batch column. RD operated into a batch mode operation, batch reactive distillation (BREAD), is more interesting than that operated into a continuous mode.

Absolutely, the operation of BREAD offers many benefits such as the higher conversion and selectivity, lower energy consumption and capital investment. On the other hand, the control of the BREAD is particularly difficult due to the dynamic interaction of reaction kinetics, mass transfer and thermodynamic properties, as a result that the control of BREAD is really a difficult task. There are many advanced control techniques that have been developed and applied to control in the BREAD process. However, the mathematical models are essentials to design the model based, therefore the synthesizing of the mathematical models is considered, firstly. The modeling of batch reactive distillation have been applied to achieve the high quality of product by the optimization technique based on an objective function in the optimization problem that depends on the nature of the problem [1–6].

Although the production composition control can be employed directly by using on-line measured composition, this measurement is expensive, difficult to maintain, necessitating frequent calibrations. Moreover, it introduces measurement delay such as NIR spectroscopic sensor, gas chromatograph analyzer, and etc.. Then the temperature measurement is suitable than the composition measurement. Disadvantage of a direct temperature control is the product composition maybe off-spec because composition will be known at the end of the batch. In distillation column, the tray temperature does not correspond exactly to the compositions [7]. As a result that the composition estimation is one of the techniques that can be applied [8–10].

There are many techniques to infer compositions from the temperature data and the control for the batch distillation and batch reactive distillation processes; for example, an extended Luenberger Observer (ELO) with a conventional PI controller [11], an extended Kalman filter (EKF) [12, 13], a Kalman filter based on multiple reduced order models with a model predictive control based on reduced order model [14, 15], and an artificial neural network (ANN) estimator [16, 17]. When composition control for batch reactive distillation is focused, it has not been much addressed; a model predictive control based on the traveling wave phenomena of simplified model [18], the MPC based on the artificial neural network [17], and the MPC based on the simplified mathematical model [19].

In this study, the development of dynamic composition estimation with model based controller in the batch reactive distillation process is presented. The design of an estimator based on EKF estimator is supported by a simple model of batch reactive distillation that includes reaction kinetics and thermodynamics. The estimation composition performance is evaluated in the term of integral of the absolute error (IAE) between the estimated and actual compositions in the reflux drum and reboiler. In addition, the model based control, model predictive control (MPC), based on a realistic model is considered. The control performances of MPC are compared with a conventional control technique under the nominal and mismatch case.

2 Mathematical Model

In a conventional batch reactive distillation as show in Fig. 1, it has total ten trays, including the reboiler (1st tray) and condenser (10th tray). A reaction mixture is charged into a vessel and heat is added so that reaction takes place to form products and the vapours move up the column. The reaction occurs in liquid phase in reboiler, on all plates and in condenser. The chemical reaction and distillation proceed simul-

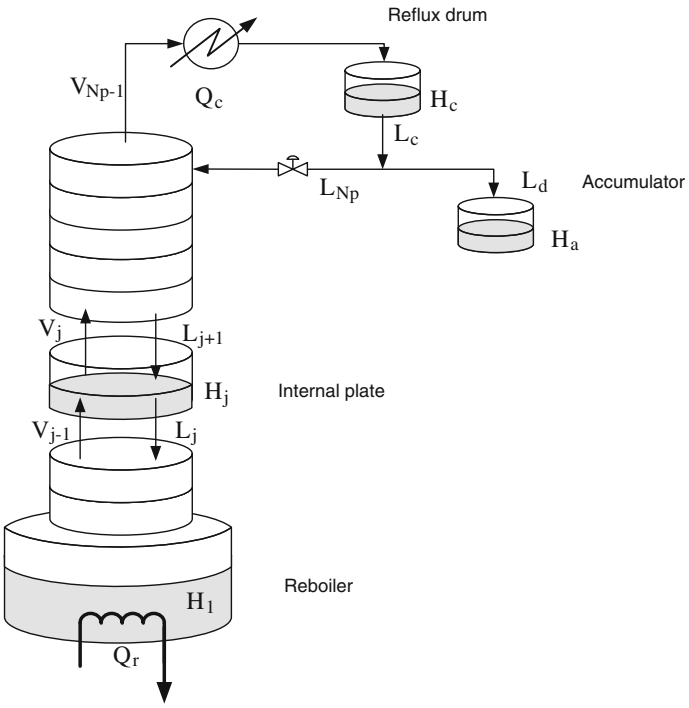


Fig. 1 Conventional batch reactive distillation column

taneously and the product is collected from the condenser. The derivation of model assumes that constant molar holdups on the plates, negligible vapour holdup, no chemical reactions in the vapour phase, the initial state of the column is the steady state total reflux condition with no reactions, constant operating pressure, perfect mixing and equilibrium on all plates, reaction occurred on the plates, and in the condenser and reboiler, fast energy dynamics, and total condensation with no subcooling [1]. To achieve the mathematical models, the material and energy can be balanced around the reboiler, the internal plate, and the reflux drum.

- Reboiler

$$\frac{dH_1}{dt} = L_2 - V_1 + \Delta n_1 H_1 \tag{1}$$

$$\frac{d(H_1 x_{1,i})}{dt} = L_2 x_{2,i} - V_1 y_{1,i} + R_{1,i} H_1, \ where\ i = 1, \ldots, Nc \tag{2}$$

$$\frac{d(H_1 h_1^l)}{dt} = L_2 h_2^l - V_1 h_2^l + Qr \tag{3}$$

- Internal plate $where\ i = 1, ..., Nc\ and\ j = 2 : Np - 1$

$$\frac{dH_j}{dt} = L_{j+1} - L_j + V_{j-1} - V_j + \Delta n_j H_j \tag{4}$$

$$\frac{d\left(H_j x_{j,i}\right)}{dt} = L_{j+1} x_{j+1,i} - L_j x_{j,i} + V_{j-1} y_{j-1,i} - V_j y_{j,i} + R_{j,i} H_j \tag{5}$$

$$\frac{d(H_j h_j^l)}{dt} = L_{j+1} h_{j+1}^l - L_j h_j^l + V_{j-1} h_{j-1}^v - V_j h_j^v \tag{6}$$

- Condenser and Distillate Accumulator

$$\frac{dH_{Np}}{dt} = V_{Np-1} - L_c + \Delta n_{Np} H_{Np} \tag{7}$$

$$\frac{d\left(H_{Np} x_{Np,i}\right)}{dt} = V_{Np-1} y_{Np-1,i} - x_{Np,i} L_c + R_{Np,i} H_{Np} \tag{8}$$

$$\frac{d(H_{Np} h_{Np}^l)}{dt} = V_{Np-1} h_{Np-1}^v - L_c h_{Np}^l - Qc \tag{9}$$

$$\frac{dH_a}{dt} = L_d \tag{10}$$

$$\frac{d\left(H_a x_{a,i}\right)}{dt} = L_d x_{Np,i}, \ where \ i = 1, ..., Nc \tag{11}$$

$$L_d = \left(V_{Np-1} + \Delta n_{Np} H_{Np}\right)\left(1 - r_f\right) \tag{12}$$

where r_f is the reflux ratio, and y is a vapour mol fraction that can calculated by the vapour-liquid equilibrium as;

$$y_{j,i} = K_{j,i} x_{j,i} \tag{13}$$

where $K_{j,i}$, vapour-liquid equilibrium constant, is related to the pure component vapor pressure by the relation:

$$K_{j,i} = \frac{P^*_{j,i}}{P} \tag{14}$$

where P is the column pressure, and $P^*_{j,i}$ is the vapour pressure of component i in plate j. The vapour-liquid equilibrium relation can be used to compute bubble point temperature that $\sum_i^{Nc} y_{j,i} = 1$. The other variables such as h^l, and h^v can be calculated by the relation of enthalpy respectively.

$$h^l_j = f(x_j, T_j, P) \tag{15}$$

$$h^v_j = f(y_j, T_j, P) \tag{16}$$

The esterification reaction of ethanol with acetic acid is reversible reaction as

$$Acetic\ acid(1) + Ethanol(2) \rightleftharpoons Ethylacetate(3) + Water(4)$$

The boiling temperatures of acetic acid, ethanol, ethyl acetate and water are: 391.1, 351.5, 350.3 and 373.2 K, respectively. Ethyl acetate can be removed firstly by distillation to overcome restriction given by chemical reaction equilibrium therefore the reactants conversion will be improved. The column specification and the vapour-liquid equilibrium and kinetic data [1] are given in Table 1.

2.1 Optimal Reflux Policy

In this work, optimal product composition profiles in the condenser are prior determined by maximizing productivity that is subject to the process model Eqs. (1)–(16) and specified product purity is greater than 0.9. To solve a dynamic optimization problem, the reflux ratio is selected to be the decision variable into a finite set in which a piecewise constant function is utilized and reboiler heat duty give constant value along the operation time. It is assumed the operations are dived into 16 intervals to discretize the profile. The optimal composition profiles are shown in Fig. 2.

Table 1 Column specification for ethanol esterification process vapour-liquid equilibrium and kinetic data for ethanol esterificartion reaction

No. of ideal stages (including reboiler and condenser)	10
Total fresh feed	5.0 kmol
Condenser holdup	0.1 kmol
Internal plates holdup	0.0125 kmol
Feed composition	
Acetic acid/ Ethanol/ Ethyl acetate/ Water	0.45/0.45/00/0.10
Column pressure	1.013 bar
Reboiler heat duty	32 MJ/h
Vapour liquid equilibrium	
Acetic acid (1); $K_1 = \left(2.25 \times 10^{-2}\right) T - 7.812, \quad T > 347.6$ K	
$K_1 = 0.001, \quad T \leq 347.6$ K	
Ethanol (2); $\log K_2 = -2.3 \times 10^3 / T + 6.588$	
Ethyl acetate (3); $\log K_3 = -2.3 \times 10^3 / T + 6.742$	
Water (4); $\log K_4 = -2.3 \times 10^3 / T + 6.484$	
Kinetic data	
$R = k_{r1} C_1 C_2 - k_{r2} C_3 C_4$	
where R is rate of reaction, gmol/(l min), and C is concentration for ith component, gmol/l. The rate constants are:	
$k_{r1} = 4.76 \times 10^{-4}$ L/(gmol min)	
$k_{r2} = 1.63 \times 10^{-4}$ L/(gmol min)	

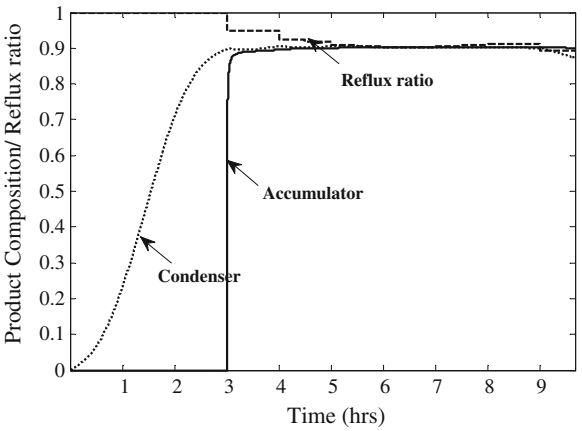


Fig. 2 Optimal composition profiles in the condenser and accumulator with reflux ratio profiles for 16 intervals

3 Dynamic Composition Estimation

The model assumptions to design a dynamic composition estimation based on simplified mathematical models are constant liquid molar holdup on tray and in reflux drum, and constant vapour and liquid flow rates. The simplified equations are given by

$$\frac{dx_{1,i}}{dt} = \frac{r_f V}{H_1}(x_{2,i} - x_{1,i}) - \frac{V}{H_1}(y_{1,i} - x_{1,i}) + R_{1,i} - x_{1,i}\Delta n_1 \tag{17}$$

$$\frac{dx_{j,i}}{dt} = \frac{r_f V}{H_j}(x_{j+1,i} - x_{j,i}) + \frac{V}{H_j}(y_{j-1,i} - y_{j,i}) + R_{j,i} - x_{j,i}\Delta n_j \tag{18}$$

$$\frac{dx_{Np,i}}{dt} = \frac{V}{H_{Np}}(y_{Np-1,i} - x_{Np,i}) + R_{Np,i} - x_{Np,i}\Delta n_{Np} \tag{19}$$

where $i = 1, ..., Nc$ *, and* $j = 2 : Np - 1$

A state vector is $[x_{1,2}, x_{2,2}, \ldots, x_{10,2}, x_{1,3}, x_{2,3}, \ldots, x_{10,3}, x_{1,4}, x_{2,4}, \ldots, x_{10,4}]^T$ in which only three components are considered (ethanol, ethyl acetate and water). The mole fractions acetic acid can be obtained by

$$x_{j,1} = 1 - \sum_{i=2}^{4} x_{j,i} \tag{20}$$

The measurement equations are derived from Antoine's equations [19]. Estimated compositions are directly used by the EKF with available temperatures measurements. Moreover the EKF is also applied to offer the estimated process parameters to handle unknown/uncertain parameters. In the chemical process, the achievement of the mathematical models accuracy to illustrate the process behaviour is not possible because of the limitation of the experimentations. Therefore, the design of the model based control has to concern about uncertain parameter.

In this work, the reaction rate constant is regarded. The sensitivity analysis of the reaction rate constant is shown in Fig. 3. It has been found that the forward reaction constant (k_{r1}) decreased by 30 % from its nominal value is the most sensitive parameter. Here, state equation appended for parameter estimation is:

$$\frac{dk_{r1}}{dt} = 0 \tag{21}$$

The initial condition for model to support the EKF estimator is used with a feed charge of 5 kmol, condenser holdup of 0.1 kmol, and tray holdup of 0.0125 kmol. The diagonal elements of P_0 and Q are selected as 1×10^{-4} for inferential state estimation and 7×10^4 for process parameter. The diagonal elements of R are defined as 10^9 [19].

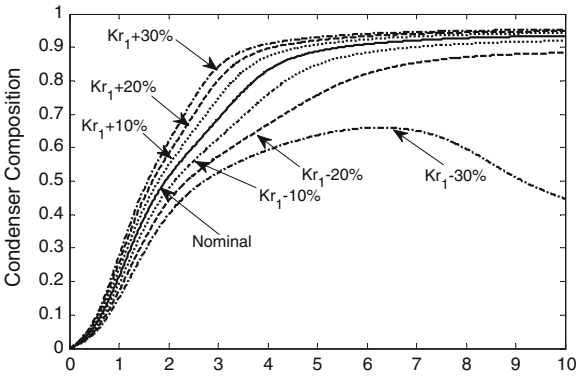


Fig. 3 Sensitivity analysis of the reaction rate constant

4 Model Predictive Control (MPC) Algorithm

Due to its high performance, model based control method has received a great deal of attention to control chemical processes. The formulation of the model based controller, model predictive controller, (MPC) bases on solving an on-line optimal control problem to minimize the problem referred to an objective function which is the sum of squares of deviation of the set point and predicted values on outputs and inputs over the prediction horizon (P). The optimization decision variables are control variable (r_f) for M time steps.

$$\min_{u(k),\ldots,u(k+M-1)} \sum_{i=k=1}^{k+P} \left[\left(y_{sp,i} - y_{pred,i} \right)^2 W_1 + \left(\Delta u_i \right)^2 W_2 \right] \tag{22}$$

Subject to process models: Eqs. (17–19)
Bound on the manipulated variable: $r_{f,\min} \leq r_f \leq 1$
And the end point constraint: $x_{d,3}(t + t_f) = x_{d,3sp}$

where W_1, and W_2 is a weighting on output, and input, respectively. The control structure proposed in this work is shown in Fig. 4. The MPC controller with the dynamic compositions estimator is used to control the product composition to following the desired profile using a reflux ratio as a manipulated variable.

5 Simulation Results

The set point profile (composition of ethyl acetate in the reflux drum) is prior determined by dynamic optimization to solve an optimization problem based on an objective function to maximize productivity (61.74 kg in 9.67 h) for a specified product purity is greater than 0.9.

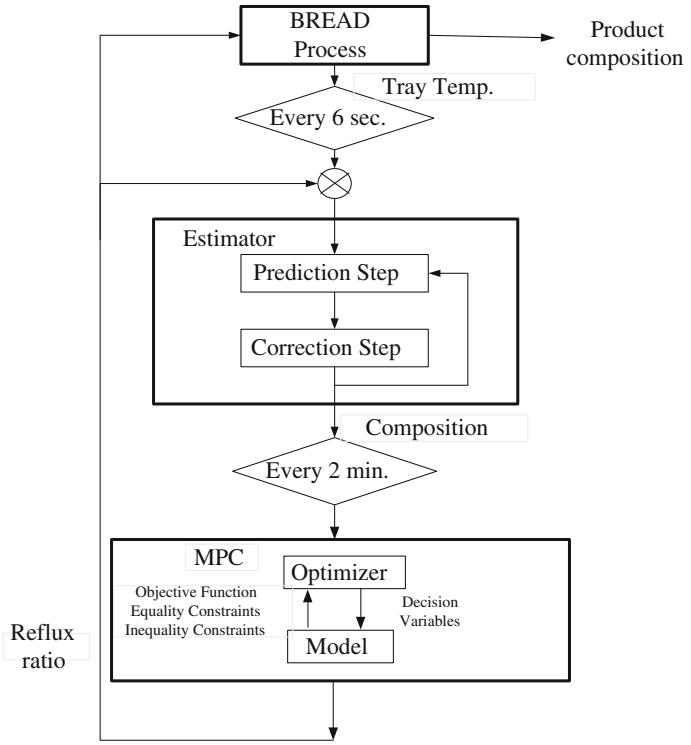


Fig. 4 Block diagram of model based control with dynamic composition estimation for BREAD process

Table 2 Parameters which have been varied

Case 1; 0.45/0.45/0.00/0.10
Case 2; 0.45/0.45/0.02/0.08
Case 3; 0.41/0.41/0.02/0.16

The compositions estimation in the reflux drum and reboiler can be estimated by the measurements of each tray temperature corrupted with a zero mean random Gaussian noise having a standard deviation of 0.5 K. It is well known that the initial compositions of the reactants may vary from batch to batch and also the initial reactant concentrations may not correctly represent the actual compositions. Therefore, the estimator must be able to start with approximation initial conditions and converge to actual conditions in the column. The different initial mixture compositions cases are presented in Table 2. The estimation performance of the EKF can be evaluated in terms of integral of the absolute error (IAE) between the estimated and actual compositions in the reflux drum and reboiler as given in Table 3.

In case 1, the initial condition of the estimation is equal to the initial condition for the simulation ($\hat{x}_0 = x_0 = 0.45/0.45/0.00/0.10$). It can be seen that the estimated

Table 3 Summary of IAE values for the dynamic composition estimation based on EKF estimator

	IAE							
	Acetic acid (1)		Ethanol (2)		Ethyl acetate (3)		Water (4)	
	x_d	x_b	x_d	x_b	x_d	x_b	x_d	x_b
Case 1	0.040	0.031	0.230	0.026	0.210	0.040	0.031	0.230
Case 2	2.77	11.79	26.15	9.82	10.59	2.77	11.79	26.15
Case 3	8.120	15.75	30.82	13.83	10.89	8.12	15.75	30.82

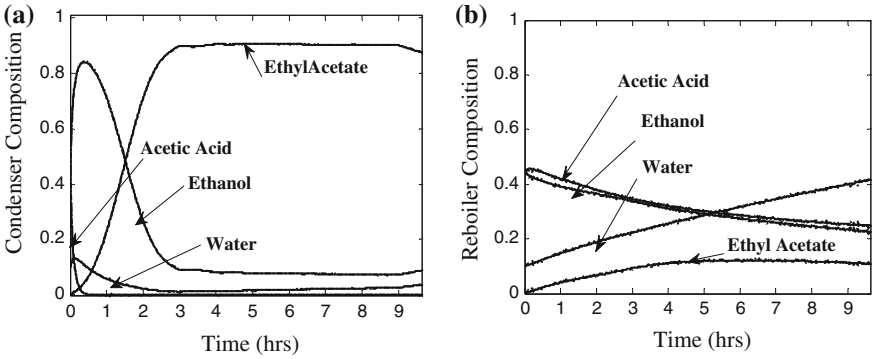


Fig. 5 **a** Actual and estimated composition in distillate, **b** Actual and estimated composition in reboiler; (actual (*solid*); estimated (*dash*)) in case 1

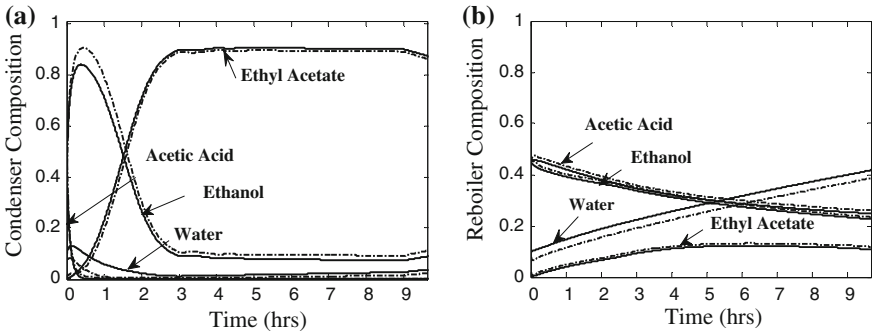


Fig. 6 **a** Actual and estimated composition in distillate, **b** Actual and estimated composition in reboiler; (actual (*solid*); estimated (*dash*)) in case 2

compositions in the reflux drum and reboiler are closed agreement between the actual and estimated responses as shown in Fig. 5a and b, respectively.

When the state estimation is also evaluated by applying it for recycle batches, a fresh feed into the reboiler of 5 kmol with the different initial condition cases to estimate are considered. The actual and estimated compositions in the condenser holdup tank and the reboiler for case 2 ($\hat{x}_0 = 0.45/0.45/0.02/0.08$) and case 3 ($\hat{x}_0 = 0.41/0.41/0.02/0.16$) are shown in Figs. 6 and 7, respectively. It can be seen that the performances of the EKF estimator become worse, if the initial mixture compositions

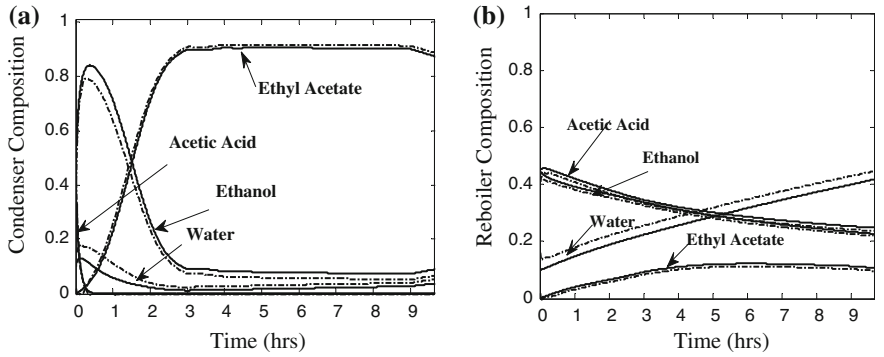


Fig. 7 **a** Actual and estimated composition in distillate, **b** Actual and estimated composition in reboiler; (actual (*solid*); estimated (*dash*)) in case 3

Table 4 Tuning parameter of controllers

PID controller	MPC controller
$K_c = 0.02$	$W_1 = 5,\ W_2 = 0.01$
$\tau_I = 20,\ \tau_D = 0.2$	$P = 15,\ M = 15$

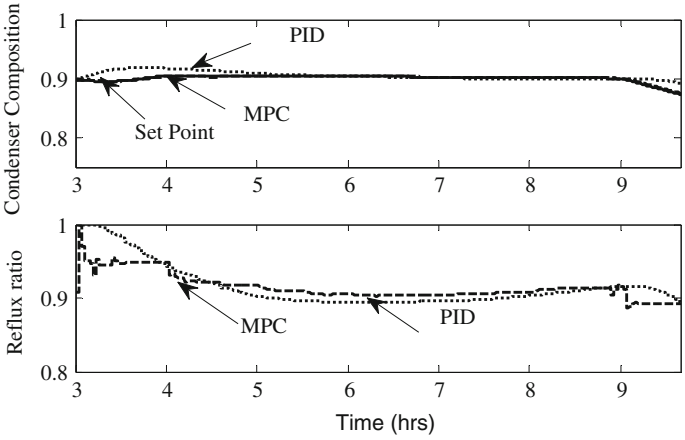


Fig. 8 Product composition profile under controller with dynamic composition estimator in nominal case, and reflux profile

in the estimation are more different than the initial mixture compositions in the simulation.

The model based controller (MPC) is activated at every 2 min after total reflux period to control the composition of ethyl acetate in the reflux drum tracking the desired profile. Tuning parameters of MPC based on the rigorous model and PID controller are presented in Table 4.

The control responses of MPC and PID controllers in the nominal case, all parameters correctly specified, are shown in Fig. 8 and the product quantities are summarized

Table 5 Performance of controllers and product quantity at final time

Case study	Controller with dynamic composition estimator		
	PID	MPC	MPC with parameter estimator
	IAE (product (kg))	IAE (product (kg))	IAE (product (kg))
Nominal	0.045 (59.83)	0.004 (61.48)	0.004 (61.48)
Parameter mismatch	0.371 (37.13)	0.319 (39.41)	0.004 (61.48)

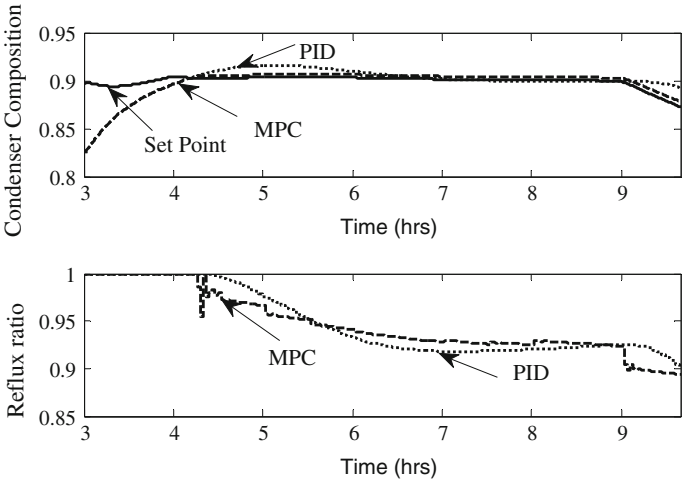


Fig. 9 Product composition profile under controller with dynamic composition estimator in mismatch case of −30 % kr1, and reflux profile

in Table 5. While, the amount of product under the MPC controller based on rigorous mathematical models is greater than the amount of product under the PID controller about 2.76 %, the amount of product under the MPC controller is greater than the amount of product under the PID controller about 0.81 % [19].

Absolutely, the design of the model based controller has to be concerned about parameter uncertainty. In the actual plant, plant/model mismatches and uncertainty in process parameters subsist. The robustness tests regarding plant/model mismatches and uncertainty must be carried out. In this work, the forward reaction rate constant decreased 30 % from its real value will be considered. The control responses of the PID and MPC controllers are illustrated in Fig. 9. Although the MPC and PID controllers can track the desired profile in the mismatch cases, the amounts of the product under the PID and MPC controllers decreases to −37.94 %, −35.90 % respectively.

The EKF has to estimate the forward reaction rate constant that is employed in the MPC formulation. The control response of the MPC controller with the dynamic

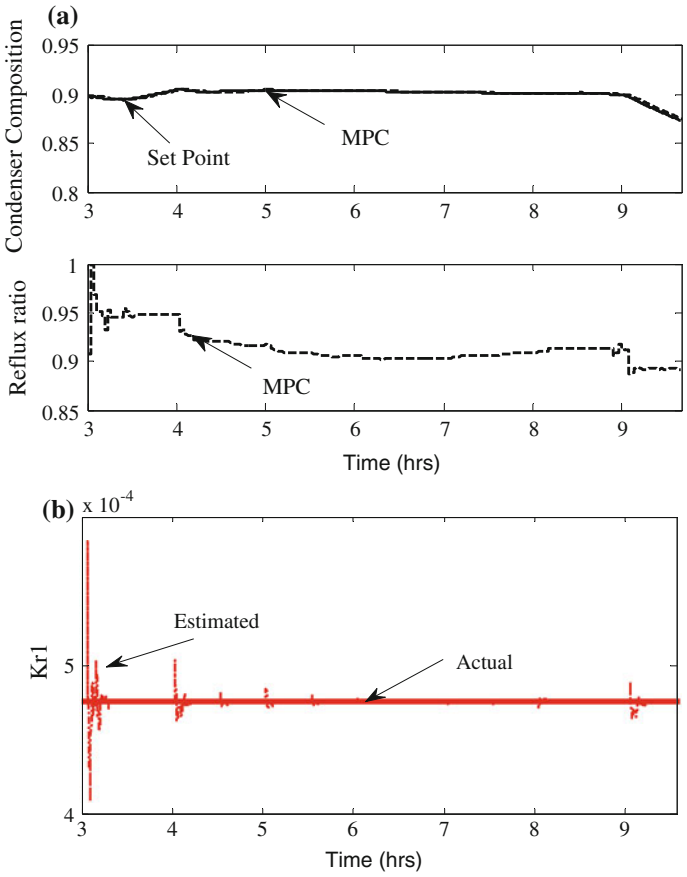


Fig. 10 **a** Product composition profile under controller with dynamic composition and parameter estimator in mismatch case of −30 % kr1 and reflux profile, **b** Estimate value of kr1

composition and parameter estimator in the mismatch case of the rate constant is shown in Fig. 10a and the estimated parameter (kr1) is shown in Fig. 10b. The control performances of the MPC controllers with the EKF for state estimation only, and state and parameter estimation are presented in Table 5.

6 Conclusion

The high product quality in batch reactive distillation improved by model based control with dynamic composition estimator has been proposed. The profile of ethyl acetate composition in the reflux drum is determined by a dynamic optimization to be the set point into the controllers. The design of the dynamic composition estimation (EKF) based on simplified mathematical models coupled with a bubble point

calculation are applied. Moreover, vapour flow rate and holdups of tray and drum are constant. The estimate performances of the EKF are investigated the influence of changing in the initial compositions. The EKF can provide good estimates of compositions in the reflux drum and reboiler. The control performance of the MPC based on rigorous mathematical models with the dynamic composition estimator is the best; maximum product and high performance. In the unknown/uncertain parameters (forward reaction rate constant), the estimator is still able to provide accurate compositions. As a result, the MPC based on rigorous mathematical models with the dynamic composition estimator is still robust and applicable in real plants.

Acknowledgments This work is supported by Thailand Research Fund through the Royal Golden Jubilee Ph.D. Program and Chulalongkorn University.

References

1. Mujtaba IM, Macchietto S (1997) Efficient optimization of batch distillation with chemical reaction using polynomial curve fitting techniques. Ind Eng Chem Res 36:2287–2295
2. Sorensen E, Skogestad S (1994) Control strategies for reactive batch distillation. J Process Control 4:205–217
3. Sorensen E, Macchietto S, Stuart G, Skogestad S (1996) Optimal control and on-line operation of reactive batch distillation. Comput Chem Eng 20:1491–1498
4. Wajge RM, Reklaitis GV (1999) RBDOPT: a general-purpose object-oriented module for distributed campaign optimization of reactive batch distillation. Chem Eng J 75:57–68
5. Konakom K, Saengchan A, Kittisupakorn P, Mujtaba IM (2011) Use of a batch reactive distillation with dynamic optimization strategy to achieve industrial grade ethyl acetate. AIP Conf Proc 1373:262–275
6. Kittisupakorn P, Jariyaboon K, Weerachaipichasgul W (2013) Optimal high purity acetone production in a batch extractive distillationColumn. Lecture notes in engineering and computer science: proceedings of the international multiConference of engineers and computer scientists 2013, Hong Kong, 13–15 Mar 2013, pp 143–147
7. Kano M, Showchaiya N, Hasebe S, Hashimoto I (2003) Inferential control of distillation compositions: selection of model and control configuration. Control Eng Pract 11:927–933
8. Ramesh K, Hisyam A, Aziz N, Abd Shukor SR (2012) Nonlinear model predictive control of a distillation column using wavenet based hammerstein model. Eng Lett 20:330–335
9. Jassar S, Liao Z, Zhao L (2010) Data quality in hybrid neuro-fuzzy based soft-sensor models: an experimental study. IAENG Int J Comput Sci 37:1
10. Kittisupakorn P, Hussain MA (2000) Model predictive control for the reactant concentration control of a reactor, Korean J Chem Eng 17:368–372
11. Quintero-Marmol E, Luyben WL, Georgakis C (1991) Application of an extended Luenberger observer to the control of multicomponent batch distillation. Ind Eng Chem Res 30:1870–1880
12. Venkateswarlu C, Avantika S (2001) Optimal state estimation of multicomponent batch distillation. Chem Eng Sci 56:5771–5786
13. Venkateswarlu C, Jeevan Kumar B (2006) Composition estimation of multicomponent reactive batch distillation with optimal sensor configuration. Chem Eng Sci 62:5560–5574
14. Kaewpradit P, Kittisupakorn P, Thitiyasook P, Mujtaba IM (2008) Dynamic composition estimation for a ternary batch distillation. Chem Eng Sci 63:3309–3318
15. Weerachaipichasgul W, Kittisupakorn P, Saengchan A, Konakom K, Mujtaba IM (2010) Batch distillation control improvement by novel model predictive control. J Ind Eng Chem 16:305–313

16. Bahar A, Ozgen C (2010) State estimation and inferential control for a reactive batch distillation column. Eng Appl Artif Intell 23:262–270
17. Konakom K, Saengchan A, Kittisupakorn P, Mujtaba IM (2011) Neural network-based controller design of a batch reactive distillation column under uncertainty. Asia-Pac J Chem Eng 7(3):361–377
18. Balasubramhanya LS, Doyle FJ III, Balasubramhanya LS (2000) Nonlinear model-based control of a batch reactive distillation column. J Process Control 10:209–218
19. Weerachaipichasgul W, Kittisupakorn P, Mujtaba IM (2013) Improvement of multicomponent batch reactive distillation under parameter uncertainty by inferential State with model predictive control. Lecture notes in engineering and computer science: proceedings of the international multiConference of engineers and computer scientists 2013, Hong Kong, 13–15 Mar 2013, pp 121–126

An Iterative Process for Solving the Constrained Convex Optimization Problem via Fixed Point Methods

Tanom Chamnarnpan and Poom Kumam

Abstract In this paper, we introduce the iterative scheme for finding a common element of the set of fixed pointsand the set of equilibrium problems for nonexpansive mappings. We provide algorithm which strong convergence theorems are obtained in Hilbert spaces. Then, we apply these algorithm to solve some convex optimization problems. The results of this paper extend and improve several results presented in the literature in the recent past.

Keywords Equilibrium problem · Fixed point · Hilbert space · Nonexpansive mapping · Optimization problem · Strong convergence

1 Introduction

Equilibrium problem has emerged as an interesting branch of the applied mathematics. This theory has become a rich source of inspiration and motivation for the study of a large number of problems arising in economics, optimization and operation research in a general.

Consider a real Hilbert space H with inner product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$. Let C be a nonempty closed convex subset of H and $F : C \times C \to R$ is a real function with $F(x, y) = 0$ for all $x \in C$. The "so-called" equilibrium problem for function F is to find a point $x^* \in C$ such that

$$F\langle x^*, y\rangle \geq 0, \quad \forall y \in C. \tag{1}$$

T. Chamnarnpan (✉) · P. Kumam
Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi (KMUTT), 126 Pracha-Uthit Rd., Bang Mod, 10140 Thung Khru, Bangkok, Thailand
e-mail: poom.kum@kmutt.ac.th

T. Chamnarnpan
e-mail: tchamnarnpan538@gmail.com

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 275, DOI: 10.1007/978-94-007-7684-5_18,

Denote the solutions of the equilibrium problem Eq. (1) by $EP(F)$.

In this work, Let $S : C \to C$ be a mapping, we may assume that $Fix(S) \neq \emptyset$, which $Fix(S)$ is closed and convex. So there exists a unique $x^* \in Fix(S)$ satisfies the following :

$$\|x^*\| = \min\{\|x\| : x \in Fix(S)\}.$$

That is, x^* is the minimum-norm fixed point of S.

We recall the basic concept of mappings as shown in the following:

A mapping S from C into itself is said to be a nonexpansive mapping if

$$\|Sx - Sy\| \leq \|x - y\|$$

for any $x, y \in C$.

Since 1967, Halpern introduced an explicit iterative scheme as shown in the following:

$$x_{n+1} = \alpha_n u + (1 - \alpha_n)Sx_n, \ \forall n \geq 0,$$

where $\{\alpha_n\} \subset [0, 1]$. He proved that the convergence theorem which the Halpern's iterative method do find the minimum-norm fixed point x^* of S if $0 \in C$.

In 1997, Combettes and Hirstoaga [1] introduced an iterative scheme of finding the best approximation to the initial data when $EP(F)$ is nonempty and prove a strong convergence theorem.

A typical problem is to minimize a quadratic function over the set of the fixed points of a nonexpansive mapping in a real Hilbert space H:

$$\min_{x \in C} \frac{1}{2}\langle Ax, x\rangle - \langle x, b\rangle, \tag{2}$$

where C is the fixed point set of a nonexpansive mapping T on H and b is a given point in H. In 2003, Xu [3] proved that the sequence $\{x_n\}$ defined by the iterative method below, with the initial guess $x_0 \in H$, chosen arbitrarily:

$$x_{n+1} = (I - \alpha_n A)Tx_n + \alpha_n u, \quad n \geq 0 \tag{3}$$

converges strongly to the unique solution of the minimization problem.

In 2004, Xu studied the iteration process $\{x_n\}$ so called viscosity approximation method as shown in the following:

$$x_{n+1} = \alpha_n f(x_n) + (1 - \alpha_n)Sx_n, \ for\ n \geq 1, \tag{4}$$

where $\{\alpha_n\} \subset (0, 1)$ and $f : C \to C$ is a contraction ($i.e. \exists k \geq 0\, such\, that \|f(x) - f(y)\| \leq k\|x - y\|\ \forall x, y \in C$). He also proved the strong convergence theorem of the sequence $\{x_n\}$ which generated by the above scheme under the appropriate conditions.

In 2007, Takahashi and Takahashi [4] introduced an iterative scheme by the viscosity approximation method for finding a common element of the set of solution Eq. (1) and the set of fixed points of a nonexpansive mapping in a Hilbert space. Let $S : C \to H$ be a nonexpansive mapping. Starting with arbitrary initial $x_1 \in H$, define sequences $\{x_1\}$ and $\{u_n\}$ recursively by

$$\begin{cases} F(u_n, y) + \frac{1}{r}\langle y - u_n, u_n - x\rangle \geq 0, \quad \forall y \in C, \\ x_{n+1} = \alpha_n f(x_n) + (1 - \alpha_n) S u_n, \quad \forall n \in N. \end{cases} \tag{5}$$

They proved that under certain appropriate conditions imposed on $\{\alpha_n\}$ and $\{r_n\}$, the sequences $\{x_n\}$ and $\{u_n\}$ converge strongly to $z \in F(S) \bigcap EP(F)$, where $z = P_{F(S)\cap EP(F)} f(z)$.

In 2011, Yao and Xu [5] independently introduced two iterative methods for finding the minimum-norm fixed point of nonexpansive mapping which is defined on closed convex subset C of H. The proposed algorithms are based on the well-known Browder's iterative method [6] and Halpern's iterative method [7].

Recently, Chamnarnpan and Kumam [12] introduced an explicit method for finding the least norm of fixed points for strict pseudo mappings by using the projection technique. They provide algorithm which strong convergence theorems are obtained in Hilbert spaces. Then, they apply these algorithm to solve some convex optimization problems.

Motivated and inspired by the previous mentioned researches, we present the new strong convergence theorems for approximating minimum-norm fixed point of a nonexpansive mapping and equilibrium problem.

Consequencely, we prove that the sequence $\{x_n\}$ generated by our algorithm converges strongly to the element of minimal norm fixed points of a nonexpansive mapping and the solution of an equilibrium problem. As application, we provide iterative processes for solving the constrained convex optimization problem.

2 Preliminaries

This section collects some lemmas which will be used in proofs for main results in the next section.

Lemma 1 [2] *Assume* $\{a_n\}$ *is a sequence of nonnegative real numbers such that*

$$a_{n+1} \leq (1 - \alpha_n)a_n + \alpha_n \delta_n, \ \ n \geq 0,$$

where $\{\alpha_n\}$ *is a sequence in* $(0, 1)$ *and* $\{\delta_n\}$ *is a sequence in* R *such that*

(a) $\sum_{n=1}^{\infty} \alpha_n = \infty$
(b) $\limsup_{n\to\infty} \frac{\delta_n}{\alpha_n} \leq 0$ *or* $\sum_{n=1}^{\infty} |\delta_n| < \infty$.

Then $\lim_{n\to\infty} a_n = 0$.

For solving the equilibrium problem of a bifunction $F : C \times C \to R$, let us assume that F satisfies the following conditions:

(A1) $F(x, x) = 0$ for all $x \in C$;
(A2) F is monotone, i.e., $F(x, y) + F(y, x) \leq 0$ for any $x, y \in C$;
(A3) for each $x, y, z \in C$, $\lim_{t \to 0} F(tz + (1 - t)x, y) \leq F(x, y)$;
(A4) for each $x \in C$, $y \mapsto F(x, y)$ is convex and lower semicontinuous.

The following lemma appears implicitily in [8]

Lemma 2 [1] *Let C be a nonempty closed convex subset of H and let F be a bifunction of $C \times C$ into R satisfying (A1)–(A4). Let $r > 0$ and $x \in H$. Then, there exist $z \in C$ such that*

$$F(z, y) + \frac{1}{r}\langle y - z, z - x\rangle \geq 0, \quad \forall y \in C. \tag{6}$$

Lemma 3 [1] *Assume that $F : C \times C \to R$ satisfying (A1)–(A4). For $r > 0$ and $x \in H$, define a mapping $T_r : H \to C$ as follows*

$$T_r(x) = \{z \in C : F(z, y) + \frac{1}{r}\langle y - z, z - x\rangle \geq 0, \quad \forall y \in C\} \tag{7}$$

for all $z \in H$. Then, the followimg hold:

1. *T_r is single-valued;*
2. *T_r is firmly nonexpansive, i.e., $\|T_r x - T_r y\|^2 \leq \langle T_r x - T_r y, x - y\rangle$; for any $x, y \in H$,*
3. *$F(T_r) = EP(F)$;*
4. *$EP(F)$ is closed and convex.*

Lemma 4 [9] *(Demiclosedness principle of nonexpansive mapping) Let $S : C \to C$ a nonexpansive mapping with $Fix(S) \neq \emptyset$. If $x_n \rightharpoonup x$ and $(I - S)x_n \to 0$, then $x = Sx$.*

Lemma 5 [10] *Let $\{x_n\}$ and $\{y_n\}$ be bounded sequences in a Banach space E and let $\{\beta_n\}$ be a sequence in $[0, 1]$ with $0 < \liminf \beta_n \leq \limsup \beta_n < 1$. Suppose $x_{n+1} = \beta_n y_n + (1 - \beta_n)x_n$ for all $n \geq 0$ and*

$$\limsup_{n \to \infty}(\|y_{n+1} - y_n\| - \|x_{n+1} - x_n\|) \leq 0. \tag{8}$$

Then $\lim_{n \to \infty} \|y_n - x_n\| = 0$.

3 Main Result

In this section,we shall prove our main theorem as follows:

Theorem 1 *Let H be a real Hilbert space, C be a nonempty closed convex subset of H. Let $F : C \times C \to R$ be a bifunction satisfying the conditions (A1)–(A4), $S : H \to H$ be a nonexpansive mapping with $\Omega := F(S) \bigcap EP(F) \neq \emptyset$. Let $\{\alpha_n\}$ be a sequence in $(0, 1]$, $Lamda \in (0, 1)$ and $\{r_n\} \subset (0, \infty)$ be a real sequence satisfying the following conditions:*

(1) $\lim_{n\to\infty} \alpha_n = 0$; $\Sigma_{n=0}^{\infty}\alpha_n = \infty$; $\Sigma_{n=1}^{\infty}|\alpha_{n+1} - \alpha_n| < \infty$,
(2) $0 < r < r_n$ *for all* $n \geq 0$ *and* $\Sigma_{n=0}^{\infty}|r_n - r_{n+1}| < \infty$, $\liminf_{n\to\infty} r_n > 0$;

where r is a positive constant. For any given $x_0 \in H$, let $\{x_n\}$ and $\{u_n\}$ be the sequences defined by

$$\begin{cases} F(u_n, y) + \frac{1}{r_n}\langle y - u_n, u_n - x\rangle \geq 0, \quad \forall y \in C. \\ x_{n+1} = (1 - \alpha_n)[\lambda S x_n + (1 - \lambda)T r_n x_n] \end{cases} \tag{9}$$

for all $n \geq 0$. Then the sequence $\{x_n\}$ converges strongly to a fixed point of S which is a minimal norm and the unique solution of the equilibrium problem.

We divide the proof of theorem 3 as follows:

Proof **First** we prove that the sequence $\{x_n\}$ and $\{u_n\}$ are bounded in C. From the definition of T_{r_n} in Lemma 3, we assume that $u_n = T_{r_n}x_n$. therefore, for any $q \in \Omega$, we have

$$\|u_n - q\| = \|T_{r_n}x_n - T_{r_n}q\| \leq \|x_n - q\| \tag{10}$$

From Eq. (9) and Eq. (10), we get

$$\begin{aligned} \|x_{n+1} - q\| &= \|(1 - \alpha_n)[\lambda S x_n + (1 - \lambda)u_n] - q\| \\ &= \|(1 - \alpha_n)[(1 - \lambda)(u_n - q) + \lambda(S x_n - q)] - \alpha_n q\| \\ &\leq (1 - \alpha_n)[(1 - \lambda)\|(u_n - q)\| + \lambda\|(x_n - q)\|] + \alpha_n\|q\| \\ &\leq (1 - \alpha_n)[(1 - \lambda)\|(x_n - q)\| + \lambda\|(x_n - q)\|] + \alpha_n\|q\| \\ &= (1 - \alpha_n)\|(x_n - q)\| + \alpha_n\|q\| \\ &\leq \max\{\|(x_n - q)\|, \|q\|\} \\ &\leq ... \\ &\leq \max\{\|(x_0 - q)\|, \|q\|\}. \end{aligned}$$

Hence, we can conclude that

$$\|x_{n+1} - q\| \leq \max\{\|(x_0 - q)\|, \|q\|\},$$

for all $n \geq 0$. This implies that $\{x_n\}$ is bounded sequence in H. By Eq. (10), $\{u_n\}$ is a bounded sequence in C and so $\{T_{r_n}x_n\}$, $\{Sx_n\}$ are bounded in H.

Next, we make an estimation for the sequence $\{\|u_{n+1} - u_n\|\}$. By the definition of T_r, $u_n = T_{r_n}x_n$ and $u_{n+1} = T_{r_{n+1}}x_{n+1}$. We have

$$F(u_{n+1}, y) + \frac{1}{r_{n+1}}\langle y - u_{n+1}, u_{n+1} - x\rangle \geq 0, \quad \forall y \in C; \tag{11}$$

$$F(u_n, y) + \frac{1}{r_n}\langle y - u_n, u_n - x\rangle \geq 0, \quad \forall y \in C; \tag{12}$$

Substituting $y = u_{n+1}$ into Eq. (12) and $y = u_n$ into Eq. (11), we get

$$F(u_{n+1}, u_n) + \frac{1}{r_{n+1}}\langle u_n - u_{n+1}, u_{n+1} - x_{n+1}\rangle \geq 0, \tag{13}$$

$$F(u_n, u_{n+1}) + \frac{1}{r_n}\langle u_{n+1} - u_n, u_n - x_n\rangle \geq 0. \tag{14}$$

By the condition (A2), we have

$$\left\langle u_{n+1} - u_n, \frac{u_n - x_n}{r_n} - \frac{u_{n+1} - x_{n+1}}{r_{n+1}}\right\rangle \geq 0, \tag{15}$$

$$\left\langle u_{n+1} - u_n, u_n - u_{n+1} + u_{n+1} - x_n - \frac{r_n}{r_{n+1}}(u_{n+1} - x_{n+1})\right\rangle \geq 0. \tag{16}$$

Since $\liminf_{n\to\infty} r_n > 0$, we assume that there exists a real number r such that $r_n > r > 0$ for all $n \in N$. Thus, we have

$$\begin{aligned}\|u_{n+1} - u_n\|^2 &\leq \left\langle u_{n+1} - u_n, x_{n+1} - x_n + \left(1 - \frac{r_n}{r_{n+1}}\right)(u_{n+1} - x_{n+1})\right\rangle \\ &\leq \|u_{n+1} - u_n\|\left\{\|x_{n+1} - x_n\| + \left|1 - \frac{r_n}{r_{n+1}}\right|\|u_{n+1} - x_{n+1}\|\right\}.\end{aligned}$$

Thus, we obtain

$$\begin{aligned}\|u_{n+1} - u_n\| &\leq \|x_{n+1} - x_n\| + \frac{1}{r_{n+1}}|r_{n+1} - r_n| \cdot \|u_{n+1} - x_{n+1}\| \\ &\leq \|x_{n+1} - x_n\| + \frac{1}{r}\left|r_{n+1} - r_n\right| \cdot M,\end{aligned} \tag{17}$$

where $M = \sup\{\|u_n - x_n\| : n \in N\}$.

Next, We claim that $\|T_{r_n}x_n - x_n\| \to 0$. Since S is a nonexpansive mapping and from Eq. (9) and Eq. (17), we have

$$
\begin{aligned}
\|x_{n+1}-x_n\| &= \|(1-\alpha_n)[\lambda S x_n+(1-\lambda)T_{r_n}x_n] \\
&\quad -((1-\alpha_{n-1})[\lambda S x_{n-1}+(1-\lambda)T_{r_{n-1}}x_{n-1}])\| \\
&= \|(1-\alpha_n)\lambda S x_n+(1-\alpha_n)\lambda S x_{n-1}-(1-\alpha_n)\lambda S x_{n-1} \\
&\quad -(1-\alpha_{n-1})\lambda S x_{n-1}+(1-\alpha_n)(1-\lambda)T_{r_n}x_n \\
&\quad -(1-\alpha_n)(1-\lambda)T_{r_{n-1}}x_{n-1}+(1-\alpha_n)(1-\lambda)T_{r_{n-1}}x_{n-1} \\
&\quad -(1-\alpha_{n-1})(1-\lambda)T_{r_{n-1}}x_{n-1}\| \\
&\le (1-\alpha_n)\lambda\|Sx_n-Sx_{n-1}\|+\|((1-\alpha_n)-(1-\alpha_{n-1}))\lambda S x_{n-1}\| \\
&\quad +(1-\alpha_n)(1-\lambda)\|T_{r_n}x_n-T_{r_{n-1}}x_{n-1}\| \\
&\quad +\|((1-\alpha_n)-(1-\alpha_{n-1}))(1-\lambda)T_{r_{n-1}}x_{n-1}\| \\
&= (1-\alpha_n)\lambda\|Sx_n-Sx_{n-1}\|+|\alpha_n-\alpha_{n-1}|\lambda\|Sx_{n-1}\| \\
&\quad +(1-\alpha_n)(1-\lambda)\|T_{r_n}x_n-T_{r_{n-1}}x_{n-1}\| \\
&\quad +|\alpha_n-\alpha_{n-1}|(1-\lambda)\|T_{r_{n-1}}x_{n-1}\| \\
&\le (1-\alpha_n)\lambda\|x_n-x_{n-1}\|+|\alpha_n-\alpha_{n-1}|\lambda\|Sx_{n-1}\| \\
&\quad +(1-\alpha_n)(1-\lambda)\|u_n-u_{n-1}\| \\
&\quad +|\alpha_n-\alpha_{n-1}|(1-\lambda)\|T_{r_{n-1}}x_{n-1}\| \\
&\le \|(1-\alpha_n)\lambda\|x_n+x_{n-1}\|+|\alpha_n-\alpha_{n-1}|\lambda\cdot K \\
&\quad +(1-\alpha_n)(1-\lambda)\|u_n-u_{n-1}\| \\
&\quad +|\alpha_n-\alpha_{n-1}|(1-\lambda)\cdot K \\
&\le (1-\alpha_n)\lambda\|x_n-x_{n-1}\|+(\alpha_n-\alpha_{n-1})\cdot K \\
&\quad +(1-\alpha_n)(1-\lambda)\left\{\|x_n-x_{n-1}\|+\frac{1}{r}\left|r_n-r_{n-1}\right|\cdot M\right\}
\end{aligned}
\tag{18}
$$

where $K=\sup\{\|Sx_{n-1}\|+\|T_{r_{n-1}}x_{n-1}\|:n\in N\}\le\infty$.

By condition (1), (2) and Lemma 1, we obtain

$$
\lim_{n\to\infty}\|x_{n+1}-x_n\|=0,\ as\ n\to\infty. \tag{19}
$$

It follows from Eqs. (17), (19) and the condition Eq. (2) that

$$
\|u_{n+1}-u_n\|\to 0,\ as\ n\to\infty. \tag{20}
$$

Furthermore, for any $q\in F(S)\bigcap EP(F)$, from Lemma 3, we get

$$
\begin{aligned}
\|u_n-q\|^2 &= \|T_{r_n}x_n-T_{r_{nq}}\|^2 \\
&\le \langle T_{r_n}x_n-T_{r_{nq}},x_n-q\rangle \\
&= \langle u_n-q,x_n-q\rangle \\
&= \frac{1}{2}\{\|u_n-q\|^2+\|x_n-q\|^2-\|x_n-u_n\|^2\}.
\end{aligned}
$$

Therefore, we get

$$\|u_n - q\|^2 = \|x_n - q\|^2 - \|x_n - u_n\|^2. \tag{21}$$

Since $\alpha_n \to 0$, and $\|x_n - x_{n+1}\| \to 0$, we have

$$\|x_n - u_n\| \to 0 \, as \, n \to \infty. \tag{22}$$

Let $y_n = \frac{(1-\alpha_n)\lambda S x_n}{\alpha_n + (1-\alpha_n)\lambda}$, then the iterative sequence Eq. (9) is equivalent to

$$x_{n+1} = (\alpha_n + (1-\alpha_n)\lambda) y_n + (1 - \alpha_n - (1-\alpha_n)\lambda) x_n. \tag{23}$$

Since $\lim_{n\to\infty}(\alpha_n + (1-\alpha_n)\lambda) = \lambda$, then

$$\begin{aligned}
\|y_n - q\| &= \left\| \frac{(1-\alpha_n)\lambda S x_n}{\alpha_n + (1-\alpha_n)\lambda} - q \right\| \\
&= \left\| \frac{(1-\alpha_n)\lambda S x_n - (\alpha_n + (1-\alpha_n)\lambda) q}{\alpha_n + (1-\alpha_n)\lambda} \right\| \\
&= \left\| \frac{(1-\alpha_n)\lambda S x_n - \alpha_n q - (1-\alpha_n)\lambda q}{\alpha_n + (1-\alpha_n)\lambda} \right\| \\
&\le \frac{(1-\alpha_n)\lambda \|x_n - q\| - \alpha_n \|q\|}{\alpha_n + (1-\alpha_n)\lambda} \\
&= \frac{\alpha_n}{\alpha_n + (1-\alpha_n)\lambda} \|q\| + (1 - \frac{\alpha_n}{\alpha_n + (1-\alpha_n)\lambda}) \|x_n - q\| \\
&\le \max\{\|x_n - q\|, \|q\|\}.
\end{aligned}$$

Thus, $\{y_n\}$ is bounded. Hence by nonexpansiveness of S, we have

$$\begin{aligned}
\|y_{n+1} - y_n\| - \|x_{n+1} - x_n\| &= \left\| \frac{(1-\alpha_{n+1})\lambda S x_{n+1}}{\alpha_{n+1} + (1-\alpha_{n+1})\lambda} - \frac{(1-\alpha_n)\lambda S x_n}{\alpha_n + (1-\alpha_n)\lambda} \right\| \\
&\quad - \|x_{n+1} - x_n\| \\
&\le \frac{(1-\alpha_{n+1})\lambda}{\alpha_{n+1} + (1-\alpha_{n+1})\lambda} \|S x_{n+1} - S x_n\| \\
&\quad + \left| \frac{(1-\alpha_{n+1})\lambda}{\alpha_{n+1} + (1-\alpha_{n+1})\lambda} - \frac{(1-\alpha_n)\lambda}{\alpha_n + (1-\alpha_n)\lambda} \right| \|S x_n\| \\
&\quad - \|x_{n+1} - x_n\| \\
&\le \left(\frac{(1-\alpha_{n+1})\lambda}{\alpha_{n+1} + (1-\alpha_{n+1})\lambda} - 1 \right) \|x_{n+1} - x_n\| \\
&\quad + \left| \frac{(1-\alpha_{n+1})\lambda}{\alpha_{n+1} + (1-\alpha_{n+1})\lambda} - \frac{(1-\alpha_n)\lambda}{\alpha_n + (1-\alpha_n)\lambda} \right| \|S x_n\|.
\end{aligned}$$

From $\{x_n\}$ and $\{S x_n\}$ are bounded sequences and $\lim_{n\to\infty} \alpha_n = 0$, then

$$\limsup_{n\to\infty}(\|y_{n+1}-y_n\|-\|x_{n+1}-x_n\|)\le 0. \tag{24}$$

By Lemma 5, we obtain that $\lim_{n\to\infty}\|y_n-x_n\|=0$. Therefore,

$$\lim_{n\to\infty}\|x_{n+1}-u_n\|=\lim_{n\to\infty}(\alpha_n+(1-\alpha_n)\lambda)\|y_n-u_n\|=0. \tag{25}$$

On the other hand, we consider

$$\begin{aligned}\|x_n-Sx_n\| &\le \|x_n-x_{n+1}\|+\|x_{n+1}-Sx_n\|\\ &= \|x_n-x_{n+1}\|+\|(1-\alpha_n)(\lambda Sx_n+(1-\lambda)u_n)-Sx_n\|\\ &\le \|x_n-x_{n+1}\|+(1-\alpha_n)(1-\lambda)\|x_n-Sx_n\|+\alpha_n\|Sx_n\|.\end{aligned}$$

It follows that

$$\begin{aligned}\|x_n-Sx_n\| \le{}& \frac{1}{1-(1-\alpha_n)(1-\lambda)}\|x_n-x_{n+1}\|\\ &+\frac{1}{1-(1-\alpha_n)(1-\lambda)}\alpha_n\|Sx_n\|\to 0\ as\ n\to\infty.\end{aligned}$$

Next, We prove that $\limsup_{n\to\infty}\langle x^*-x_n,x^*\rangle\le 0$.
Since $\{x_n\}$ is bounded. Then, we can take a subsequence $\{x_{n_i}\}$ of $\{x_n\}$ such that

$$\limsup_{n\to\infty}\langle x^*-x_n,x^*\rangle=\lim_{i\to\infty}\langle x^*-x_{n_i},x^*\rangle.$$

Again, since $\{x_n\}$ is bounded, without loss of generality, we may assume that $x_{n_i}\rightharpoonup x'$. Consequently,

$$\limsup_{n\to\infty}\langle x^*-x_n,x^*\rangle=\langle x^*-x',x^*\rangle\le 0.$$

From Eq. (19), we get

$$\limsup_{n\to\infty}\langle x^*-x_{n+1},x^*\rangle=\langle x^*-x^{'},x^*\rangle\le 0.$$

Notice that $\lim_{n\to\infty}\|x_n-Sx_n\|=0$. By the demiclosedness principle of a nonexpansive mapping S, we have $x'\in Fix(S)$. Since $x^*=EP(F)$. It follows from the properties of a nonexpansive mapping that

$$\limsup_{n\to\infty}\langle x^*-x_n,x^*\rangle=\langle x^*-x',x^*\rangle\le 0. \tag{26}$$

By Eq. (9), we have

$$\begin{aligned}\|x_{n+1} - (1-\alpha_n)x^*\|^2 &= \|[(1-\alpha_n)\lambda S x_{n+1} + (1-\lambda)U_n] - (1-\alpha_n)x^*\| \\ &= \|(1-\alpha_n)[\lambda S x_n + (1-\lambda)U_n] - x^*\| \\ &\leq (1-\alpha_n)\|\lambda(Sx_n - x^*) + (1-\lambda)(U_n - x^*)\| \\ &\leq (1-\alpha_n)\lambda\|x_n - x^*\| + (1-\lambda)\|x_n - x^*\| \\ &\leq (1-\alpha_n)\|x_n - x^*\|. \end{aligned} \tag{27}$$

Observe that

$$\|x_{n+1} - (1-\alpha_n)x^*\|^2 \geq \|x_{n+1} - x^*\|^2 - 2\alpha_n\langle x_{n+1} - x^*, x^*\rangle. \tag{28}$$

Therefore by Eq. (27) and Eq. (28), we get

$$\|x_{n+1} - x^*\|^2 \leq (1-\alpha_n)\|x_n - x^*\|^2 + 2\alpha_n\langle x_{n+1} - x^*, x^*\rangle. \tag{29}$$

By the condition of (2) and the inequality Eq. (26), we can apply Lemma 1 to inequality Eq. (29) and conclude that $\{x_n\}$ converges strongly to x^* as $n \to \infty$ that is, the minimum-norm fixed point of S. This completes the proof. □

Remark 1 Theorem 1 also improve (see [5], Theorem 3.2), in which the restrictions $\lim_{n\to\infty} \frac{\alpha_n}{\alpha_{n+1}} = 1$ is removed.

Corollary 1 *Let H be a real Hilbert space, C be a nonempty closed convex subset of H. Let $S : H \times H \to R$ be a bifunction satisfying the conditions (A1)–(A4), $S : C \to H$ be a nonexpansive mapping with $\Omega := F(S) \neq \emptyset$. Let $\{\alpha_n\}$ be a sequence in* $[0, 1]$ *satisfying the following conditions:*

(1) $\lim_{n\to\infty} \alpha_n = 0$;
(2) $\Sigma_{n=0}^{\infty}\alpha_n = \infty$; $\Sigma_{n=1}^{\infty}|\alpha_{n+1} - \alpha_n| < \infty$.

For any given $x_0 \in H$, let $\{x_n\}$ and $\{u_n\}$ be the sequences defined by

$$x_{n+1} = (1-\alpha_n)[\lambda S x_n + (1-\lambda)T_{r_n}x_n] \tag{30}$$

for all $n \geq 0$. Then the sequence $\{x_n\}$ converges strongly to a fixed point of S which is a minimal norm and the unique solution of the equilibrium problem.

Proof Put $F(x, y) = 0$ for all $x, y \in C$ and $r_n = 1$. Then, from Theorem 1 the sequence $\{x_n\}$ generated in Corollary 1 converges strongly to fixed point of S. □

4 Applications to Convex Optimization Problem

In this section,we apply the proposed methods for approximating the minimum-norm solution of convex function and split feasibility problems. Let's recall that standard constrained convex optimization problem as follows :

$$\text{find } x^* \in C, \text{such that } f(x^*) = \min_{x \in C} f(x), \tag{31}$$

where $f : C \to R$ is a convex, Fréchet differentiable function, C is closed convex subset of H.

It is known that the above optimization problem is equivalent to the following variational inequality:

$$\text{find } x^* \in C, \text{such that } \langle v - x^*, \nabla f(x^*) \rangle \geq 0, \forall\, v \in C, \tag{32}$$

where $\nabla f : H \to H$ is the gradient of f.

It is well-known that the optimality condition Eq. (32) is equivalent to the following fixed point problem:

$$x^* = P_C(I - \mu \nabla f)x^*,$$

where P_C is the metric projection onto C and $\mu > 0$ is positive constant. Based on the fixed point problem, we deduce the projected gradient method.

$$\begin{cases} x_0 \in C, \\ x_{n+1} = x_n - \mu \nabla f(x_n), \quad n \geq 0. \end{cases} \tag{33}$$

Using Theorem 3, we immediately obtain the following result.

Theorem 2 *Assume that the solution set of* Eq. (31) *is nonempty. Let the objective function f be convex, fréchet differentiable and it gradient ∇f is Lipschitz continuous with Lipschitz constant L. In addition, if $0 \in C$ or C is closed convex cone. Let $\mu \in (0, \frac{2}{L})$ and define a sequence $\{x_n\}$ by following*

$$x_{n+1} = (1 - \alpha_n)((I - \mu \nabla f)(x_n) + (1 - \lambda)x_n), \quad n \geq 0$$

where $\lambda \in (0, 1)$ and the sequence $\{\alpha_n\} \subset (0, 1)$ satisfies conditions in Theorem 3. Then the sequence $\{x_n\}$ converges strongly to the minimum-norm solution of the minimization Eq. (31).

Proof Since ∇f is lipschitz continuous with constant L, then well-known that the $P_C(I - \mu \nabla f)$ is nonexpansive mapping. Replace the mapping S with $P_C(I - \mu \nabla f)$ and take $T_{r_n} = I$ in Eq. (30). Therefore, the conclusion of this Theorem 2 follows from Corollary 1 immediately. □

Conclusion

In this paper we obtained a new strong convergence theorem for approximating minimum-norm fixed point of a nonexpansive mappings and equilibrium problem

for an α-inverse strongly monotone operator in a real Hilbert space. Furthermore, as application, we also obtained an iterative process for solving the constrained convex optimization problem.

Acknowledgments The authors would like to thank the Office of the Higher Education Commission for the supports under the Higher Education Research Promotion project. Also, the first author would like to thank the faculty of science, King Mongkut's University of Technology Thonburi for financial support.

References

1. Combettes PL, Hirstoaga SA (1997) Equilibrium programming using proximal-like algorithm. Math Program 78:29–41
2. Xu HK (2002) Iterative algorithms for nonlinear operators. J London Math Soc 66:240–256
3. Xu HK (2003) An iterative approach to quadratic optimization. J Optim Theory Appl 116:659–678
4. Takahashi S, Takahashi W (2007) Viscosity approximation methods for equilibrium problems and fixed point problems in Hilbert spaces. J Math Anal Appl 331(1):506–515
5. Yao YH, Xu HK (2011) Iterative methods for finding minimum-norm fixed points of nonexpansive mappings with applications. Optimization 60(6):645–654
6. Browder FE (1967) Convergence theorems for sequences of nonlinear operators in Banach space. Math Z 100:201–225
7. Halpern B (1967) Fixed points of nonexpanding maps. Math Z 73:961–975
8. Blum E, Oettli W (1994) From optimization and variational inequalities to equilibrium problems. Math Student 63:123–145
9. Xu HK (2002) Another control condition in an iterative method for nonexpansive mappings. Bull Aust Math Soc 65:109–113
10. Iiduka H, Takahashi W, Toyoda M (2004) Approximation of solutions of variational inequalities for monotone mapping. Pan-American Math J 14:49–61
11. Suzuki T (2005) Strong convergence of Krasnoselkii and Mann's type sequences for one parameter nonexpansive semigroups without Bochner integrals. J Math Anal Appl 305:227–239
12. Chamnarnpan T, Kumam P (2013) Iterative algorithm for minimum-norm of fixed point for nonexpansive mapping and convex optimization problems. Lecture notes in engineering and computer science In: Proceedings of the international multiConference of engineers and computer scientists, Hong Kong, Vol. 1, pp. 148–151

A Finite Difference Method for Electrostatics with Curved Boundaries

David Edwards

Abstract Difficulties of the finite difference method occur when a boundary does not lie on the mesh points in the overlaid meshpoint array but passes between them. To overcome this limitation and at the same time to effect high precision solutions to the curvilinear problem a solution is described which extends the internal space of the geometry to the other side of the enclosing boundary by analytic continuation. The boundary potential itself is incorporated into the values of near mesh points by interpolation from the boundary. Using this technique precisions of the order of $\sim 10^{-13}$ have been obtained for the concentric sphere geometry. Thus a fundamental limitation of the finite difference method has been removed.

Keywords Analytic continuation · Curved boundaries · Cylindrically symmetric electrostatics · FDM · Finite difference method · High precision

1 Introduction

The finite difference method (FDM) is a simple computational tool for finding the solution to boundary value problems by an iterative method [1]. The solution is a function, having fixed values on the boundary, satisfying a differential equation at all interior points. The method involves overlaying a set of equally spaced meshpoints over the geometry and then relaxing the mesh.

For boundaries lying on meshpoints, a multi region process has been previously described in a series of papers [2–8] demonstrating the high precision capabilities of the FDM process when both multi regions and high order algorithms are used.

A serious limitation affecting the precisions obtained by the current FDM method occurs when the boundary passes between meshpoints as need occur for curvilinear

D. Edwards (✉)
Member of the IJL research center, Newark, VT, USA
e-mail: dej122842@gmail.com

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 275, DOI: 10.1007/978-94-007-7684-5_19,

boundaries. This article will report the solution to this problem which was presented in [9].

Without loss of generality the discussion will be restricted to cylindrical symmetric electrostatics in which both Laplace's equation and cylindrical symmetry are assumed. The solution of this problem will have immediate applicability to field of electron and ion optics.

1.1 Background, the FDM Process

In order to both standardize our notation and emphasize certain features of FDM, a quick overview of the FDM process is useful [2]. Consider Fig. 1 in which an array of mesh points is overlaid on a geometry represented by three connected line segments all lying on either rows or columns of meshpoints. This geometry represents a closed cylinder in three dimensions. It is noted that potentials of meshpoints falling on the boundary are constant through the relaxation process. Points strictly within the geometry will be designated as ingeometry points.

In order to relax such a mesh, the points of the mesh are stepped through in a sequential manner. At each point the value of the potential is evaluated by means an algorithm using the values of the surrounding mesh points as input. This process is continued until all values have been determined. Iterations are continued until a suitable end criterion is met for the potentials within the net; i.e. the values at all meshpoints have stopped changing. When the end criterion is met, the mesh is said to be relaxed and the potentials within the mesh determined.

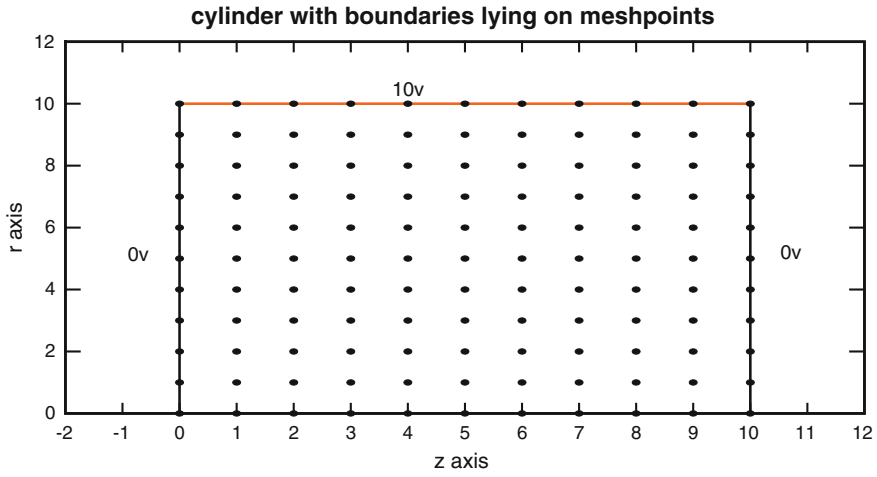


Fig. 1 A cylinder is shown with its boundaries lying on meshpoints and the potentials on the various segments are indicated

1.2 Boundaries Lying Between Meshpoints, Low Order Estimates

If Fig. 1 is slightly modified by having the outer boundary pass between meshpoints rather than lie on meshpoints, a problem for the relaxation process is immediately created, i.e. no mesh point has the boundary potential. If we consider the boundary points as outermost points of the mesh, the values of these points must somehow be estimated in order that the above process is defined. A zeroth order estimate would be to set the value of these outermost points to the potential of the nearest boundary and would allow the mesh to be relaxed as described above. This estimate has the benefit of simplicity in implementation while providing a process solution converging to the problem solution in the limit of high mesh densities. The downside is its inherent lack of precision, since the geometry itself has been modified and no longer corresponds to the model geometry.

A slightly improved estimate [10] of the potential of the outermost points is to place non integral mesh points on the boundary. Using these points, algorithms for points near the boundary can be found but include two additional parameters describing the vertical and horizontal distances from the boundary. These additional parameters make the algorithms for near boundary points considerably more complex and although providing a better estimate for the potential near a boundary than the zeroth order estimate, this complexity will likely preclude its application to any but the lowest order algorithms.

1.3 The Algorithm Development Process

The algorithm development process has been described previously [2–8] and only a brief summary of its essential features is presented here in order to familiar the reader the need for a required set of mesh points as input to any algorithm used in the relaxation process.

About any mesh point in the mesh overlay of the geometry there is assumed to be a power series expansion of the potential v(r, z) as a function of the relative coordinates r, z with respect to the particular mesh point (in this notation the potential at the position of the mesh point itself is v (0, 0)).

The power series expansion of v(r, z) is written:

$$\mathrm{v(r,\,z)} = \mathrm{c0} + \mathrm{c1} * \mathrm{z} + \mathrm{c2} * \mathrm{r} + \cdots + \mathrm{c64} * \mathrm{z} * \mathrm{r\hat{}8} + \mathrm{c65} * \mathrm{r\hat{}8} + \cdots + O(\mathrm{j}) \quad (1)$$

where O(j) (read order j) means terms of order $r^k z^l$ are neglected for k+l $>$ j. j is called the order of the particular class of algorithms generated by this power series. For example for an order 8 algorithm, there are 45 cj's in the above expansion. Requiring that v(r, z) satisfy Laplace's equation in a neighborhood of meshpoint produces *one equation* involving the coefficients cj's and powers of r and z. Further requiring that this equation be true at any point in the neighborhood of the meshpoint

implies that the coefficient of terms $r^k z^l$ are zero and results in a set of 28 linear equations in which only the coefficients cj appear (along with any parameters from Laplace's equation). Thus an additional 17 (45-28) additional equations are required for a solution to the entire set of cjs.

In order to find these additional equations it is noticed that if Eq. (1) is evaluated at a neighbor bj of the central mesh point its value-v(rj, zj)- may be found using Eq. (1), where v(rj, zj) is the value of the meshpoint bj and assumed known. In this way by forming one equation from each of 17 neighboring meshpoints, 17 additional equations are found and the set of 45 equations determined. As this set is linear in cj all cjs can be determined by using the techniques of linear algebra. (It should be noted that the set of selected neighboring meshpoints must provide a consistent set of linear equations for a solution to be determined. This consistency is established during the solution of the equation set. Further the resulting algorithm must give a stable solution when used in the relaxation process as will be discussed below. This set of selected meshpoints is in fact highly degenerate since there are many such sets that will satisfy the above requirements.)

One such set of selected meshpoints used in the 8th order algorithm is shown in Fig. 2. Seen is that meshpoints in the second surrounding ring of the central meshpoint *must be available*, the implication of this observation will be discussed in below. As the actual solution for any cj involves several hundred terms, the visualization of the solution itself is not instructive.

The solution ck depends on both the selection of meshpoints {bj} and "a" the distance of the meshpoint from the axis (a parameter used in Laplace's equation). It may be written:

$$\mathrm{ck} = \mathrm{ck\_coeff\_b0(a)} * \mathrm{b0} + \mathrm{ck\_coeff\_b1(a)} * \mathrm{b1} + \cdots \tag{2}$$

where coeff_bj(a) is a truncated power series in a, the highest power of a which depends upon the order of the algorithm.

There are two situations that will be encountered in the following: the first is one in which the potential at a meshpoint itself is desired. In this situation only c0 need be determined since c0 = v (0, 0). The second situation is one for which the potential is desired at a point somewhere in the vicinity of a meshpoint for which all cj's must be found. the solution for external points

When the geometry of Fig. 1 is relaxed, points one unit from the boundary need to be relaxed using a general mesh point algorithm. Thus using the 8th order algorithm of Fig. 2 on a point one unit below the upper line segment, the two upper points of

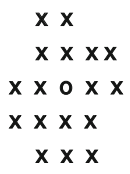


Fig. 2 The set of 17 mesh points around the central mesh point o for a general 8th order mesh point algorithm

Fig. 2 are required but are not available. If the upper boundary were in fact to pass between the meshpoints, then the upper 6 points of Fig. 2 would not be available. The solution to this dilemma is to place meshpoints on the other side of the boundary insuring that any point one unit from a boundary will have the necessary potentials available.

Laplace's boundary value problem is known to have an analytic solution interior to the geometry and that an analytic function may be analytically continued across a boundary by requiring that the function and all of its derivatives be continuous at the boundary [11]. In this way the space of meshpoints is extended to the other side of the boundary. It is noted that in this extended space the boundary itself is assumed to be continuous with continuous derivatives and that Laplace's equation is also applicable to any point in the continued space. The analytically continued meshpoints are designated as external points and since the differential equation itself is applicable in this region, the algorithms for these points may be found as described above with the exception of points closest to the boundary. These points are interpolated from the boundary itself by using Eq. (1). In this manner the potential of the boundary is incorporated into net by folding it into the potentials of the points near the boundary.

The set of external points are explicitly constructed from the ingeometry points near the boundary by requiring every such point have a complete set of four surrounding rings available. This will ensure that an algorithm applied to any such point will have the necessary neighboring potentials available.

In order to be able to create separate algorithms for select types of external points, the external points themselves are classified in the following manner: *Near* points are defined as points of the extended space closest to the boundary. *Middle* points are those external points that have near points as neighbors, *far* having middle points as neighbors, etc. This will allow algorithms at any external point to be created based on its neighbors near_far classification.

In addition each near_far classification is further divided into subtypes depending on the configuration of its surrounding neighbors that have a similar or lower near far classification as the point itself. In this way a point anywhere in the geometry being in a similar neighborhood as other points with the same sub classification will use the same algorithm thereby allowing algorithms to be generated depending upon the point's local neighborhood.

2 The Algorithms for External Points

2.1 Algorithms for Middle, Far, … Points

The potential at a middle, far and veryfar pt is found from the coefficient c0 itself, and depends only upon the potential at its neighboring meshpoints and the distance "a" of the meshpoint from the axis, It turns out that the stability of an algorithm is considerably less sensitive to the selection of its neighboring mesh points if feedback

from points further from the boundary is reduced or eliminated. This can be done by considering points in the selection having a nearfartype less than or equal to that of the meshpoint itself. Thus for example the selection of neighboring of points for a middle point should not include far, or veryfar points. This rule markedly simplifies the search for stable algorithms.

Using Eq. (2) the expression for c0 may be rewritten:

$$\mathrm{c0} = \Sigma_{\mathrm{j}}\,(\mathrm{c0\_coeff\_bj\,(a)} * \mathrm{bj})\,, \tag{3}$$

where bj is the potential of the jth neighbor.

It is noted that for any given meshpoint ck_coeff_bj(a) is a calculated numerical constant dependent on "a" and is constant throughout the relaxation process while bj is the potential of the jth neighbor and changes during the relaxation process. Using this decomposition, the coefficient of each bj, c0_coeff_bj(a), need be calculated once at the start of the relaxation process and used in subsequent iteration cycles.

2.2 Near Point Algorithms

As mentioned previously the algorithms for near points are fundamentally different than the other points in that they involve interpolating the potential from the boundary to the point itself. To do this the difference in the potential between the meshpoint and the boundary is determined and this difference is simply added to the boundary potential to find the potential at the meshpoint. Details of this calculation are given below for a meshpoint whose relative distance to the boundary is rb, zb.

v (r, z) from Eq. (1) can be written:

$$\mathrm{v\ (r,\ z)} = \mathrm{c0} + \Sigma_{\mathrm{k}}\mathrm{ck} * \mathrm{fk\,(r,\ z)\,,\ k\ =\ 1\ to\ kmax}$$

where fk (r, z) is found from the kth term in the following sequence:

$$\mathrm{f1(r,\,z) = z,\ f2(r,\,z) = r,\ f3(r,\,z) = z\text{^}2,\ f4(r,\,z) = z * r,\ f5(r,\,z) = r\text{^}2,\ \ldots}$$

and kmax is the index of the last term in the sequence $r^j z^l$, l+j = algorithm order.

After evaluating v(r, z) at a point on the boundary – v (rb, zb) = vb – the potential at the near meshpoint v(0, 0) can be written:

$$\mathrm{v\,(0,\,0) = vb} - \Sigma_{\mathrm{k}}\mathrm{ck} * \mathrm{fk\,(rb,\ zb)\ sum\ k\ =\ 1\ to\ kmax,\ kmax}$$

Inserting (3) for ck it is found after rearranging the summations that:

$$\mathrm{v\,(0,\,0) = vb} - \Sigma_{\mathrm{j}}\,(\Sigma_{\mathrm{k}}\,\{\,\mathrm{ck\_coeff\_bj\,(a)} * \mathrm{fk\,(rb,\ zb)}\} * \mathrm{value\_bj}\,)$$

Defining: $\mathrm{coeff\_bj\ (a)} = \Sigma_{\mathrm{k}}\,\{\,\mathrm{ck\_coeff\_bj\,(a)} * \mathrm{fk\,(rb,\ zb)}\}$

which is seen to be independent of k depending only on "a".
Finally:

$$\mathrm{v}\,(0,0) = \mathrm{vb}\ - \Sigma_{\mathrm{j}}\mathrm{coeff\_bj}\ (\mathrm{a}) * \mathrm{bj}. \tag{4}$$

The distinct advantage of this formulation is that the calculation involves only one coefficient, coeff_bj (a), which similar to the description above for c0, again being calculated for the meshpoint once prior to the start of the iteration process. In fact during the relaxation process there is no time penalty for the interpolation required by the near points as compared with the calculation for c0 for any other point in the net.

3 Stability Tests

Order 2, 4, 6, and 8 algorithms were created for the various classifications of meshpoints described above and subsequently used in the relaxation of various test meshes. Although the algorithms themselves could be readily found, when applied to certain geometries stability problems were frequently encountered. The general feature of an instability was the observation of unbounded growth of the mesh values during successive iterations of process. In view of this an algorithm is considered stable if the end criterion is met for the relaxation process for all geometries in a collection of test meshes.

The collection of meshes created for stability tests consisted of two types. The first was one constructed of line segments while the second constructed using concentric spheres. These test nets provided a variety of different neighboring configurations for each point classification.

3.1 Linear Segmented Geometries

Linear segmented geometries were constructed with one segment being above, below, to the right of, and to the left of the ingeometry points. The other segments were either horizontal or vertical. This type of construction was made so that within any particular geometry only one type of meshpoint classification would be tested. The potential on all segments was set at 10 so that the after relaxed potential at any point in the net would also be 10. This selection of the boundary potentials provided a zeroth order check on the algorithm itself and on the availability of its required neighbors.

As an example consider: for each non horizontal or vertical segment the segment starts and ends on meshpoints and its angle is varied in 1° increments. A spawned set of geometries with the segment at the same angle is formed by slightly incrementing

the position of the starting and ending points. Further sets of geometries are formed by subdividing the angle itself by fractions of a degree. In each of these geometries a given type of meshpoint would have a different surrounding neighborhood and hence be tested under a variation of the configuration of the meshpoints in its neighborhood. This construction generated over 3000 distinct geometries and was the stability test set for line segmented geometries.

Stable algorithm sets were found for order 2, 4, 6, and 8. It should be noted that while for orders 2 and 4 the stable algorithms were easily determined, finding stable algorithms for the higher order algorithms became progressively more difficult. The procedure was one of trial and error and involves selecting a set of neighboring meshpoints, finding the algorithm, and determining its stability. If unstable another set of mesh points was selected and the process continued until a stable algorithm was found. As mentioned above the search was simplified by minimizing the feedback from meshpoints with a higher nearfartype when possible for all except near types for which a reasonably symmetric set of neighboring points could be used. It should be noted that even for the high order algorithms the stable algorithm was in fact highly degenerate meaning an algorithm created using many different neighbor sets would also be stable.

It is clear that the connection point between any two is a singular point in the potential space. This is true as the potential is not differentiable at this point. The best one can do for points in the vicinity of the singular points is to estimate the potential. This is done here using a low order algorithm, a procedure introduced and described previously [8].

3.2 Geometries with Curved Boundaries

To simulate a geometry with general curved boundaries, concentric hemispheres were selected (representing concentric spherical shells in 3 dimensions) an example of which is shown in Fig. 3.

A collection of test sets for concentric spheres were made consisting of several hundred geometries of this configuration type. Tests were done over this set using the stable algorithms found for the segmented geometries and, with few exceptions, were found to be stable for the concentric sphere geometries. For these exceptions, slight modifications to the algorithms were made and the modified algorithms retested for stability.

In this manner stable algorithms were found orders 2 through 8 for both the line segmented and concentric sphere test sets. Being stable is of course a necessary condition for a useful solution, but gives little indication as to the precision of the process. The next section will discuss methods for determining the process precision.

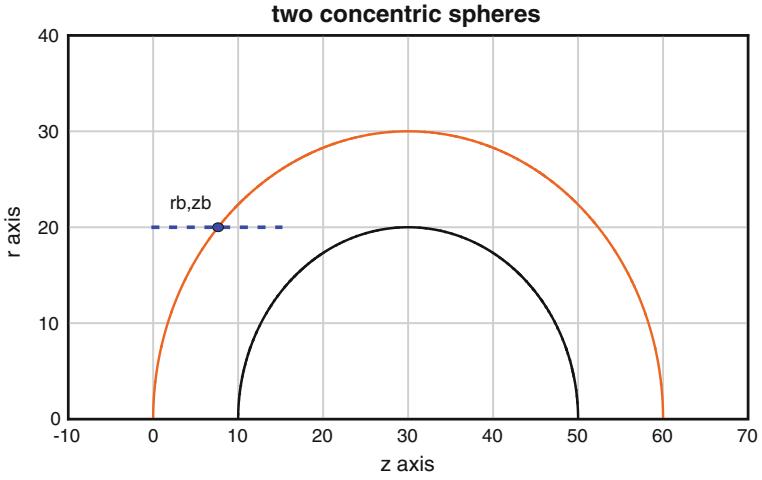


Fig. 3 A geometry consisting of 2 concentric spheres along with a *horizontal line* at r = 20 intersecting the outer boundary at rb, zb

4 Precision Tests

Three types of precision tests have been made. The first, a zeroth order test, already mentioned and completed during the stability trials, was to set all boundary potentials to the same value and to verify that all relaxed potentials attained that value. The second was to estimate the potential on the boundary from the potentials in the relaxed mesh at meshpoints near the boundary. The third and the one most indicative of the actual precision was to find the absolute error for the relaxed mesh using the concentric sphere geometry since the errors can be determined for this geometry as theoretical values for the potential at all points within the geometry are known.

4.1 Low Order Precision Tests: Boundary Position Estimation from the Relaxed Mesh

This first low precision test is to see whether an estimate for the boundary position from the relaxed mesh potential data is reasonably close to the actual position of model boundary. As an example a simple segmented geometry similar to that shown in Fig. 1 is created in which all segments except the left fall on meshpoints. The left vertical boundary being at z = 0.6 falls in between the overlaid mesh point array. The geometry is shown in Fig. 4 and the relaxed potentials determined.

In Fig. 5 the resultant potential values along the line r = 4 are shown near the left boundary along with a fourth order fit to the values at the meshpoints lying on this line. From the location of the point z(r = 4, v = 10) on the fitted curve the z coordinate

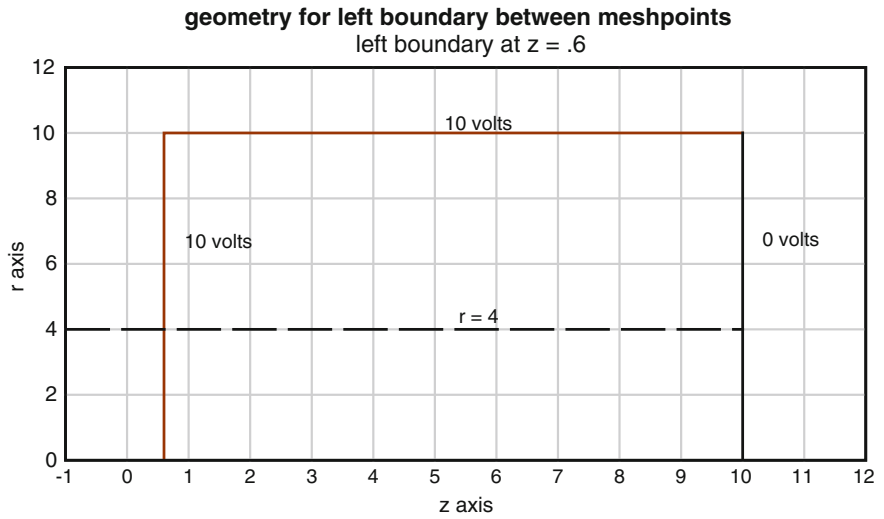


Fig. 4 Shows the geometry used for testing the position of the *left boundary* from a fit to the potentials along a *horizontal line* at r = 4. The *left vertical* boundary falls in between meshpoints

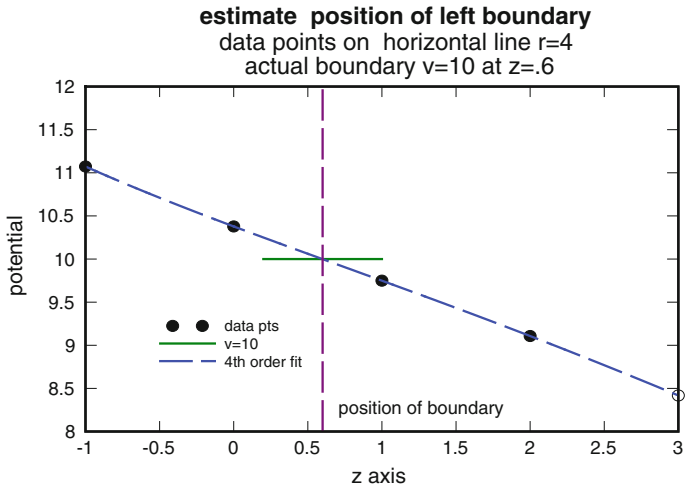


Fig. 5 A plot of the potential at meshpoints near the *left vertical* boundary along the line r = 4. (See Fig. 4). For a discussion see text

of the left boundary line can be inferred. It was found to be 0.598 which is within 0.002 of the position of the actual boundary, 0.600. The residual error in locating the position of the boundary from the relaxed data is likely due to two causes; first, the coarseness of the mesh overlay and second, the presence of singular points at the upper two corners of the geometry. Within these limitations however, the above

test shows that relaxed potential distribution reasonably predicts the position of the model boundary.

Also seen in Fig. 5 is the linear nature of the solution near the boundary which results from the requirement that the analytically continued solution have continuous derivatives at the boundary.

A similar test was made from the relaxed solution of the concentric sphere geometry of Fig. 3. The potential at meshpoints along the line r = 20 and near the upper sphere (the short line shown in Fig. 3) is plotted in Fig. 6 together with a 5th order fit to the data and as described above an estimate of the position of the boundary can be inferred from the location of the point zb(r = 20, v = 10) on the fitted curve.

From this figure it is seen that zb (rb = 20, v = 10) = 7.63932 whereas the theoretical value of the intersection of the line r = 20 with the outer sphere is 7.639320. Thus the position of the actual boundary is again close to that inferred from the after relax potentials indicating that the inferred boundary is reasonably close to the model boundary.

That the estimate in the latter example the boundary position is much closer to that of the geometric model than in the former is likely due to the absence of singular points in concentric sphere geometries.

4.2 Higher Order Precision Tests Using Spheres

In order to do more meaningful precision tests the error in the potential at a selection of points within a soluble geometry must be found. The fact that the segmented

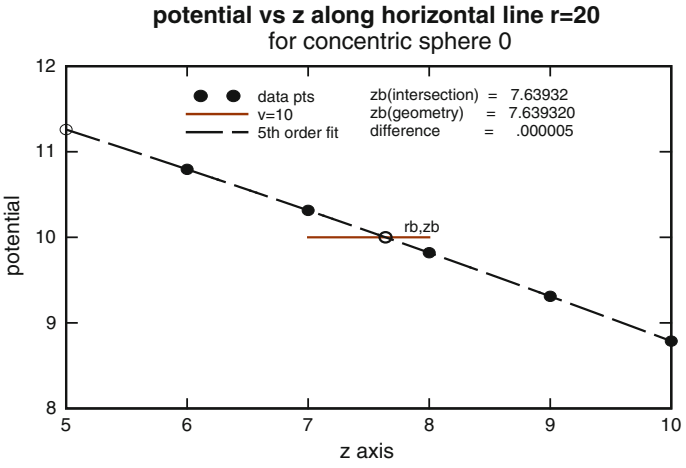


Fig. 6 The solid points are a plot of the after relax potentials at meshpoints along a *horizontal line* in sphere 0 at r = 20 and near z = 7 (see Fig. 3). A 5th order fit to the data is given by the *solid curve*

geometries both contain singular points and in general do not have soluble solutions makes them not useful for these tests. However concentric sphere geometries are suitable as they both have a known theoretical solution and contain no singular points. A collection of concentric sphere geometries were made consisting of the geometry of Fig. 3, the remaining elements of the collection created by scaling from this geometry using a scale factor of 2. Thus con_sphere 0 has Rin = 20, Rout = 30, con_sphere 1 (40, 60), etc. Each geometry has been relaxed with the order 8 algorithm and the errors measured for points within .5 units of the median plane ((Rin + Rout)/2). From the point error data an average error may easily be found for a particular concentric and is plotted in Fig. 7 versus the density points at measured that point, the density of the points given by 1/(Dout*Rout), where Dout = 2 * Rout.

Seen is the super linear increase of precision with density; namely for the density increase of one order of magnitude the error decreases by over 3 orders of magnitude.

Although Fig. 7 was taken from data for the measurement sphere in the median plane of the geometry, varying the radius of this plane from the inner to outer shell had little effect on the error. Similar results pertained to the other test spheres as well, i.e. the error at the median sphere represented to within a factor of 2 or 3 the errors over the entire space, with the error distribution becoming flatter for higher densities.

As a line drawn from any point to the center of the concentric sphere makes an angle alpha with respect to the z axis the distribution of errors vs angle along points in the test sphere may be made and is shown in Fig. 8, again for the highest density geometry studied.

Seen is that the average error for points within 1/2 unit of the median sphere is found to be $\sim 4.4 * 10^{-14}$ being essentially independent of the point's position on the test sphere.

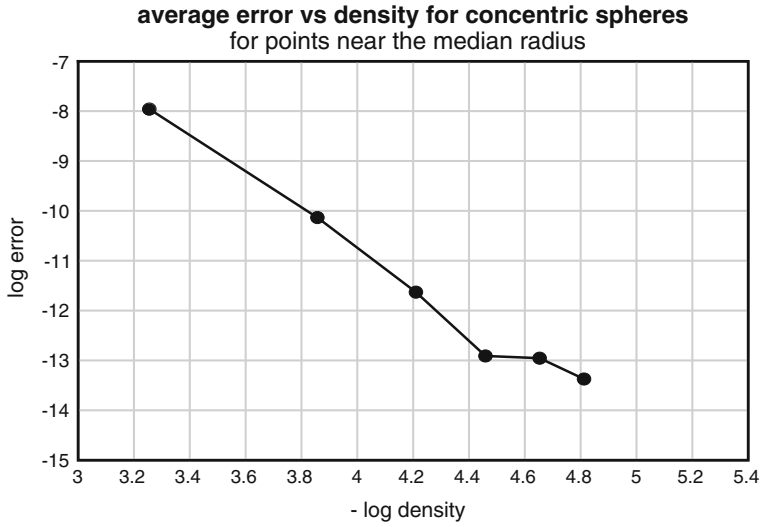


Fig. 7 Shows the average error versus density for points near the median radius of the concentric spheres in the test set

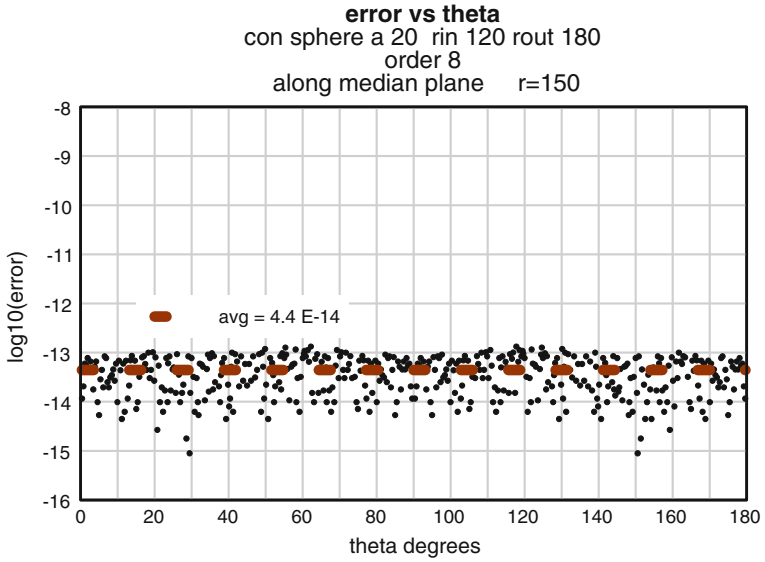


Fig. 8 Plots the error versus theta for our highest density geometry

5 Summary and Conclusion

Extending FDM to curvilinear geometries has been enabled by expanding the meshpoints overlaid on the geometry to points on the other side of the boundary. This was done by extending the potential space by analytic continuation to points across the boundary. The process of creating the required algorithms for the expanded set of meshpoints has been described in some detail, the algorithm itself classified as to the order of the power series used to represent the potential near a meshpoint. Stable algorithms have been created for orders 2, 4, 6, and 8 and tested using a large selection of test geometries. The precision of the solution has been determined for the concentric sphere geometry for which there is a known solution. It was found for the higher mesh density overlays that the precision for the order 8 algorithm was $\sim< 10^{-13}$ within the region between the spheres and in fact for the maximum density studied the average error for points near the median plane was found to be 4.24×10^{-14}.

The implication of this study is that the technique described in this report for dealing with curvilinear boundaries is capable of precisions of the order of 10^{-13} and hence providing a high precision solution to the curvilinear boundary value problem.

References

1. Heddle DWO (2000) Electrostatic lens systems, 2nd ed. Institute of Physics Publishing, ISBN 0-7503-0697-1, pp 32–60
2. Edwards D Jr (1983) Accurate calculations of electrostatic potentials for cylindrically symmetric Lenses. Rev Sci Instrum 54:1229–1235
3. Jr Edwards David (2007) High precision electrostatic potential calculations for cylindrically symmetric lenses. Rev Sci Instrum 78:1–10
4. Edwards D Jr (2008) High precision multiregion FDM calculation of electrostatic potential. Adv Ind Eng Oper Res, Springer. ISBN: 978-0-387-74903-7
5. Edwards D Jr (2008) Single point FDM algorithm development for points one unit from a metal surface. In: Proceedings of international multi conference of engineers and computer scientists, Hong Kong
6. Edwards D Jr (2007) Accurate potential calculations for the two tube electrostatic lens using a multiregion FDM method. In: Proceedings EUROCON 2007, Warsaw, Sept 9–13
7. Edwards D Jr (2010) The use of shadow regions in multi region FDM: high precision cylindrically symmetric electrostatics. In: Taniar D et al (eds) ICCSA 2010, Part II, LNCS 6017. Springer, Berlin Heidelberg, pp 1–13
8. Edwards D Jr (2011) Highly accurate potential calculations for cylindrically symmetric geometries using multi-region FDM: a review. Nuclear Instrum Methods Phys Res A, Elsevier B V 2011, pp 283–291
9. Edwards D Jr (2013) A finite difference method for cylindrically symmetric electrostatics having curvilinear boundaries, lecture notes in engineering and computer science. In: Proceedings of The international multi conference of engineers and computer scientists, 13–15 March, 2013, Hong Kong, pp 1161–1167
10. Khursheed A (1999) The finite element method in charged particle optics. Kluwer Academic Publishers, pp 46–60. ISBN 0-7923-8611-6
11. Erwin K (1962) Advanced engineering mathematics. Wiley, New York

Modified Iterative Scheme for Multivalued Nonexpansive Mappings, Equilibrium Problems and Fixed Point Problems in Banach Spaces

Uamporn Witthayarat, Kriengsak Wattanawitoon and Poom Kumam

Abstract In this research, we modified iterative scheme for finding common element of the set of fixed point of total quasi-ϕ-asymptotically nonexpansive multivalued mappings, the set of solution of an equilibrium problem and the set of fixed point of relatively nonexpansive mappings in Banach spaces. In addition, the strong convergence for approximating common solution of our mentioned problems is proved under some mild conditions. Our results extend and improve some recent results announced by some authors. We divide our research details into three main sections including Introduction, Preliminaries, Main Results. First, we introduce the backgrounds and motivations of this research and follow with the second section, Preliminaries, which mention about the tools that will be needed to prove our main results. In the last section, Main Results, we propose the theorem and corollary which is the most important part in our research.

Keywords Banach space · Equilibrium problem · Fixed point problem · Hybrid projection method · Multivalued nonexpansive mapping · Strong convergence

1 Introduction

Let C be a nonempty closed convex subset of a real Banach space E. A mapping $t : C \to C$ is said to be *nonexpansive* if $\|tx - ty\| \leq \|x - y\|$ for all $x, y \in C$.

U. Witthayarat · K. Wattanawitoon · P. Kumam (✉)
King Mongkut's University of Technology Thonburi, 126 Pracha-Uthit Rd., Bang Mod, Thung Khru, Bangkok, Thailand
e-mail: poom.kum@kmutt.ac.th

K. Wattanawitoon
e-mail: kriengsak.wat@rmutl.ac.th

U. Witthayarat
e-mail: u.witthayarat@hotmail.com

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 275, DOI: 10.1007/978-94-007-7684-5_20,

We denote by $F(t)$ the set of fixed points of t, that is $F(t) = \{x \in C : x = tx\}$. A mapping t is said to be an asymptotic fixed point of t (see [1]) if C contains a sequence $\{x_n\}$ which converges weakly to p such that $\lim_{n\to\infty} \|x_n - tx_n\| = 0$. The set of asymptotic fixed points of t will be denoted by $\widehat{F}(t)$. A mapping t from C into itself is said to be *relatively nonexpansive* [2–4] if $\widehat{F}(t) = F(t)$ and $\phi(p, tx) \leq \phi(p, x)$ for all $x \in C$ and $p \in F(t)$.

The asymptotic behavior of a relatively nonexpansive mapping was studied in [5, 6].

Let $N(C)$ and $CB(C)$ denote the family of nonempty subsets and nonempty closed bounded subsets of C, respectively. Let $H : CB(C) \times CB(C) \to \mathbf{R}^+$ be the Hausdorff distance on $CB(C)$, that is

$$H(A, B) = max\{\sup_{a\in A} dist(a, A), \sup_{b\in B} dist(b, B)\},$$

for every $A, B \in CB(C)$, where $dist(a, B) = \inf\{\|a - b\| : b \in B\}$ is the distance from the point a to the subset B of C.

A multi-valued mapping $T : E \to CB(C)$ is said to be nonexpansive if

$$H(Tx, Ty) \leq \|x - y\|,$$

for all $x, y \in C$. An element $p \in C$ is called a fixed point of $T : C \to CB(C)$, if $p \in Tp$. The set of fixed point T is denoted by $F(T)$.

A point $p \in C$ is said to be an *asymptotic fixed point* of $T : C \to CB(C)$, if there exists a sequence $\{x_n\} \subset C$ such that $x_n \rightharpoonup x \in E$ and $d(x_n, Tx_n) \to 0$. Denote the set of all asymptotic fixed points of T by $\widehat{F}(T)$. T is said to be *relatively nonexpansive*, if $F(T) \neq \emptyset$, $\widehat{F}(T) = F(T)$ and $\phi(p, z) \leq \phi(p, x)$, $\forall x \in C$, $p \in F(T)$, $z \in Tz$. A mapping T is said to bo*closed*, if for any sequence $\{x_n\} \subset C$ with $x_n \to x \in C$ and $d(y, Tx_n) \to 0$ then $d(y, Tx) \to 0$. T is said to be *quasi-ϕ-nonexpansive* if $F(T) \neq \emptyset$ and $\phi(p, z_n) \leq \phi(p, x)$, $\forall x \in C$, $p \in F(T)$, $z_n \in T^n(x)$. T is said to be *quasi-ϕ-asymptotically nonexpansive* if $F(T) \neq \emptyset$ and there exists a real sequence $k_n \subset [1, +\infty)$, $k_n \to 1$ such that

$$\phi(x, z_n) \leq k_n\phi(p, x), \forall x \in C, p \in F(T), z_n \in T^n x. \tag{1}$$

A mapping T is said to be *total quasi-ϕ-asymptotically nonexpansive* if $F(T) \neq \emptyset$ and there exists nonnegative real sequence $\{v_n\}, \{\mu_n\}$ with $v_n, \mu_n \to 0$ as $n \to \infty$ and a strictly increasing continuous function $\zeta : \mathbf{R}^+ \to \mathbf{R}^+$ with $\zeta(0) = 0$ such that

$$\phi(x, z_n) \leq \phi(p, x) + v_n\zeta(\phi(p, x)) + \mu_n, \quad \forall x \in C,\ p \in F(T),\ z_n \in T^n x. \tag{2}$$

A mapping T is said to be *uniformly L -Lipschitz continuous*, if there exists a constant $L > 0$ such that $\|x_n - y_n\| \leq L\|x - y\|$, where $x, y \in C$, $x_n \in T^n x$, $y_n \in T^n y$.

Let E be a real Banach space, E^* the dual space of E. Let C be a nonempty closed convex subset of E and f a bifunction from $C \times C$ to $\mathbf{R}$, where $\mathbf{R}$ denotes the set of

numbers. The *equilibrium problem* (for short, EP) is to find $p \in C$ such that

$$f(p, y) \geq 0, \quad \forall y \in C. \tag{3}$$

The set of solutions of (3) is denoted by $EP(f)$. There are several other problems, for example, the complementarity problem, fixed point problem and optimization problem, which can also be written in the form of an EP. In other words, the EP is an unifying model for several problems arising in physics, engineering, science, optimization, economics, etc.

In 2008, by using a (new) hybrid method, Takahashi et al. [7] proved the following theorem.

Theorem 1 (Takahashi et al. [7]). *Let H be a Hilbert space and let C be a nonempty closed convex subset of H. Let $\{T_n\}$ and T be families of nonexpansive mappings of C into itself such that $\cap_{n=1}^{\infty} F(T_n) := F(\mathcal{T}) \neq \emptyset$ and let $x_0 \in H$. Suppose that $\{T_n\}$ satisfies the NST-condition (I) with $\mathcal{T}$. For $C_1 = C$ and $x_1 = P_{C_1} x_0$, define a sequence $\{x_n\}$ of C as follows:*

$$\begin{cases} y_n = \alpha_n x_n + (1 - \alpha_n) T_n x_n, \\ C_{n+1} = \{z \in C_n : \|y_n - z\| \leq \|x_n - z\|\}, \\ x_{n+1} = P_{C_{n+1}} x_0, \qquad n \in \boldsymbol{N}, \end{cases} \tag{4}$$

where $0 \leq \alpha < 1$ for all $n \in \boldsymbol{N}$ and $\{T_n\}$ is said to satisfy the NST-condition (I) with $\mathcal{T}$ if for each bounded sequence $\{z_n\} \subset C$, $\lim_{n\to\infty} \|z_n - T_n z_n\| = 0$ implies that $\lim_{n\to\infty} \|z_n - T z_n\| = 0$ for all $T \in \mathcal{T}$. Then, $\{x_n\}$ converges strongly to $P_{F(\mathcal{T})} x_0$.

Note that, recently, many authors try to extend the above result from Hilbert spaces to a Banach space setting.

Let E be a real Banach space with dual E^*. Denote by $\langle \cdot, \cdot \rangle$ the duality product. The *normalized duality mapping* J from E to 2^{E^*} is defined by $Jx = \{f \in E^* : \langle x, f \rangle = \|x\|^2 = \|f\|^2\}$, for all $x \in E$. The function $\phi : E \times E \to \mathbf{R}$ is defined by

$$\phi(x, y) = \|x\|^2 - 2\langle x, Jy \rangle + \|y\|^2, \qquad \text{for all } x, y \in E. \tag{5}$$

On the other hand, Matsushita and Takahashi [8] introduced the following iteration: a sequence $\{x_n\}$ defined by

$$x_{n+1} = \Pi_C J^{-1}(\alpha_n J x_n + (1 - \alpha_n) J T x_n), \quad n = 0, 1, 2, \ldots, \tag{6}$$

where the initial guess element $x_0 \in C$ is arbitrary, $\{\alpha_n\}$ is a real sequence in $[0, 1]$, T is a relatively and Π_C denotes the generalized projection from E onto a closed convex subset C of E. Under some suitable conditions, they proved that the sequence $\{x_n\}$ converges weakly to a fixed point of T.

Moreover, Matsushita and Takahashi [9] proposed the following modification of iteration Eq. (6) in a Banach space E:

$$\begin{cases} x_0 = x \in C, \qquad \text{chosen arbitrarily}, \\ y_n = J^{-1}(\alpha_n Jx_n + (1-\alpha_n)JTx_n), \\ C_n = \{z \in C : \phi(z, y_n) \le \phi(z, x_n)\}, \\ Q_n = \{z \in C : \langle x_n - z, Jx - Jx_n\rangle \ge 0\}, \\ x_{n+1} = \Pi_{C_n \cap Q_n} x, \qquad n = 0, 1, 2, \ldots, \end{cases} \tag{7}$$

and proved that the sequence $\{x_n\}$ converges strongly to $\Pi_{F(T)}x$.

In 2012, Chang et al. [10] modified the Halpern-type iteration algorithm for total quasi-ϕ-asymptotically nonexpansive mappings to have the strong convergence under a limit condition in Banach spaces. Recently, Tang and Chang [11] introduce the concept of total quasi-ϕ-asymptotically nonexpansive multivalued mappings in Banach spaces , let $\{x_n\}$ be a sequence generated by

$$\begin{cases} x_0 \in C, \ \text{ is arbitrary}, \\ C_0 = C, \\ y_n = J^{-1}(\alpha_n Jx_n + (1-\alpha_n)JTz_n), \\ z_n = J^{-1}(\beta_n Jx_n + (1-\beta_n)JTw_n), \\ C_{n+1} = \{v \in C_n : \phi(v, y_n) \le \phi(v, x_n) + \xi_n\}, \\ x_{n+1} = \Pi_{C_{n+1}} x_0, \end{cases} \tag{8}$$

$\forall n \ge 0$, where $w_n \in T^n x_n$, $\xi_n = \nu_n \sup_{p\in F} \zeta(\phi(p, x_n)) + \mu_n$ and showed that the sequence $\{x_n\}$ converges strongly to $\Pi_F x_0$.

Recently, Wattanawitoon et al. [12], obtained the strong convergence theorem for finding a common element of the zero point set of a maximal monotone operator and the fixed point set of a sequence for multivalued nonexpansive mappings in the framework of Banach spaces.

In this paper, motivated by Tang and Chang [11] and Wattanawitoon et al. [12], we prove strong convergence theorems for fixed points of a sequence for multivalued nonexpansive mappings, equilibrium problems and fixed point problems in Banach spaces by using the hybrid projection methods. Our results extend and improve the recent results by Tang and Chang [11] and many others.

2 Preliminaries

In this section, we will recall some basic concepts and useful well known results.

A Banach space E is said to be *strictly convex* if

$$\left\| \frac{x+y}{2} \right\| < 1, \tag{9}$$

for all $x, y \in E$ with $\|x\| = \|y\| = 1$ and $x \ne y$. It is said to be *uniformly convex* if for any two sequences $\{x_n\}$ and $\{y_n\}$ in E such that $\|x_n\| = \|y_n\| = 1$ and

$$\lim_{n\to\infty} \|x_n + y_n\| = 2, \tag{10}$$

$\lim_{n\to\infty} \|x_n - y_n\| = 0$ holds.

Let $U = \{x \in E : \|x\| = 1\}$ be the unit sphere of E. Then the Banach space E is said to be *smooth* if

$$\lim_{t\to 0} \frac{\|x + ty\| - \|x\|}{t}, \tag{11}$$

exists for each $x, y \in U$. It is said to be *uniformly smooth* if the limit is attained uniformly for $x, y \in E$.

In our work, the concept duality mapping is very important. Here, we list some known facts, related to the duality mapping J, as following:

(a) E (E^*, resp) is uniformly convex if and only if E^* (E, resp.) is uniformly smooth.
(b) $J(x) \neq \emptyset$ for each $x \in E$.
(c) If E is reflexive, then J is a mapping of E onto E^*.
(d) If E is strictly convex, then $J(x) \cap J(y) \neq \emptyset$ for all $x \neq y$.
(e) If E is smooth, then J is single valued.
(f) If E is uniformly smooth, then J is uniformly norm to norm continuous on each bounded subset of E.
(g) If E is a Hilbert space, then J is the identity operator.

For more information, the readers may consult [13, 14].

If C is a nonempty closed convex subset of real a Hilbert space H and $P_C : H \to C$ is the *metric projection*, then P_C is nonexpansive. Alber [15] has recently introduced a *generalized projection* operator Π_C in a Banach space E which is an analogue representation of the metric projection in Hilbert spaces.

The generalized projection $\Pi_C : E \to C$ is a map that assigns to an arbitrary point $x \in E$ the minimum point of the functional $\phi(y, x)$, that is, $\Pi_C x = x^*$, where x^* is the solution to the minimization problem

$$\phi(x^*, x) = \min_{y\in C} \phi(y, x).$$

Notice that the existence and uniqueness of the operator Π_C is followed from the properties of the functional $\phi(y, x)$ and strict monotonicity of the mapping J, and moreover, in the Hilbert spaces setting we have $\Pi_C = P_C$. It is obvious from the definition of the function ϕ that

$$(\|y\| - \|x\|)^2 \leq \phi(y, x) \leq (\|y\| + \|x\|)^2, \qquad \text{for all } x, y \in E. \tag{12}$$

Remark 1 If E is a strictly convex and a smooth Banach space, then for all $x, y \in E$, $\phi(y, x) = 0$ if and only if $x = y$, see Matsushita and Takahashi [9].

To obtain our results, following lemmas is of important.

Lemma 1 (Kamimura and Takahashi [16]). *Let E be a uniformly convex and smooth real Banach space and let $\{x_n\}$, $\{y_n\}$ be two sequences of E. If $\phi(x_n, y_n) \to 0$ and either $\{x_n\}$ or $\{y_n\}$ is bounded, then $\|x_n - y_n\| \to 0$.*

Lemma 2 (Kamimura and Takahashi [16]). *Let E be a uniformly convex and smooth Banach space and let $r > 0$. Then there exists a continuous, strictly increasing and convex function $g : [0, 2r] \to [0, \infty)$ such that $g(0) = 0$ and*

$$g(\|x - y\|) \leq \phi(x, y),$$

for all $x, y \in B_r = \{z \in E : \|z\| \leq r\}$.

Lemma 3 (Alber [15]). *Let E be a reflexive, strict convex and smooth real Banach space, let C be a nonempty closed convex subset of E and let $x \in E$. Then*

$$\phi(y, \Pi_C x) + \phi(\Pi_C x, x) \leq \phi(y, x), \qquad \forall y \in C. \tag{13}$$

For solving the equilibrium problem for a bifunction $f : C \times C \to \mathbf{R}$, let us assume that f satisfies the following conditions:

(A1) $f(x, x) = 0$ for all $x \in C$;
(A2) f is monotone, i.e., $f(x, y) + f(y, x) \leq 0$ for all $x, y \in C$;
(A3) for each $x, y, z \in C$,

$$\lim_{t \downarrow 0} f(tz + (1 - t)x, y) \leq f(x, y);$$

(A4) for each $x \in C$, $y \mapsto f(x, y)$ is convex and lower semi-continuous.

Lemma 4 (Blum and Oettli [17]). *Let C be a closed convex subset of a smooth, strictly convex, and reflexive Banach space E, let f be a bifunction from $C \times C$ to R satisfying $(A1)$-$(A4)$, and let $r > 0$ and $x \in E$. Then, there exists $z \in C$ such that*

$$f(z, y) + \frac{1}{r}\langle y - z, Jz - Jx\rangle \geq 0, \quad \forall y \in C.$$

The following lemma was also given by Combettes in [18].

Lemma 5 *Let C be a closed convex subset of a uniformly smooth, strictly convex, and reflexive Banach space E, and let f be a bifunction from $C \times C$ to $\boldsymbol{R}$ satisfying $(A1)$-$(A4)$. For $r > 0$ and $x \in E$, define a mapping $S_r : E \to C$ as follows:*

$$S_r x = \{z \in C : f(z, y) + \frac{1}{r}\langle y - z, Jz - Jx\rangle, \quad \forall y \in C\}$$

for all $x \in C$. Then the following hold:

1. S_r is single-valued;
2. S_r is a firmly nonexpansive-type mapping, i.e., for all $x, y \in E$, $\langle S_r x - S_r y, JS_r x - JS_r y\rangle \leq \langle S_r x - S_r y, Jx - Jy\rangle$
3. $F(S_r) = EP(f)$;
4. $EP(f)$ is closed and convex.

Lemma 6 (Takahashi and Zembayashi [19]). *Let C be a closed convex subset of a smooth, strictly convex, and reflexive Banach space E, let f be a bifunction from $C \times C$ to R satisfying $(A1)$-$(A4)$, and let $r > 0$. Then, for $x \in E$ and $q \in F(S_r)$,*

$$\phi(q, S_r x) + \phi(S_r x, x) \leq \phi(q, x).$$

3 Main Results

In this section, we propose the modified iterative scheme for solving fixed point problem for multivalued nonexpansive mappings and equilibrium problems. Its strong convergence theorem has been proved as shown in the following.

Theorem 2 *Let E be a real uniformly smooth and uniformly convex Banach space and let C be a nonempty closed convex subset of E. Let $T : C \to CB(C)$ be a closed and total quasi-ϕ-asymptotically nonexpansive multivalued mapping with nonnegative real sequence $\{v_n\}, \{\mu_n\}$ and a strictly increasing continuous function $\zeta : \boldsymbol{R}^+ \to \boldsymbol{R}^+$ such that $\mu_1 = 0$, $v_n \to 0$, $\mu_n \to 0$ as $n \to \infty$ and $\zeta(0) = 0$, let $t : C \to C$ be a relatively nonexpansive mapping, let $f : C \times C \to \boldsymbol{R}$ be a bifunction satisfying conditions (A1)-(A4) such that $F := F(T) \cap F(t) \cap EP(f)$ and let a sequence $\{x_n\}$ in C by the following algorithm:*

$$\begin{cases} x_0 \in C, \ \text{chosen arbitrarity and } \ C_0 = C, \\ u_n = S_{r_n} z_n, z_n \in T^n x_n \\ w_n = J^{-1}(\beta_n J x_n + (1-\beta_n) J u_n) \\ y_n = J^{-1}(\alpha_n J x_n + (1-\alpha_n) J t w_n), \\ C_{n+1} = \{z \in C_n : \phi(z, y_n) \leq \alpha_n \phi(z, x_n) + (1-\alpha_n)\phi(z, w_n) \\ \qquad\qquad \leq \gamma_n \phi(z, x_n) + (1-\gamma_n)\phi(p, z_n) \leq \phi(z, x_n) + \xi_n\}, \\ x_{n+1} = \Pi_{C_{n+1}} x_0, \end{cases} \tag{14}$$

for $n \in \boldsymbol{N} \cup \{0\}$, where J is the single-valued duality mapping on E and $\xi_n = v_n \sup_{u^* \in F} \zeta(\phi(u^*, x_n)) + \mu_n$, $\gamma_n = \alpha_n + \beta_n + \alpha_n \beta_n$. The coefficient sequence $\{\alpha_n\}, \{\beta_n\} \subset [0, 1]$ satisfying

(i) $0 < \beta_1 \leq \beta_n \leq \beta_2 < 1$,
(ii) $0 \leq \alpha_n \leq \alpha < 1$.

Then $\{x_n\}$ converges strongly to $\Pi_F x_0$, where Π_F is the generalized projection from C onto F.

Proof We first show that C_{n+1} is closed and convex for each $n \geq 0$. Obviously, from the definition of C_{n+1}, we see that C_{n+1} is closed for each $n \geq 0$. Now we show that C_{n+1} is convex for any $n \geq 0$. Since

$$\phi(z, y_n) \leq \phi(z, x_n) + \xi_n \Longleftrightarrow 2\langle v, J x_n - J y_n \rangle + \|y_n\|^2 - \|x_n\|^2 - \xi_n \leq 0,$$

this implies that C_{n+1} is a convex set.

Next, we show that $\{x_n\}$ is bounded and $\{\phi(x_n, x_0)\}$ is convergent sequence. Put $u_n = S_{r_n} z_n$, $\forall n \geq 0$, let $p \in F := F(T) \cap F(t) \cap EP(f)$. By Eq. (14), we obtain

$$\begin{aligned}
\phi(p, y_n) &= \phi(p, J^{-1}(\alpha_n Jx_n + (1-\alpha_n)Jtw_n)) \\
&\leq \|p\|^2 - 2\alpha_n\langle p, Jx_n\rangle - 2(1-\alpha_n)\langle p, Jtw_n\rangle + \alpha_n\|x_n\|^2 + (1-\alpha_n)\|tw_n\|^2 \\
&\quad -\alpha_n(1-\alpha_n)g\|Jx_n - Jtw_n\|^2 \\
&= \alpha_n\phi(p, x_n) + (1-\alpha_n)\phi(p, tw_n) - \alpha_n(1-\alpha_n)g\|Jx_n - Jtw_n\|^2 \\
&\leq \alpha_n\phi(p, x_n) + (1-\alpha_n)\phi(p, w_n) - \alpha_n(1-\alpha_n)g\|Jx_n - Jtw_n\|^2 \\
&\leq \alpha_n\phi(p, x_n) + (1-\alpha_n)\phi(p, w_n), \qquad (15)
\end{aligned}$$

and

$$\begin{aligned}
\phi(p, u_n) &= \phi(p, S_{r_n} z_n) \\
&\leq \phi(p, z_n) \\
&= \phi(p, T^n x_n) \\
&\leq \phi(p, x_n) + \nu_n\zeta(\phi(p, x_n)) + \mu_n \\
&= \phi(p, x_n) + \nu_n \sup_{u^* \in F} \zeta(\phi(u^*, x_n)) + \mu_n \\
&= \phi(p, x_n) + \xi_n, \qquad (16)
\end{aligned}$$

where $\xi_n = \nu_n \sup_{u^* \in F} \zeta(\phi(u^*, x_n)) + \mu_n$.
By the assumptions of $\{\nu_n\}$, $\{\mu_n\}$, we obtain

$$\xi_n = \nu_n \sup_{u^* \in F(T)} \zeta(\phi(u^*, x_n)) + \mu_n \to 0, \quad \text{as } n \to \infty. \qquad (17)$$

By Eqs. (14) and (16),

$$\begin{aligned}
\phi(p, w_n) &= \phi(p, J^{-1}(\beta_n Jx_n + (1-\beta_n)Ju_n)) \\
&\leq \|p\|^2 - 2\beta_n\langle p, Jx_n\rangle - 2(1-\beta_n)\langle p, Ju_n\rangle + \beta_n\|x_n\|^2 + (1-\beta_n)\|u_n\|^2 \\
&\quad -\beta_n(1-\beta_n)g\|Jx_n - Ju_n\|^2 \\
&= \beta_n\phi(p, x_n) + (1-\beta_n)\phi(p, u_n) - \beta_n(1-\beta_n)g\|Jx_n - Ju_n\|^2 \\
&= \beta_n\phi(p, x_n) + (1-\beta_n)\phi(p, z_n) - \beta_n(1-\beta_n)g\|Jx_n - Ju_n\|^2 \\
&\leq \beta_n\phi(p, x_n) + (1-\beta_n)[\phi(p, x_n) + \xi_n] - \beta_n(1-\beta_n)g\|Jx_n - Ju_n\|^2
\end{aligned}$$

$$\begin{aligned} &= \phi(p, x_n) + \xi_n - \beta_n(1-\beta_n)g\|Jx_n - Ju_n\|^2 \\ &= \phi(p, x_n) + \xi_n. \end{aligned} \tag{18}$$

Substituting Eqs. (18) into (15), we have

$$\begin{aligned} \phi(p, y_n) &\leq \alpha_n\phi(p, x_n) + (1-\alpha_n)[\beta_n\phi(p, x_n) + (1-\beta_n)\phi(p, z_n) \\ &\quad -\beta_n(1-\beta_n)g\|Jx_n - Ju_n\|^2] \\ &\leq \gamma_n\phi(p, x_n) + (1-\gamma_n)\phi(p, z_n) - (1-\alpha_n)\beta_n(1-\beta_n)g\|Jx_n - Ju_n\|^2 \\ &\leq \gamma_n\phi(p, x_n) + (1-\gamma_n)[\phi(p, x_n) + \xi_n] - (1-\alpha_n)\beta_n(1-\beta_n)g\|Jx_n - Ju_n\|^2 \\ &\leq \phi(p, x_n) + \xi_n - (1-\alpha_n)\beta_n(1-\beta_n)g\|Jx_n - Ju_n\|^2 \\ &\leq \phi(p, x_n) + \xi_n, \forall p \in F, \end{aligned} \tag{19}$$

where $\gamma_n = \alpha_n + \beta_n + \alpha_n\beta_n$. This means that, $p \in C_{n+1}$ for all $n \geq 0$. As consequently, the sequence $\{x_n\}$ is well defined. Moreover, since $x_n = \Pi_{C_n}x_0$ and $x_{n+1} \in C_{n+1} \subset C_n$, we get

$$\phi(x_n, x_0) \leq \phi(x_{n+1}, x_0),$$

for all $n \geq 0$. Therefore, $\{\phi(x_n, x_0)\}$ is nondecreasing. By definition of x_n and Lemma 3, we have

$$\phi(x_n, x_0) = \phi(\Pi_{C_n}x_0, x_0) \leq \phi(p, x_0) - \phi(p, \Pi_{C_n}x_0) \leq \phi(p, x_0), \tag{20}$$

for all $p \in \bigcap_{n=0}^{\infty} F(T_n) \subset C_n$. Thus, $\{\phi(x_n, x_0)\}$ is a bounded sequence. Moreover, by Eq. (12), we know that $\{x_n\}$ is bounded. So, $\lim_{n\to\infty}\phi(x_n, x_0)$ exists. Again, by Lemma 3, we have

$$\begin{aligned} \phi(x_{n+1}, x_n) &= \phi(x_{n+1}, \Pi_{C_n}x_0) \\ &\leq \phi(x_{n+1}, x_0) - \phi(\Pi_{C_n}x_0, x_0) \\ &= \phi(x_{n+1}, x_0) - \phi(x_n, x_0), \end{aligned}$$

for all $n \geq 0$. Thus, $\phi(x_{n+1}, x_n) \to 0$ as $n \to \infty$. Since $\{x_n\}$ is bounded and x is reflexive, there exists a subsequence $\{x_{n_i}\} \subset \{x_n\}$ such that $x_{n_i} \rightharpoonup q \in C$.

From C_n is closed and convex and $C_{n+1} \subset C_n$, this implies that C_n is weakly closed and $q \in C_n$, for each $n \geq 0$. In view of $x_{n_i} = \Pi_{C_{n_i}}x_0$, we have $\phi(x_{n_i}, x_0) \leq \phi(q, x_0)$, $\forall n_i \geq 0$. Since the norm $\|\cdot\|$ is weakly lower semi-continuous, we have

$$\begin{aligned} \liminf_{n_i\to\infty}\phi(x_{n_i}, x_0) &= \liminf_{n_i\to\infty}\|x_{n_i}\|^2 - 2\langle x_{n_i}, Jx_0\rangle + \|x_0\|^2 \\ &\geq \|q\|^2 - 2\langle q, Jx_0\rangle + \|x_0\|^2 \\ &= \phi(q, x_0). \end{aligned}$$

So, $\phi(q, x_0) \leq \liminf_{n_i\to\infty}\phi(x_{n_i}, x_0) \leq \limsup_{n_i\to\infty}\phi(x_{n_i}, x_0) \leq \phi(q, x_0)$. This implies that $\lim_{n_i\to\infty}\phi(x_{n_i}, x_0) = \phi(q, x_0)$. By Kadec-Klee property of E, we have $\lim_{n_i\to\infty}x_{n_i} = q$, as $n_i \to \infty$.

Since the sequence $\{\phi(x_n, x_0)\}$ is convergent and $\lim_{n_i\to\infty}\phi(x_{n_i}, x_0) = \phi(q, x_0)$, which implies that $\lim_{n\to\infty}\phi(x_n, x_0) = \phi(q, x_0)$. If there exists some subsequence $\{x_{n_j}\} \subset \{x_n\}$ such that $x_{n_j} \rightharpoonup q^*$, then from Lemma 3

$$\begin{aligned}\phi(q, q^*) &= \lim_{n_i,n_j\to\infty}\phi(x_{n_i}, x_{n_j})\\ &= \lim_{n_i,n_j\to\infty}\phi(x_{n_i}, \Pi_{C_{n_j}}x_0)\\ &\le \lim_{n_i,n_j\to\infty}[\phi(x_{n_i}, x_0) - \phi(\Pi_{C_{n_j}}x_0, x_0)]\\ &= \lim_{n_i,n_j\to\infty}[\phi(x_{n_i}, x_0) - \phi(x_{n_j}x_0, x_0)]\\ &= \phi(q, x_0) - \phi(q, x_0) = 0.\end{aligned}$$

This implies that $q = q^*$ and so

$$\lim_{n\to\infty} x_n = q. \tag{21}$$

By definition of $\Pi_{C_n}x_0$, we have

$$\begin{aligned}\phi(x_{n+1}, x_n) &= \phi(x_{n+1}, \Pi_{C_n}x_0)\\ &\le \phi(x_{n+1}, x_0) - \phi(\Pi_{C_n}x_0, x_0)\\ &= \phi(x_{n+1}, x_0) - \phi(x_n, x_0).\end{aligned} \tag{22}$$

From $\lim_{n\to\infty}\phi(x_n, x_0)$ exists, we have

$$\lim_{n\to\infty}\phi(x_{n+1}, x_n) = 0. \tag{23}$$

It follows from Lemma 1, we get

$$\lim_{n\to\infty}\|x_{n+1} - x_n\| = 0. \tag{24}$$

By definition of C_n and $x_{n+1} = \Pi_{C_{n+1}}x_0 \in C_{n+1}$, we have $\phi(x_{n+1}, y_n) \le \phi(x_{n+1}, x_n) + \xi_n$. It follows from Eq. (23) and $\xi_n \to 0$ as $n \to \infty$ that $\lim_{n\to\infty}\phi(x_{n+1}, y_n) = 0$. Again from Lemma 1, we have

$$\lim_{n\to\infty}\|x_{n+1} - y_n\| = 0. \tag{25}$$

By Eqs. (24) and (25), we also have $\lim_{n\to\infty}\|y_n - x_n\| = 0$. Since J is uniformly norm-to-norm continuous, we obtain

$$\lim_{n\to\infty}\|Jy_n - Jx_n\| = 0. \tag{26}$$

From Eq. (19), for $u^* \in F(T)$ and $z_n \in T^n x_n$, we have

$$\phi(p, y_n) \leq \phi(p, x_n) + \xi_n - (1-\alpha_n)\beta_n(1-\beta_n)g\|Jx_n - Ju_n\|^2, \tag{27}$$

and hence

$$(1-\alpha_n)\beta_n(1-\beta_n)g\|Jx_n - Ju_n\|^2 \leq \phi(p, x_n) - \phi(p, y_n) + \xi_n. \tag{28}$$

On the other hand, we note that

$$\begin{aligned}\phi(p, x_n) - \phi(p, y_n) &= \|x_n\|^2 - \|y_n\|^2 - 2\langle p, Jx_n - Jy_n\rangle \\ &\leq \|x_n - y_n\|(\|x_n + y_n\|) + 2\|p\|\|Jx_n - Jy_n\|. \end{aligned}\tag{29}$$

It follows from $\|x_n - y_n\| \to 0$ and $\|Jx_n - Jy_n\| \to 0$, that

$$\phi(p, x_n) - \phi(p, y_n) \to 0, \quad \text{as } n \to \infty. \tag{30}$$

Since condition (i), (ii), Eqs. (17) and (30), it follows from Eq. (28) $g\|Jx_n - Ju_n\| \to 0$, as $n \to \infty$. If follows from the property of g that $\lim_{n\to\infty}\|Jx_n - Ju_n\| = 0$. Since J^{-1} is uniformly norm-to-norm continuous, we have

$$\lim_{n\to\infty}\|x_n - u_n\| = 0. \tag{31}$$

From definition of C_n, we obtain

$$\gamma_n\phi(z, x_n) + (1-\gamma_n)\phi(z, z_n) \leq \phi(z, x_n) + \xi_n \Leftrightarrow \phi(z, z_n) \leq \phi(z, x_n) + \xi_n.$$

Since $x_{n+1} = \Pi_{C_{n+1}}x_0 \in C_{n+1}$, we have $\phi(x_{n+1}, z_n) \leq \phi(x_{n+1}, x_n) + \xi_n$. It follows from Eqs. (17) and (23) that $\lim_{n\to\infty}\phi(x_{n+1}, z_n) = 0$. From Lemma 1, we get

$$\lim_{n\to\infty}\|x_{n+1} - z_n\| = 0. \tag{32}$$

By using the triangle inequality, we have

$$\|u_n - z_n\| \leq \|u_n - x_n\| + \|x_n - x_{n+1}\| + \|x_{n+1} - z_n\|.$$

By Eqs. (24), (31) and (32), we also have

$$\lim_{n\to\infty}\|u_n - z_n\| = 0. \tag{33}$$

Since J is uniformly norm-to-norm continuous, we have

$$\lim_{n\to\infty}\|Ju_n - Jz_n\| = 0. \tag{34}$$

From definition of C_n, we have

$$\alpha_n\phi(z, x_n)+(1-\alpha_n)\phi(z, w_n) \leq \phi(z, x_n)+\xi_n \Leftrightarrow \phi(z, w_n) \leq \phi(z, x_n)+\xi_n. \quad (35)$$

Since $x_{n+1} = \Pi_{C_{n+1}} x_0 \in C_{n+1}$, we have

$$\phi(x_{n+1}, w_n) \leq \phi(x_{n+1}, x_n) + \xi_n. \quad (36)$$

It follows from Eq. (23) and $\xi_n \to 0$ as $n \to \infty$ that $\lim_{n\to\infty} \phi(x_{n+1}, w_n) = 0$.

From Lemma 1, we have

$$\lim_{n\to\infty} \|x_{n+1} - w_n\| = 0. \quad (37)$$

By using the triangle inequality, we get

$$\|w_n - x_n\| \leq \|w_n - x_{n+1}\| + \|x_{n+1} - x_n\|, \quad (38)$$

again by Eqs. (24) and (37), we also have

$$\lim_{n\to\infty} \|w_n - x_n\| = 0. \quad (39)$$

From Eqs. (15) and (18), that

$$\begin{aligned}\phi(p, y_n) &\leq \alpha_n\phi(p, x_n) + (1 - \alpha_n)[\phi(p, x_n) + \xi_n] - \alpha_n(1 - \alpha_n)g\|Jx_n - Jtw_n\| \\ &\leq \phi(p, x_n) + \xi_n - \alpha_n(1 - \alpha_n)g\|Jx_n - Jtw_n\|, \end{aligned} \quad (40)$$

and hence

$$\alpha_n(1 - \alpha_n)g\|Jx_n - Jtw_n\| \leq \phi(p, x_n) - \phi(p, y_n) + \xi_n. \quad (41)$$

By condition (ii), Eqs. (17) and (30), we obtain that $g\|Jx_n - Jtw_n\| \to 0$, as $n \to \infty$. If follows from the property of g that $\|Jx_n - Jtw_n\| \to 0$, as $n \to \infty$. Since J^{-1} is uniformly norm-to-norm continuous, we have

$$\|x_n - tw_n\| \to 0, \text{ as } n \to \infty. \quad (42)$$

By Eqs. (39) and (42), we have $\lim_{n\to\infty} \|w_n - tw_n\| = 0$. Since $\{x_n\}$ is bounded and $x_{n_i} \rightharpoonup q \in C$. It follows from Eq. (39), we have $w_{n_i} \rightharpoonup q$ as $i \to \infty$ and t be relatively nonexpansive. We have that $q \in \widehat{F}(t) = F(t)$. By Eqs. (31) and (33), we have $\lim_{n\to\infty} \|x_n - z_n\| = 0$. From Eq. (21), we get

$$z_n \to q \text{ as } n \to \infty. \quad (43)$$

From $u_n \in T^n x_n$ and let $\{s_n\}$ be a sequence generate by

$$\begin{cases} s_2 \in Tz_1 \subset T^2 x_1; \\ s_3 \in Tz_2 \subset T^3 x_2; \\ s_4 \in Tz_3 \subset T^4 x_3; \\ \vdots \\ s_n \in Tz_{n-1} \subset T^n x_{n-1}; \\ s_{n+1} \in Tz_n \subset T^{n+1} x_n; \\ \vdots \end{cases} \tag{44}$$

By the assumption that T is uniformly L-Lipschitz continuous and any $z_n \in T^n x_n$ and $a_{n+1} \in Tz_n \subset T^{n+1} x_n$, we have

$$\begin{aligned} \|s_{n+1} - z_n\| &\leq \|s_{n+1} - z_{n+1}\| + \|z_{n+1} - x_{n+1}\| + \|x_{n+1} - x_n\| + \|x_n - z_n\| \\ &\leq L\|x_n - x_{n+1}\| + \|z_{n+1} - x_{n+1}\| + \|x_{n+1} - x_n\| + \|x_n - z_n\| \\ &\leq (L+1)\|x_n - x_{n+1}\| + \|z_{n+1} - x_{n+1}\| + \|x_n - z_n\|. \end{aligned}$$

From Eqs. (21), (24) and (43) that

$$\lim_{n\to\infty} \|s_{n+1} - z_n\| = 0 \text{ and } \lim_{n\to\infty} s_{n+1} = q. \tag{45}$$

In view of closeness of T, it yields that $q \in Tq$. Therefore $q \in F(T)$.

By Eq. (34), where $r_n > 0$

$$\lim_{n\to\infty} \frac{\|J(u_n) - J(z_n)\|}{r_n} = 0. \tag{46}$$

Also, we obtain

$$f(u_n, y) + \frac{1}{r_n}\langle y - u_n, J(u_n) - J(z_n)\rangle \geq 0, \forall y \in C.$$

Hence,

$$f(u_{n_i}, y) + \frac{1}{r_{n_i}}\langle y - u_{n_i}, Ju_{n_i} - Jz_{n_i}\rangle \geq 0, \quad \forall y \in C.$$

From the (A2), we note that

$$\|y - u_{n_i}\|\frac{\|Ju_{n_i} - Jz_{n_i}\|}{r_{n_i}} \geq \frac{1}{r_{n_i}}\langle y - u_{n_i}, Ju_{n_i} - Jz_{n_i}\rangle \geq -f(u_{n_i}, y) \geq f(y, u_{n_i}), \quad \forall y \in C.$$

Taking the limit as $n \to \infty$ in above inequality and from (A4) and $u_{n_i} \rightharpoonup q$, we have $f(y, q) \leq 0, \quad \forall y \in C$. For $0 < t < 1$ and $y \in C$, define $y_t = ty + (1-t)q$. Noticing that $y, q \in C$, we obtains $y_t \in C$, which yields that $f(y_t, q) \leq 0$. It follows from (A1) that

$$0 = f(y_t, y_t) \leq tf(y_t, y) + (1-t)f(y_t, \hat{x}) \leq tf(y_t, y).$$

That is, $f(y_t, y) \geq 0$.

Let $t \downarrow 0$, from (A3), we obtain $f(q, y) \geq 0, \quad \forall y \in C$. Therefore $q \in EP(f)$. Hence, $q \in F(T) \cap F(t) \cap EP(f)$.

Finally, we show that $x_n \to q = \Pi_F x_0$. Let $p^* = \Pi_F x_0$. Since $p^* \in F \subset C_n$ and $x_n = \Pi_{C_n} x_0$, we have

$$\phi(x_n, x_0) \leq \phi(p^*, x_0), \forall n \geq 0.$$

This implies that

$$\phi(q, x_0) = \lim_{n\to\infty} \phi(x_n, x_0) \leq \phi(p^*, x_0). \tag{47}$$

In view of definition of $\Pi_F x_0$, we have $q = p^*$. Therefore, $x_n \to q = \Pi_F x_0$. This completes the proof. □

Corollary 1 *Let E be a real uniformly smooth and uniformly convex Banach space and let C be a nonempty closed convex subset of E. Let $T : C \to CB(C)$ be a closed and total quasi-ϕ-asymptotically nonexpansive multivalued mapping with nonnegative real sequence $\{v_n\}, \{\mu_n\}$ and a strictly increasing continuous function $\zeta : \boldsymbol{R}^+ \to \boldsymbol{R}^+$ such that $\mu_1 = 0$, $v_n \to 0$, $\mu_n \to 0$ as $n \to \infty$ and $\zeta(0) = 0$, let $t : C \to C$ be a relatively nonexpansive mapping such that $F := F(T) \cap F(t)$ and let a sequence $\{x_n\}$ in C by the following algorithm:*

$$\begin{cases} x_0 \in C, \ \text{chosen arbitrarity and } C_0 = C, \\ w_n = J^{-1}(\beta_n J x_n + (1-\beta_n) J z_n), z_n \in T^n x_n \\ y_n = J^{-1}(\alpha_n J x_n + (1-\alpha_n) J t w_n), \\ C_{n+1} = \{z \in C_n : \phi(z, y_n) \leq \alpha_n \phi(z, x_n) + (1-\alpha_n)\phi(z, w_n) \\ \qquad\qquad \leq \gamma_n \phi(z, x_n) + (1-\gamma_n)\phi(p, z_n) \leq \phi(z, x_n) + \xi_n\}, \\ x_{n+1} = \Pi_{C_{n+1}} x_0, \end{cases} \tag{48}$$

for $n \in \boldsymbol{N} \cup \{0\}$, where J is the single-valued duality mapping on E and $\xi_n = v_n \sup_{u^* \in F} \zeta(\phi(u^*, x_n)) + \mu_n$, $\gamma_n = \alpha_n + \beta_n + \alpha_n\beta_n$. The coefficient sequence $\{\alpha_n\}, \{\beta_n\} \subset [0, 1]$ satisfying

(i) $0 < \beta_1 \leq \beta_n \leq \beta_2 < 1$,
(ii) $0 \leq \alpha_n \leq \alpha < 1$.

Then $\{x_n\}$ converges strongly to $\Pi_F x_0$, where Π_F is the generalized projection from C onto F.

4 Conclusion

In this research, we are motivated by Tang and Chang [11] and Wattanawitoon et al. [12], we have extended and combined their schemes to be more general and improved by changing the mentioned problem from zero of maximal monotone operators to equilibrium problem which will be more effective tools for solving the problems in other related disciplines of sciences. Furthermore, the results on investigating the convergence of the sequence defined by our scheme was proved and got the strong convergence under some mild conditions.

Acknowledgments The authors would like to thank the referees for the valuable suggestions which helped to improve this manuscript. K. Wattanawitoon gratefully acknowledges support provided by the King Mongkut's University of Technology Thonburi (KMUTT) during the second author's stay at the King Mongkut's University of Technology Thonburi (KMUTT) as a post doctoral fellow (KMUTT-Post-doctoral Fellowship).

References

1. Reich S (1996) A weak convergence theorem for the alternating method with Breman distance. In: Kaetsatos AG (ed) Theory and applications of nonlinear operators of accretive and monotone type. Marcel Dekker, New York, pp 313–318
2. Nilsrakoo W, Saejung S (2008) Strong convergence to common fixed points of countable relatively quasi-nonexpansive mappings. Fixed Point Theor Appl 2008(312454):19
3. Su Y, Wang D (2008) Strong convergence of monotone hybrid algorithm for hemi-relatively nonexpansive mappings. Fixed Point Theor Appl 2008(284613):8
4. Zegeye H, Shahzad N (2009) Strong convergence for monotone mappings and relatively weak nonexpansive mappings. Nonlinear Anal 70:2707–2716
5. Butnariu D, Reich S, Zaslavski AJ (2001) Asymptotic behavior of relatively nonexpansive operators in Banach spaces. J Math Anal Appl 7(2):151–174
6. Cens Y, Reich S (1996) Iterations of paracontractions and firmly nonexpansive operators with applications to feasibility and optimization. Optimization 37:323–339
7. Takahashi W, Takeuchi Y, Kubota R (2008) Strong convergence theorems by hybrid methods for families of nonexpansive mappings in Hilbert spaces. J Math Anal Appl 341:276–286
8. Matsushita S, Takahashi W (2004) Weakly and strong convergence theorems for relatively nonexpansive mappings in a Banach space. Fixed Point Theor Appl 2004:37–47
9. Matsushita S, Takahashi W (2005) A Strong convergence theorem for relatively nonexpansive mappings in a Banach space. J Approximation Theor 134(2):257–266
10. Chang SS, Lee HWJ, Chan CK, Zhang WB (2012) A modified Halpern-type iteration algorithm for totally quasi-ϕ-asymptotically nonexpansive mappings with applications. Appl Math Comput 218:6489–6497
11. Tang J, Chang SS (2012) Strong convergence theorems for total quasi-ϕ-asymptotically nonexpansive multi-value mappings in Banach spaces. Fixed Point Theor Appl 2012:63. doi:10.1186/1687-1812-2012-63
12. Wattanawitoon K, Witthayarat U, Kumam P (2013) Strong convergence theorems of multivalued nonexpansive mappings and maximal monotone operators in banach spaces. Lecture Notes in Engineering and Computer Science: Proceedings of The International MultiConference of Engineers and Computer Scientists 2013, Hong Kong, 13–15 Mar 2013, pp 1194–1199
13. Cioranescu I (1990) Geometry of banach spaces, duality mappings and nonlinear problems of mathematics and its applications. Kluwer Academic Publishers, Dordrecht, The Netherlands

14. Takahashi W (2000) Nonlinear functional analysis, fixed point theory and its applications. Yokohama Publishers, Yokohama, Japan
15. Alber YI (1996) Metric and generalized projection operators in Banach spaces: properties and applications, theory and applications of nonlinear operators of accretive and monotone type. In: Kartsatos AG (ed) Marcel Dekker, New York, vol 178, pp 15–50
16. Kamimura S, Takahashi W (2002) Strong convergence of a proximal-type algorithm in a Banach space. SIAM J Optim 13(3):938–945
17. Blum E, Oettli W (1994) From optimization and variational inequalities to equilibrium problems. Math Stud 63:123–145
18. Combettes PL, Hirstoaga SA (2005) Equilibrium programming in Hilbert spaces. J Nonlinear Convex Anal 6:117–136
19. Takahashi W, Zembayashi K (2008) Strong convergence theorems by a new hybrid method for equilibrium problems and relatively nonexpansive mappings. Fixed Point Theor Appl 2008(528476):11

Counting the Number of Multi-player Partizan Cold Games Born by Day d

Alessandro Cincotti

Abstract In combinatorial games, few results are known about the overall structure of multi-player games. Recently, it has been proved that multi-player games lasting at most d moves, also known as the games born by day d, forms a completely distributive lattice with respect to every partial order relation $\leq_C$, where C is an arbitrary coalition of players. In this paper, we continue our investigation concerning multi-player games and, using the strings of multi-player Hackenbush in order to construct multi-player cold games, we calculate lower and upper bounds on $S_n[d]$ equal to the number of n-player partizan cold games born by day d. In particular, we prove that if n is fixed, then $S_n[d] \in \Theta(2nd)$, but in order to establish the exact value of $S_n[d]$ further efforts are necessary.

Keywords Cold game · Combinatorial game · game · Multi-player Hackenbush · Partizan game · Surreal number

1 Introduction

Combinatorial game theory [1] is a branch of mathematics devoted to studying the optimal strategy in perfect-information games with no chance moves where typically two players move alternately. Such a theory is based on a straightforward and intuitive recursive definition of games, which yields a rich algebraic structure. Games can be added and subtracted in a natural way, forming a commutative group with a partial order. There is a special sub-group of games called *numbers* which can also be multiplied and which form a field with a total order.

A. Cincotti (✉)
School of Information Science, Japan Advanced Institute of Science and Technology,
1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
e-mail: cincotti@jaist.ac.jp

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 275, DOI: 10.1007/978-94-007-7684-5_21,

When combinatorial game theory is generalized to multi-player games, the problem of coalition arises. A coalition makes it hard to have a simple game value in any additive algebraic structure. To circumvent the coalition problem in multi-player games, different approaches have been proposed [12–14, 16] with various restrictive assumptions about the rationality of one's opponents and the formation and behavior of coalitions.

Alternatively, Propp [15] adopts in his work an agnostic attitude toward such issues, and seeks only to understand in what circumstances one player has a winning strategy against the combined forces of the others. Cincotti [5–10] presents a theory to classify and analyze multi-player partizan games and their algebraic structure adopting the same attitude. Such a theory is an extension of Conway's theory of partizan games [11] and, as a consequence, it is both a theory of games and a theory of numbers. Multi-player partizan games can be added in a natural way, forming a commutative monoid and including a special sub-monoid called cold games or *numbers*. For further details about multi-player partizan games, please refer to [5].

The number of surreal numbers born by day d is $S_2[d] = 2^{d+1} - 1$ and lower and upper bounds on two-player games are given by Wolfe and Fraser in [17]. Moreover, weak lower and upper bounds on three-player partizan cold games are given in [3].

The article is organized as follows. In Sect. 2, we give a practical rule to calculate the value of a generic string of multi-player Hackenbush. In Sect. 3, we present theorems to construct multi-player cold games using the strings of multi-player Hackenbush and we give lower and upper bounds on the number of multi-player partizan cold games born by day d. Finally, in Sect. 4, we present our conclusion.

2 Multi-Player Hackenbush

Blue–Red Hackenbush is a classic combinatorial game. Every instance of this game is represented by an undirected graph such that:

- Every edge is connected via a chain of edges to a certain line called the *ground*, and
- Every edge is colored either blue or red.

Two players, called Left and Right, move alternately. Left moves by deleting any blue edge together with all the edges that are no longer connected to the ground and Right moves by deleting any red edge together with all the edges that are no longer connected to the ground. The first player unable to move because there are no edges of his/her color is the loser.

When Blue–Red Hackenbush is played on strings, it is easily solvable using Berlekamp's rule [2] or Thea van Roode's rule [1], but to determine the value of a Blue–Red Hackenbush position on a general graph is NP-hard [1].

Multi-player Hackenbush is the n-player version of Blue–Red Hackenbush. Every instance of n-player Hackenbush is represented by an undirected graph such that:

- Every edge is connected via a chain of edges to a certain line called the *ground*, and
- Every edge is labeled by an integer $j \in \{1, 2, \ldots, n\}$.

The first player moves by deleting any edge labeled 1 together with all the edges that are no longer connected to the ground, the second player moves by deleting any edge labeled 2 together with all the edges that are no longer connected to the ground, and so on. Players take turns making legal moves in cyclic fashion (1st, 2nd, . . ., nth, 1st, 2nd, . . .). When one of the n players is unable to move, then that player leaves the game and the remaining $n-1$ players continue playing in the same mutual order as before. The remaining player is the winner.

Multi-player Hackenbush played on strings is PSPACE-complete [4].

Theorem 1 *Let $g = \{g_1|g_2|\ldots|g_n\}$ be a general string of n-player Hackenbush. Then, g is a number.*

Proof By the induction hypothesis, $g_1, g_2, \ldots, g_n$ are numbers; moreover, for every pair of options g_i and g_j, we can distinguish two different sub-cases:

- If the edge removed by the option g_i is over the edge removed by the option g_j, then g_j is one of the jth options of g_i.
- If the edge removed by the option g_i is under the edge removed by the option g_j, then g_i is one of the ith options of g_j.

Therefore, in both cases we have $g_i <_i g_j$. □

Theorem 2 *Let g be a general string of n-player Hackenbush and let h_i be a string of Blue-Red Hackenbush obtained by g where the edges labeled i has been painted blue and all the other edges have been painted red. Then, $\pi_i(g) = h_i$, with $i \in \{1, 2, \ldots, n\}$.*

Proof By the induction hypothesis, for each option $\pi_i(g_i)$ there exists an option h_i^L such that $\pi_i(g_i) = h_i^L$ and vice versa. Moreover, for each option $\pi_i(g_j)$ there exists an option h_i^R such that $\pi_i(g_j) = h_i^R$, $j \neq i$ and vice versa.

It follows $\pi_i(g) = \{\pi_i(g_i)|\pi_i(g_j), j \neq i\} = \{h_i^L|h_i^R\} = h_i$. □

The previous theorem gives us a practical way to calculate the n-tuple of surreal numbers $(\pi_1(g), \pi_2(g), \ldots, \pi_n(g))$ corresponding to a general string g of n-player Hackenbush.

3 How Many Numbers Born by Day d?

We recall that in S_2 the canonical forms of a positive integer and a positive dyadic fraction are respectively

$$n = \{n-1|\,\}$$
$$\frac{2r+1}{2^{s+1}} = \left\{ \frac{r}{2^s} \middle| \frac{r+1}{2^s} \right\}$$

Moreover, the negative of a number is defined as $-g = \{-g^R| - g^L\}$.

Theorem 3 *Let* $(0, 0, x_3, \ldots, x_n)$ *be a n-tuple of integer surreal numbers such that* $x_i \in S_2[d], d > n-2, x_i < -n+2, \forall i \in \{3, \ldots, n\}$. *Then, there exists* $g \in S_n[d]$ *such that*

$$\pi_1(g) = 0,$$
$$\pi_2(g) = 0,$$
$$\pi_i(g) = x_i, \forall i \in \{3, \ldots, n\}.$$

Proof Let us consider the game $g = \{a_3, \ldots, a_n| \,|\ldots|\,\}$ where a_3, …, a_n are strings of n-player Hackenbush, as shown in Fig. 1, such that

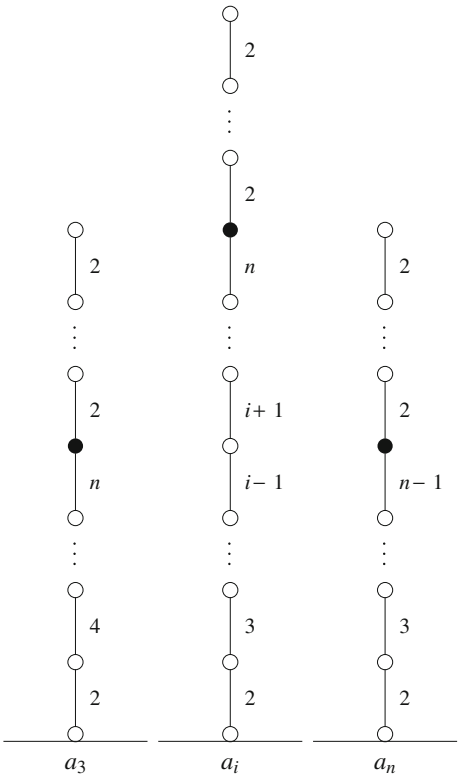


Fig. 1 The strings used to create the number corresponding to the tuple $(0, 0, x_3, \ldots, x_n)$. The *black vertex* indicates the minimum length of the string from the ground

$$\pi_1(a_i) \leq -n+2, \forall i \in \{3,\ldots,n\}$$
$$\pi_2(a_i) \geq 1/2^{n-3}, \forall i \in \{3,\ldots,n\}$$
$$\pi_i(a_i) = x_i^R \leq -n+2, \forall i \in \{3,\ldots,n\}$$
$$\pi_i(a_j) > -n+3, \forall i,j \in \{3,\ldots,n\}, i \neq j$$

We observe that $a_i \in S_n[d-1], \forall i \in \{3,\ldots,n\}$ therefore, $g \in S_n[d]$. In particular,

$$\pi_1(g) = \{\pi_1(a_3),\ldots,\pi_1(a_n)|\ \} = 0$$
$$\pi_2(g) = \{\ |\pi_2(a_3),\ldots,\pi_2(a_n)\} = 0$$
$$\pi_i(g) = \{\ |\pi_i(a_3),\ldots,\pi_i(a_n)\} = \{\ |\pi_i(a_i)\} = \{\ |x_i^R\} = x_i$$

This concludes the proof. □

Example 1 Let us define the number $g = \{a_3, a_4, a_5|\ |\ldots|\ \}$ corresponding to the tuple $(0, 0, x_3, x_4, x_5)$ where

$$x_3 = \{\ |-3\} = -4$$
$$x_4 = \{\ |-5\} = -6$$
$$x_5 = \{\ |-4\} - 5$$

The strings a_3, a_4, a_5 are shown in Fig. 2. We observe that

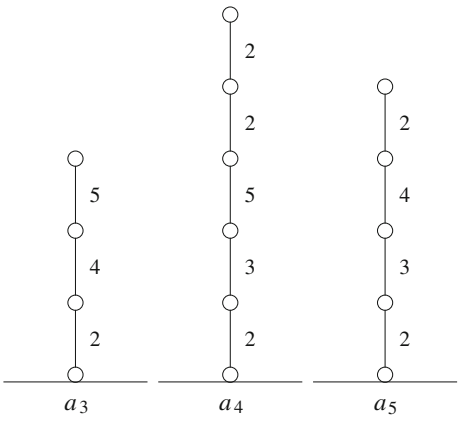


Fig. 2 The strings used to create the number corresponding to the tuple $(0, 0, -4, -6, -5)$

$$\pi_1(g) = \{\pi_1(a_3), \pi_1(a_4), \pi_1(a_5)|\ \} = \{-3, -5, -4|\ \} = 0$$
$$\pi_2(g) = \{\ |\pi_2(a_3), \pi_2(a_4), \pi_2(a_5)\} = \{\ |1/4, 1/16, 1/8\} = 0$$
$$\pi_3(g) = \{\ |\pi_3(a_3), \pi_3(a_4), \pi_3(a_5)\} = \{\ |-3, -15/16, -7/8\} = -4$$
$$\pi_4(g) = \{\ |\pi_4(a_3), \pi_4(a_4), \pi_4(a_5)\} = \{\ |-3/4, -5, -7/4\} = -6$$
$$\pi_5(g) = \{\ |\pi_5(a_3), \pi_5(a_4), \pi_5(a_5)\} = \{\ |-3/2, -15/8, -4\} = -5$$

Corollary 1 *The number of games $g \in S_n[d]$ such that $g = 1, 2\,0$ and $d > n - 2$ is at least $(d - n + 2)^{n-2}$.*

Theorem 4 *Let $(0, x_2, x_3, \ldots, x_n)$ be a n-tuple of surreal numbers such that $x_i \in S_2[d], d > n - 2, x_i < -n + 2, \forall i \in \{2, 3, \ldots, n\}$. Then, there exists $g \in S_n[d]$ such that*

$$\pi_1(g) = 0,$$
$$\pi_i(g) = x_i, \forall i \in \{2, \ldots, n\}.$$

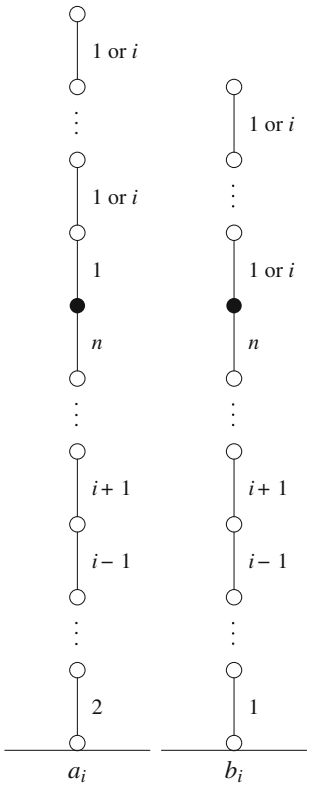


Fig. 3 The strings used to create the number corresponding to the tuple $(0, x_2, x_3, \ldots, x_n)$. The *black vertex* indicates the minimum length of the string from the ground

Proof Let us consider the game $g = \{a_2, a_3, \ldots, a_n|b_2|b_3|\ldots|b_n\}$ where a_i and b_i are two strings of n-player Hackenbush, as shown in Fig. 3, such that

$$\begin{aligned}
&\pi_1(a_i) < 0, \forall i \in \{2, \ldots, n\}\\
&\pi_1(b_i) > 0, \forall i \in \{2, \ldots, n\}\\
&\pi_i(a_i) = x_i^R \leq -n+2, \forall i \in \{2, \ldots, n\}\\
&\pi_i(a_j) > -n+3, \forall i, j \in \{2, \ldots, n\}, i \neq j\\
&\pi_i(b_i) = x_i^L < -n+2, \forall i \in \{2, \ldots, n\}\\
&\pi_i(b_j) > -n+2, \forall i, j \in \{2, \ldots, n\}, i \neq j
\end{aligned}$$

We observe that $a_i, b_j \in S_n[d-1], \forall i, j \in \{2, \ldots, n\}$; moreover, none of the following inequalities hold

$$\begin{aligned}
&a_i \geq_1 b_j, \quad \forall i, j \in \{2, \ldots, n\}\\
&b_i \geq_i a_j, \quad \forall i, j \in \{2, \ldots, n\}\\
&b_i \geq_i b_j, \quad \forall i, j \in \{2, \ldots, n\}, i \neq j
\end{aligned}$$

therefore $g \in S_n[d]$. In particular,

$$\begin{aligned}
\pi_1(g) &= \{\pi_1(a_2), \ldots, \pi_1(a_n)|\pi_1(b_2), \ldots, \pi_1(b_n)\} = 0\\
\pi_i(g) &= \{\pi_i(b_i)|\pi_i(a_2), \ldots, \pi_i(a_n), \pi_i(b_2), \ldots, \pi_i(b_{i-1}), \pi_i(b_{i+1}), \ldots, \pi_i(b_n)\}\\
&= \{\pi_i(b_i)|\pi_i(a_i)\} = \{x_i^L|x_i^R\} = x_i
\end{aligned}$$

We observe that if x_i is an integer, then the string b_i does not exist, i.e., $G_i = \emptyset$. □

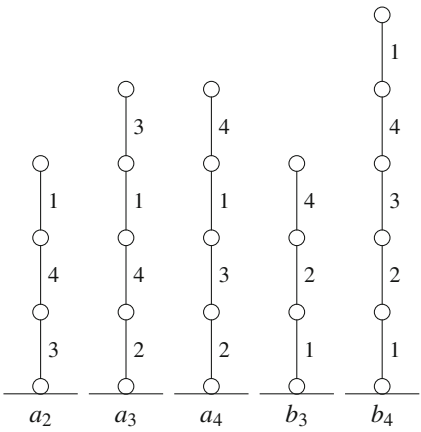


Fig. 4 The strings used to create the number corresponding to the tuple $(0, -4, -11/4, -21/8)$

Example 2 Let us define the number $g = \{a_2, a_3, a_4|b_2|b_3|b_4\}$ corresponding to the tuple $(0, x_2, x_3, x_4)$ where

$$\begin{aligned} x_2 &= \{\,|-3\} = -4 \\ x_3 &= \{-3|-5/2\} = -11/4 \\ x_4 &= \{-11/4|-5/2\} - 21/8 \end{aligned}$$

The strings a_2, a_3, a_4, b_3, and b_4 are shown in Fig. 4, the string b_2 does not exist because x_2 is an integer. We observe that

$$\begin{aligned} \pi_1(g) &= \{\pi_1(a_2), \pi_1(a_3), \pi_1(a_4)|\pi_1(b_3), \pi_1(b_4)\} \\ &= \{-3/2, -7/4, -7/4|1/4, 3/16\} = 0 \\ \pi_2(g) &= \{\,|\pi_2(a_2), \pi_2(a_3), \pi_2(a_4), \pi_2(b_3), \pi_2(b_4)\} \\ &= \{\,|-3, 1/8, 1/8, -3/4, -15/16\} = -4 \\ \pi_3(g) &= \{\pi_3(b_3)|\pi_3(a_2), \pi_3(a_3), \pi_3(a_4), \pi_3(b_4)\} \\ &= \{-3|1/4, -5/2, -7/8, -15/8\} = -11/4 \\ \pi_4(g) &= \{\pi_4(b_4)|\pi_4(a_2), \pi_4(a_3), \pi_4(a_4), \pi_4(b_3)\} \\ &= \{-11/4|-3/4, -7/8, -5/2, -3/2\} = -21/8 \end{aligned}$$

Corollary 2 *The number of games $g \in S_n[d]$ such that $g =_{(1)} 0$ and $d > n - 2$ is at least $(2^{d-n+2} - 1)^{n-1}$.*

Theorem 5 *Let $(x_1, \ldots, x_n)$ be a n-tuple of surreal numbers $\in S_2[d]$, $d > n$ such that*

$$\begin{aligned} 0 < & x_1 < 1/2^{n-1} \\ -n < & x_2 < -n + 1 \\ & x_i < -n + 1, \forall i \in \{3, \ldots, n\} \end{aligned}$$

Then, there exists $g \in S_n[d]$ such that $\pi_i(g) = x_i, \forall i \in \{1, \ldots, n\}$.

Proof Let us consider the game $g = \{a_1, \ldots, a_n|b_1, b_2|b_3| \ldots |b_n\}$ where a_i and b_i are strings of n-player Hackenbush, as shown in Fig. 5, such that

$$\begin{aligned} &\pi_1(a_1) = x_1^L \geq 0 \\ &\pi_1(a_i) < 0, \forall i \in \{2, \ldots, n\} \\ &\pi_1(b_1) = x_1^R \leq 1/2^{n-1} \\ &\pi_1(b_i) > 1, \forall i \in \{2, \ldots, n\} \\ &\pi_2(a_1) > -n + 1 \\ &\pi_2(a_2) = x_2^R \leq -n + 1 \end{aligned}$$

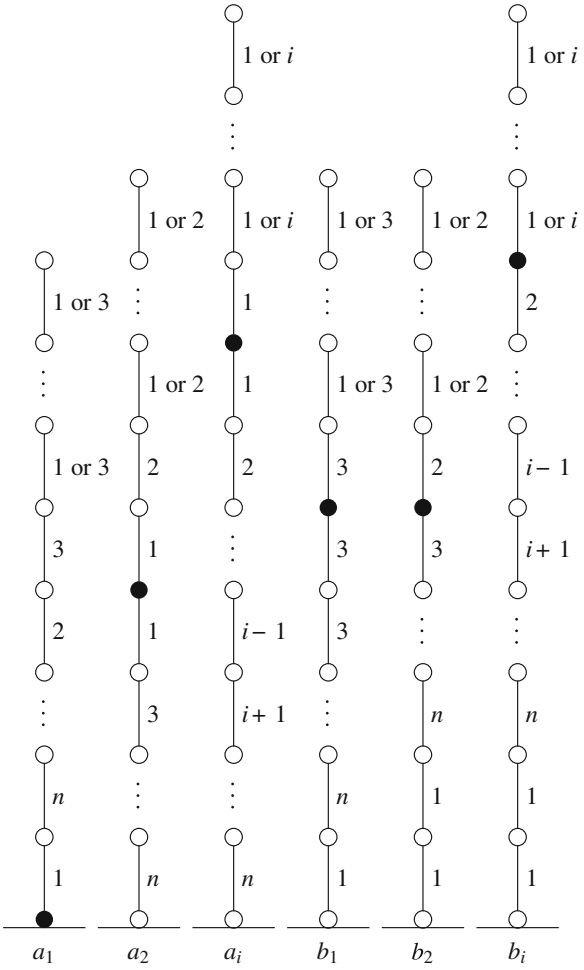


Fig. 5 The strings used to create the number corresponding to the tuple $(x_1, x_2, x_3, \ldots, x_n)$. The *black vertex* indicates the minimum length of the string from the ground

$$\pi_2(a_i) > -n+3, \forall i \in \{3, \ldots, n\}$$
$$\pi_2(b_1) \leq -n$$
$$\pi_2(b_2) = x_2^L \geq -n$$
$$\pi_2(b_i) > -n+1, \forall i \in \{3, \ldots, n\}$$
$$\pi_i(a_1) > -n+i-1, \forall i \in \{3, \ldots, n\}$$
$$\pi_i(a_2) > -n+i, \forall i \in \{3, \ldots, n\}$$
$$\pi_i(a_i) = x_i^R \leq -n+1, \forall i \in \{3, \ldots, n\}$$
$$\pi_i(a_j) > -n+i, \forall i, j \in \{3, \ldots, n\}, i \neq j$$

$$\pi_i(b_1) > -n + i - 1, \forall i \in \{3, \ldots, n\}$$
$$\pi_i(b_2) > -n + i - 2, \forall i \in \{3, \ldots, n\}$$
$$\pi_i(b_i) = x_i^L < -n + 1, \forall i \in \{3, \ldots, n\}$$
$$\pi_i(b_j) > -n + i - 2, \forall i, j \in \{3, \ldots, n\}, i \neq j$$

We observe that $a_i, b_i \in S_n[d-1], \forall i \in \{1, \ldots, n\}$; moreover, none of the following inequalities hold

$$a_i \geq_1 b_j, \quad \forall i, j \in \{1, \ldots, n\}$$
$$b_i \geq_2 a_j, \quad \forall i \in \{1, 2\}, \forall j \in \{1, \ldots, n\}$$
$$b_i \geq_2 b_j, \quad \forall i \in \{1, 2\}, \forall j \in \{3, \ldots, n\}$$
$$b_i \geq_i a_j, \quad \forall i \in \{3, \ldots, n\}, \forall j \in \{1, \ldots, n\}$$
$$b_i \geq_i b_j, \quad \forall i \in \{3, \ldots, n\}, \forall j \in \{1, 2\}$$

therefore $g \in S_n[d]$. In particular,

$$\begin{aligned}
\pi_1(g) &= \{\pi_1(a_1), \ldots, \pi_1(a_n) | \pi_1(b_1), \ldots, \pi_1(b_n)\} \\
&= \{\pi_1(a_1) | \pi_1(b_1)\} = \{x_1^L | x_1^R\} = x_1 \\
\pi_2(g) &= \{\pi_2(b_1), \pi_2(b_2) | \pi_2(a_1), \ldots, \pi_2(a_n), \pi_2(b_3), \ldots, \pi_2(b_n)\} \\
&= \{\pi_2(b_2) | \pi_2(a_2)\} = \{x_2^L | x_2^R\} = x_2 \\
\pi_i(g) &= \{\pi_i(b_i) | \pi_i(a_1), \ldots, \pi_i(a_n), \pi_i(b_1), \ldots, \pi_i(b_{i-1}), \pi_i(b_{i+1}), \ldots, \pi_i(b_n)\} \\
&= \{\pi_i(b_i) | \pi_i(a_i)\} = \{x_i^L | x_i^R\} = x_i
\end{aligned}$$

We observe that if x_i is an integer, then the string b_i does not exist, i.e., $G_i = \emptyset, i \neq 1, 2$. □

Example 3 Let us define the number $g = \{a_1, a_2, a_3, a_4 | b_1, b_2 | b_3 | b_4\}$ corresponding to the tuple (x_1, x_2, x_3, x_4) where

$$x_1 = \{1/16 | 1/8\} = 3/32$$
$$x_2 = \{-7/2 | -3\} = -13/4$$
$$x_3 = \{-15/4 | -7/2\} = -29/8$$
$$x_4 = \{-5 | -9/2\} = -19/4$$

The strings $a_1, a_2, a_3, a_4, b_1, b_2, b_3$, and b_4 are shown in Fig. 6. We observe that

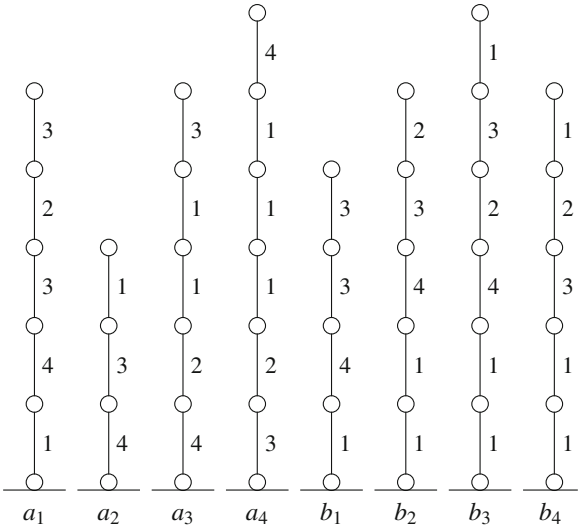


Fig. 6 The strings used to create the number corresponding to the tuple $(3/32, -13/4, -29/8, -19/4)$

$$\begin{aligned}
\pi_1(g) &= \{\pi_1(a_1), \pi_1(a_2), \pi_1(a_3), \pi_1(a_4)|\pi_1(b_1), \pi_1(b_2), \pi_1(b_3), \pi_1(b_4)\} \\
&= \{1/16, -3/2, -11/8, -19/16|1/8, 9/8, 19/16, 11/8\} = 3/32 \\
\pi_2(g) &= \{\pi_2(b_1), \pi_2(b_2)|\pi_2(a_1), \pi_2(a_2), \pi_2(a_3), \pi_2(a_4), \pi_2(b_3), \pi_2(b_4)\} \\
&= \{-4, -7/2| - 11/4, -3, -15/16, -31/32, -23/8, -11/4\} = -13/4 \\
\pi_3(g) &= \{\pi_3(b_3)|\pi_3(a_1), \pi_3(a_2), \pi_3(a_3), \pi_3(a_4), \pi_3(b_1), \pi_3(b_2), \pi_3(b_4)\} \\
&= \{-15/4| - 13/8, -3/4, -7/2, 1/32, -5/4, -11/4, -15/8\} = -29/8 \\
\pi_4(g) &= \{\pi_4(b_4)|\pi_4(a_1), \pi_4(a_2), \pi_4(a_3), \pi_4(a_4), \pi_4(b_1), \pi_4(b_2), \pi_4(b_3)\} \\
&= \{-5| - 15/16, 1/4, 1/16, -9/2, -7/8, -15/8, -31/16\} = -19/4
\end{aligned}$$

Corollary 3 *The number of games $g \in S_n[d]$ such that $g >_1 0$ and $d > n$ is at least*

$$(2^{d-n} - 1)[(2^{d-n+1} - 1)^{n-1} - (2^{d-n})^{n-1}]$$

Theorem 6 *Let $(x_1, \ldots, x_n)$ be a n-tuple of surreal numbers $\in S_2[d]$, $d > n$ such that*

$$\begin{aligned}
-n + 1 < x_1 &< -n + 3/2 \\
-n + 1 < x_2 &< -n + 3/2 \\
x_i &< -n + 1, \forall i \in \{3, \ldots, n\}
\end{aligned}$$

Then, there exists $g \in S_n[d]$ such that $\pi_i(g) = x_i, \forall i \in \{1, \ldots, n\}$.

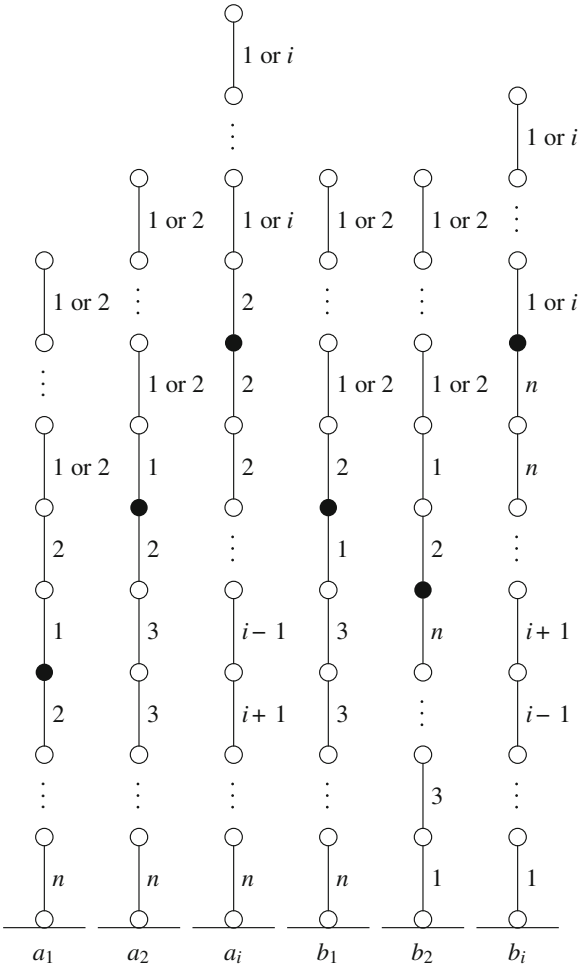


Fig. 7 The strings used to create the number corresponding to the tuple $(x_1, x_2, x_3, \ldots, x_n)$. The *black vertex* indicates the minimum length of the string from the ground

Proof Let us consider the game $g = \{a_1, \ldots, a_n | b_1, b_2 | b_3 | \ldots | b_n\}$ where a_i and b_i are strings of n-player Hackenbush, as shown in Fig. 7, such that

$$\begin{aligned}
&\pi_1(a_1) = x_1^L \geq -n + 1 \\
&\pi_1(a_2) < -n + 1 \\
&\pi_1(a_i) \leq -n + 1, \forall i \in \{3, \ldots, n\} \\
&\pi_1(b_1) = x_1^R \leq -n + 3/2 \\
&\pi_1(b_i) > 0, \forall i \in \{2, \ldots, n\} \\
&\pi_2(a_1) > -n + 2
\end{aligned}$$

$$\begin{aligned}
&\pi_2(a_2) = x_2^R \leq -n + 3/2\\
&\pi_2(a_i) > -n + 3, \forall i \in \{3, \ldots, n\}\\
&\pi_2(b_1) < -n + 1\\
&\pi_2(b_2) = x_2^L \geq -n + 1\\
&\pi_2(b_i) > -1, \forall i \in \{3, \ldots, n\}\\
&\pi_i(a_1) > -n + i, \forall i \in \{3, \ldots, n\}\\
&\pi_i(a_2) > -n + i, \forall i \in \{3, \ldots, n\}\\
&\pi_i(a_i) = x_i^R \leq -n + 1, \forall i \in \{3, \ldots, n\}\\
&\pi_i(a_j) > -n + i, \forall i, j \in \{3, \ldots, n\}, i \neq j\\
&\pi_i(b_1) > -n + i, \forall i \in \{3, \ldots, n\}\\
&\pi_i(b_2) > -i + 2, \forall i \in \{3, \ldots, n\}\\
&\pi_i(b_i) = x_i^L < -n + 1, \forall i \in \{3, \ldots, n\}\\
&\pi_i(b_j) > -i + 1, \forall i, j \in \{3, \ldots, n\}, i \neq j
\end{aligned}$$

We observe that $a_i, b_i \in S_n[d-1], \forall i \in \{1, \ldots, n\}$; moreover, none of the following inequalities hold

$$\begin{aligned}
&a_i \geq_1 b_j, \ \forall i, j \in \{1, \ldots, n\}\\
&b_i \geq_2 a_j, \ \forall i \in \{1, 2\}, \forall j \in \{1, \ldots, n\}\\
&b_i \geq_2 b_j, \ \forall i \in \{1, 2\}, \forall j \in \{3, \ldots, n\}\\
&b_i \geq_i a_j, \ \forall i \in \{3, \ldots, n\}, \forall j \in \{1, \ldots, n\}\\
&b_i \geq_i b_j, \ \forall i \in \{3, \ldots, n\}, \forall j \in \{1, 2\}
\end{aligned}$$

therefore $g \in S_n[d]$. In particular,

$$\begin{aligned}
\pi_1(g) &= \{\pi_1(a_1), \ldots, \pi_1(a_n) | \pi_1(b_1), \ldots, \pi_1(b_n)\}\\
&= \{\pi_1(a_1) | \pi_1(b_1)\} = \{x_1^L | x_1^R\} = x_1\\
\pi_2(g) &= \{\pi_2(b_1), \pi_2(b_2) | \pi_2(a_1), \ldots, \pi_2(a_n), \pi_2(b_3), \ldots, \pi_2(b_n)\}\\
&= \{\pi_2(b_2) | \pi_2(a_2)\} = \{x_2^L | x_2^R\} = x_2\\
\pi_i(g) &= \{\pi_i(b_i) | \pi_i(a_1), \ldots, \pi_i(a_n), \pi_i(b_1), \ldots, \pi_i(b_{i-1}), \pi_i(b_{i+1}), \ldots, \pi_i(b_n)\}\\
&= \{\pi_i(b_i) | \pi_i(a_i)\} = \{x_i^L | x_i^R\} = x_i
\end{aligned}$$

We observe that if x_i is an integer, then the string b_i does not exist, i.e., $G_i = \emptyset, i \neq 1, 2$. □

Example 4 Let us define the number $g = \{a_1, a_2, a_3, a_4 | b_1, b_2 | b_3 | b_4\}$ corresponding to the tuple (x_1, x_2, x_3, x_4) where

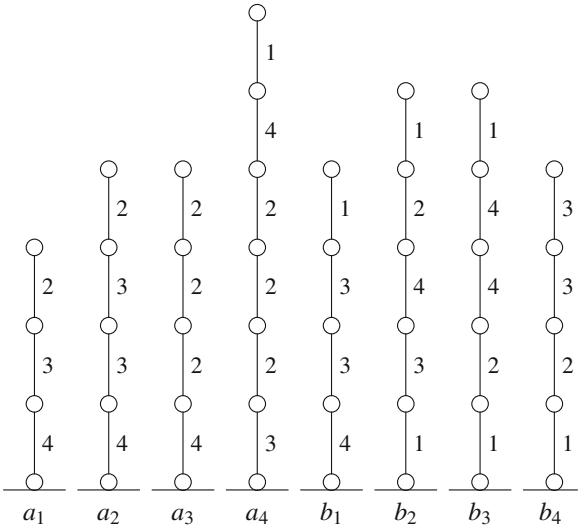


Fig. 8 The strings used to create the number corresponding to the tuple $(-11/4, -21/8, -9/2, -31/8)$

$$
\begin{aligned}
x_1 &= \{-3|-5/2\} = -11/4 \\
x_2 &= \{-11/4|-5/2\} = -21/8 \\
x_3 &= \{-5|-4\} = -9/2 \\
x_4 &= \{-4|-15/4\} = -31/8
\end{aligned}
$$

The strings a_1, a_2, a_3, a_4, b_1, b_2, b_3, and b_4 are shown in Fig. 8. We observe that

$$
\begin{aligned}
\pi_1(g) &= \{\pi_1(a_1), \pi_1(a_2), \pi_1(a_3), \pi_1(a_4)|\pi_1(b_1), \pi_1(b_2), \pi_1(b_3), \pi_1(b_4)\} \\
&= \{-3, -4, -4, -9/2|-5/2, 3/16, 3/16, 1/8\} = -11/4 \\
\pi_2(g) &= \{\pi_2(b_1), \pi_2(b_2)|\pi_2(a_1), \pi_2(a_2), \pi_2(a_3), \pi_2(a_4), \pi_2(b_3), \pi_2(b_4)\} \\
&= \{-4, -11/4|-3/2, -5/2, -1/8, -7/32, -15/16, -7/8\} = -21/8 \\
\pi_3(g) &= \{\pi_3(b_3)|\pi_3(a_1), \pi_3(a_2), \pi_3(a_3), \pi_3(a_4), \pi_3(b_1), \pi_3(b_2), \pi_3(b_4)\} \\
&= \{-5|-3/4, -3/8, -4, 1/32, -3/8, -15/16, -5/4\} = -9/2 \\
\pi_4(g) &= \{\pi_4(b_4)|\pi_4(a_1), \pi_4(a_2), \pi_4(a_3), \pi_4(a_4), \pi_4(b_1), \pi_4(b_2), \pi_4(b_3)\} \\
&= \{-4|1/4, 1/8, 1/8, -15/4, 1/8, -15/8, -11/8\} = -31/8
\end{aligned}
$$

Corollary 4 *The number of games $g \in S_n[d]$ such that $g < 0$ and $d > n$ is at least*

$$\binom{n}{2}(2^{d-n} - 1)^2(2^{d-n+1} - 1)^{n-2}$$

Table 1 Lower and upper bounds on $S_n[d]$

Class	Lower bound	Upper bound
$g = 0$	1	1
$g >_i 0$	$n(2^{d-n} - 1)[(2^{d-n+1} - 1)^{n-1} - (2^{d-n})^{n-1}], d > n$	$n[1^{n-1} + 2^{n-1} + \ldots + (2^d - 1)^{n-1}]$
$g =_{i,j} 0$	$\binom{n}{2}(d-n+2)^{n-2}, d > n-2$	$\binom{n}{2}d^{n-2}$
$g =_{(i)} 0$	$n(2^{d-n+2} - 1)^{n-1}, d > n-2$	$n(2^d - 1)^{n-1}$
$g < 0$	$\binom{n}{2}(2^{d-n} - 1)^2(2^{d-n+1} - 1)^{n-2}, d > n$	$(2^d - 1)^n$

Theorem 7 *Let $g = \{g_1, g_2, \ldots, g_n\}$ be a number. Then $\pi_i(g) + \pi_j(g) \leq 0, \forall i, j \in \{1, \ldots, n\}, i \neq j$.*

Proof We observe that $\pi_i(g_i) + \pi_j(g) < \pi_i(g_i) + \pi_j(g_i)$ and by the induction hypothesis $\pi_i(g_i) + \pi_j(g_i) \leq 0$. Analogously, $\pi_i(g) + \pi_j(g_j) < \pi_i(g_j) + \pi_j(g_j) \leq 0$ therefore no left option of $\pi_i(g) + \pi_j(g)$ is ≥ 0. □

If $g > i0$, then $\pi_i(g) > 0$ and $\pi_j(g) < 0, \forall j \in \{1, \ldots, n\}, j \neq i$ therefore we have an upper bound of $(2^d - 1)^n$. Using Theorem 7 we can refine this value obtaining $n[1^{n-1} + 2^{n-1} + \ldots + (2^d - 1)^{n-1}]$.

If $g = i, j0$, then $G_k = \emptyset, \forall k \in \{1, \ldots, n\}, k \neq i, j$ and $\pi_k(g)$ is a negative integer. Therefore, $\binom{n}{2}d^{n-2}$ is an upper bound.

4 Conclusion

Table 1 shows the lower and upper bounds on $S_n[d]$ equal to the number of n-player partizan cold games born by day d. We observe that if n is fixed, then $S_n[d] \in \Theta(2nd)$, but in order to establish the exact value of $S_n[d]$ further efforts are necessary.

References

1. Berlekamp ER, Conway JH, Guy RK (2001) Winning ways for your mathematical plays. AK Peters, Wellesley
2. Berlekamp ER (1974) The Hackenbush number system for compression of numerical data. Inform Control 26:134–140
3. Cincotti A (2006) Counting the number of three-player partizan cold games. In: van den Herik HJ, Ciancarini P, Donkers HHLM (eds) Proceedings of the conference on computers and games 2006. Springer, Turin, 29–31 May 2006, pp 181–189.
4. Cincotti A (2009) On the complexity of n-player Hackenbush. Integers 9:621–627
5. Cincotti A (2010) N-player partizan games. Theor Comput Sci 411:3224–3234
6. Cincotti A (2011) The game of n-player Cutcake. Theor Comput Sci 412:5678–5683
7. Cincotti A (2012) The lattice structure of n-player games. Theor Comput Sci 459:113–119

8. Cincotti A (2013) The structure of n-player games born by day d. In: Proceedings of the international multiconference of engineers and computer scientists 2013, IMECS 2013, Hong Kong, 13–15 March 2013. Lecture notes in engineering and computer science, pp 1113–1116.
9. Cincotti A (2013) The canonical form of multi-player combinatorial games. IAENG Int J App Math 43:77–80
10. Cincotti A (2013) Analyzing n-player Maundy Cake. Discrete Math. 313:1602–1609
11. Conway JH (2001) On numbers and games. AK Peters, Wellesley
12. Krawec WO (2012) Analyzing n-player impartial games. Int J Game Theory 41:345–367
13. Li SYR (1978) N-person Nim and N-person Moore's games. Int J Game Theory 7:31–36
14. Loeb DE (1994) Stable winning coalitions. In: Nowakowski RJ (ed) Games of no chance. Cambridge University Press, Cambridge, pp 451–471
15. Propp J (2000) Three-player impartial games. Theor Comput Sci 233:263–278
16. Straffin PD Jr (1985) Three-person winner-take-all games with Mc-Carthy's revenge rule. Coll Math J 16:386–394
17. Wolfe D, Fraser W (2004) Counting the number of games. Theor Comput Sci 313:527–532

Linear Programming Formulation of Boolean Satisfiability Problem

Algirdas Antano Maknickas

Abstract It was investigated the Boolean satisfiability (SAT) problem defined as follows: given a Boolean formula, check whether an assignment of Boolean values to the propositional variables in the formula exists, such that the formula evaluates to true. If such an assignment exists, the formula is said to be satisfiable; otherwise, it is unsatisfiable. With using of analytical expressions of multi-valued logic 2SAT boolean satisfiability was formulated as linear programming optimization problem. The same linear programming formulation was extended to find 3SAT and kSAT boolean satisfiability for k greater than 3. So, using new analytic multi-valued logic expressions and linear programming formulation of boolean satisfiability proposition that kSAT is in P and could be solved in linear time was proved.

Keywords Boolean satisfiability · Conjunctive normal form · Convex function · Linear programming · Multi-valued logic · NP problem · Time complexity.

1 Introduction

The Boolean satisfiability (SAT) problem [4] is defined as follows: given a Boolean formula, check whether an assignment of Boolean values to the propositional variables in the formula exists, such that the formula evaluates to true. If such an assignment exists, the formula is said to be satisfiable; otherwise, it is unsatisfiable. For a formula with n variables and m clauses the conjunctive normal form (CNF)

$$(X_1 \vee X_2) \wedge (X_3 \vee X_4) \wedge \cdots \wedge (X_{n-1} \vee X_n) \tag{1}$$

A. A. Maknickas (✉)
Vilnius Gediminas Technical University, Sauletekio al. 11, Vilnius, Lithuania
e-mail: algirdas.maknickas@vgtu.lt

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 275, DOI: 10.1007/978-94-007-7684-5_22,

is most the frequently used for representing Boolean formulas, where $\neg\forall X_i$ are independent and in most cases $m \geq n$. In CNF, the variables of the formula appear in literals (e.g., x) or their negation [e.g., ¬x (logical NOT ¬)]. Literals are grouped into clauses, which represent a disjunction (logical OR ∨) of the literals they contain. A single literal can appear in any number of clauses. The conjunction (logical AND ∧) of all clauses represents a formula.

Several algorithms are known for solving the 2-satisfiability problem; the most efficient of them take linear time [2, 7, 8]. Instances of the 2-satisfiability or 2SAT problem are typically expressed as 2-CNF or Krom formulas [8]

SAT was the first known NP-complete problem, as proved by Cook and Levin in 1971 [4, 9]. Until that time, the concept of an NP-complete problem did not even exist. The problem remains NP-complete even if all expressions are written in conjunctive normal form with 3 variables per clause (3-CNF), yielding the 3SAT problem. This means the expression has the form:

$$(X_1 \vee X_2 \vee X_3) \wedge (X_4 \vee X_5 \vee X_6) \wedge \cdots \wedge (X_{n-2} \vee X_{n-1} \vee X_n) \tag{2}$$

NP-complete and it is used as a starting point for proving that other problems are also NP-hard. This is done by polynomial-time reduction from 3-SAT to the other problem.

In 2010, Moustapha Diaby provided two further proofs for P = NP. His papers "Linear programming formulation of the vertex colouring problem" and "Linear programming formulation of the set partitioning problem give linear programming formulations for two well-known NP-hard problems" [5, 6]. After three years, it was shown, that linear programming formulation is possible and for boolean satisfiability [13].

2 Expressions of Multi-Valued Logic

Let describe integer discrete logic units as i, where $i \in \mathbb{Z}^+$. Let describe discrete function $g_k^n(a)$, $\forall a \in \mathbb{R}, \forall k \in \{0, 1, 2, \ldots, n-1\}$ as

$$g_k^n(a) = \lfloor a \rfloor + k \pmod{n}. \tag{3}$$

As mentioned in [10, 12] function $g_k^n(a)$ is one variable multi-valued logic generation function for multi-valued set $\{0, 1, 2, .., n-1\}$

The are n^n one variable a logic functions:

$$\begin{array}{cc|c|c} & & \lfloor a \rfloor & \\ \hline i_0 & & 0 & g^n_{i_0}(a) \\ i_1 & & 1 & g^n_{i_1}(a) \\ \rho\, i_2 & (a) = & 2 & g^n_{i_2}(a) \\ \dots & & \dots & \dots \\ i_{n-1} & & n-1 & g^n_{i_{n-1}}(a) \end{array} , \quad \begin{array}{c} i_0 \\ i_1 \\ \forall\, i_2 \\ \dots \\ i_{n-1} \end{array} \in \{0, 1, 2, \dots, n-1\} . \tag{4}$$

The are n^{n^2} two variables a, b logic functions:

$$\begin{array}{cccc} i_{0,0} & i_{0,1} & i_{0,2} & \dots i_{0,n-1} \\ i_{1,0} & i_{1,1} & i_{1,2} & \dots i_{1,n-1} \\ \mu\, i_{2,0} & i_{2,1} & i_{2,2} & \dots i_{2,n-1} \\ \dots & \dots & \dots & \dots \dots \\ i_{n-1,0} & i_{n-1,1} & i_{n-1,2} & \dots i_{n-1,n-1} \end{array} \quad (a, b) =$$

$$\begin{array}{c|cccc} \lfloor a \rfloor \backslash \lfloor b \rfloor & 0 & 1 & 2 & \dots\, n-1 \\ \hline 0 & g^n_{i_{0,0}}(a \times b) & g^n_{i_{0,1}}(a \times b) & g^n_{i_{0,2}}(a \times b) & \dots\, g^n_{i_{0,n-1}}(a \times b) \\ 1 & g^n_{i_{1,0}}(a \times b) & g^n_{i_{1,1}}(a \times b) & g^n_{i_{1,2}}(a \times b) & \dots\, g^n_{i_{1,n-1}}(a \times b) \\ 2 & g^n_{i_{2,0}}(a \times b) & g^n_{i_{2,1}}(a \times b) & g^n_{i_{2,2}}(a \times b) & \dots\, g^n_{i_{2,n-1}}(a \times b) \\ \dots & \dots & \dots & \dots & \dots \dots \\ n-1 & g^n_{i_{n-1,0}}(a \times b) & g^n_{i_{n-1,1}}(a \times b) & g^n_{i_{n-1,2}}(a \times b) & \dots\, g^n_{i_{n-1,n-1}}(a \times b) \end{array} ,$$

$$\begin{array}{cccc} i_{0,0} & i_{0,1} & i_{0,2} & \dots i_{0,n-1} \\ i_{1,0} & i_{1,1} & i_{1,2} & \dots i_{1,n-1} \\ \forall\, i_{2,0} & i_{2,1} & i_{2,2} & \dots i_{2,n-1} \\ \dots & \dots & \dots & \dots \dots \\ i_{n-1,0} & i_{n-1,1} & i_{n-1,2} & \dots i_{n-1,n-1} \end{array} \in \{0, 1, 2, \dots, n-1\} . \tag{5}$$

Using of this expressions for logic of higher dimensions with non deterministic Turing machine enables to make counter of linear time complexity [12].

The goal of this paper is the proof of proposition that kSAT is in P and could be solved in linear time. For this purpose will be used function with appropriate indexes of multi-valued logic of two variables described above.

3 2SAT is in P

Theorem 1 *If all variables are unique, equation*

$$\max \beta_2 (X_1, X_2, \dots, X_{n-1}, X_n) , \tag{6}$$

where

$$\beta_2\left(X_1, X_2, \ldots, X_{n-1}, X_n\right) = \mu\left(X_1 + X_2, \mu\left(X_3 + X_4, \ldots \mu\left(X_{n-3} + X_{n-2}, X_{n-1} + X_n\right)\right)\right) \tag{7}$$

$$\mu(a, b) = \begin{array}{c|c} \lfloor a \rfloor \backslash \lfloor b \rfloor & 0\ 1\ 2 \\ \hline 0 & 0\ 0\ 0 \\ 1 & 0\ 1\ 1 \\ 2 & 0\ 1\ 1 \end{array} \tag{8}$$

and $+$ *denotes algebraic summation could be solved for* $\forall X_i \in \{0, 1\}$ *in* $O(m)$.

Proof Let start from investigation of

$$f(x_1, x_2, \ldots, x_m) = \prod_{i=1}^{m} x_i, \ \forall x_i \in \mathbb{R} \tag{9}$$

in hyper-cube of sides [0, 1]. This function is convex, because

$$\prod_{i=1}^{m} x_i \leq \sum_{i=1}^{m} x_i \alpha^i \ , \tag{10}$$

$$1 = \sum_{i=1}^{m} \alpha^i \ . \tag{11}$$

It is obvious, on right side of Eq. (10) we have hyper-plane which goes through vertexes (0, 0, 0, .., 0) and (1, 1, 1, .., 1) of investigating hyper-cube. This hyper-plane has global maximum like a function $f(x_1, x_2, \ldots, x_m)$ and it equals 1. So we could resume that finding of global maximum of $f(x_1, x_2, \ldots, x_m)$ could be replaced by finding of global maximum of this hyper plane or

$$\max \prod_{i=1}^{m} x_i = \frac{1}{n} \max \sum_{i=1}^{n} x_i \ . \tag{12}$$

Let start to investigate Eq. (6) $\forall X_i \in \mathbb{R}, i \in \{1, 2, \ldots, n\}$. Function β_2 could be calculated within $O(m)$ [11]. According to Eqs. (8), (6) could be rewritten as follow

$$\begin{aligned} &\max \mu\left(X_1 + X_2, \mu\left(X_3 + X_4, \ldots, \mu\left(X_{n-3} + X_{n-2}, X_{n-1} + X_n\right)\right)\right) = \\ &\max \mu\left(X_1 + X_2, \max \mu\left(X_3 + X_4, \ldots, \max \mu\left(X_{n-3} + X_{n-2}, X_{n-1} + X_n\right)\right)\right) . \end{aligned} \tag{13}$$

So, m local partial maximums must satisfy equalities $\max \mu(X_{k-1} + X_k, 1) = 1$ to avoid 0 result of global maximum. This leads to inequalities $1 \leq (X_{k-1} + X_k) \leq 2$ but if we want to satisfy Eq. (12) we must leave $(X_{k-1} + X_k) = 1$ Now we could start

to solve satisfiability problem as global maximum of Eq. (6). If we have all variables unique, Eq. (13) could be solved repeating solving of system of LP equations

$$\begin{cases} \max \sum_{i=k-1}^{k} X_i \\ X_k + X_{k-1} = 1 \\ 0 \le X_{k-1} \le 1 \\ 0 \le X_k \le 1 \end{cases}. \tag{14}$$

Each of them has not zero max and could be solved using best known algorithm of linear programming [1] in $O\left(n^{3.5}\right)$ or in $O(2^{3.5})$. So all clauses should be optimized in $O\left(2^{3.5}\mathrm{m}\right)$.

3.1 Special Cases

Theorem 2 *If some of unique variables are negations* $\neg X_i$*, equation*

$$\max \beta_2 (X_1, X_2, \ldots, X_{n-1}, X_n) \tag{15}$$

could be solved for $\forall X_i \in \{0, 1\}$ *in* $O(m)$.

Proof If we have all variables unique, let replace all negations $\neg X_k$ with X'_k and all others just renamed by X'_i. Now Eq. (15) could be solved repeating solving of system of LP equations

$$\begin{cases} \max \sum_{i=k-1}^{k} X'_i \\ X'_k + X'_{k-1} = 1 \\ 0 \le X'_{k-1} \le 1 \\ 0 \le X'_k \le 1 \end{cases}. \tag{16}$$

Each of them has not zero max and could be solved in $O(2^{3.5})$. Going back to old variables do not change complexity of each solution. So all clauses should be optimized in $O\left(2^{3.5}\mathrm{m}\right)$.

Theorem 3 *If some of variables in different clauses are not unique, equation*

$$\max \beta_2 (X_1, X_2, \ldots, X_{n-1}, X_n) \tag{17}$$

could be solved for $\forall X_i \in \{0, 1\}$ *in* $O(m)$.

Proof If X_n, X_{n-1} are unique, let start to solve Eq. (17) starting from a system of LP equations

$$\begin{cases} \max \sum_{i=n-1}^{n} X_i \\ X_n + X_{n-1} = 1 \\ 0 \le X_{n-1} \le 1 \\ 0 \le X_n \le 1 \end{cases}. \tag{18}$$

Now we know two unique variables. So partial maximum is reached and system of equations Eq. (18) is solved in $O(2^{3.5})$. If $X_n = X_{n-1}$, $X_n = 1 \wedge X_{n-1} = 1$. All other sub-equations could be solved as follow: if solving sub-equations have one variable with earlier found value, LP equations

$$\begin{cases} \max \sum_{i=k-1}^{k} X_i \\ 1 \le X_k + X_{k-1} \le 2 \\ 0 \le X_{k-1} \le 1 \\ 0 \le X_k \le 1 \end{cases} \tag{19}$$

could be solved by reducing of Eq. (19) within inserting of earlier found variable value resigned to 1 (lower and upper bound could be increased in case the broken plane result to 1 earlier); if all variables of solving sub-equations in Eq. (19) aren't unique, these variables could be reassigned to 1 and solving repeated within next clause; if solving sub-equations have two unique variables, LP equations

$$\begin{cases} \max \sum_{i=k-1}^{k} X_i \\ X_k + X_{k-1} = 1 \\ 0 \le X_{k-1} \le 1 \\ 0 \le X_k \le 1 \end{cases} \tag{20}$$

could be solved. So all clauses should be optimized in $O\left(2^{3.5}\mathrm{m}\right)$.

Theorem 4 *If some of variables in different clauses are not unique and are negation each other, equation*

$$\max \beta_2 (X_1, X_2, \ldots, X_{n-1}, X_n) \tag{21}$$

could be solved for $\forall X_i \in \{0, 1\}$ *in* $O\,(v + e)$.

Proof Let mark all variables they are unique or not starting from the end of CNF. Now cycle through n variables must be repeated to rename them to X'_k so that new replaced variables marked as not unique will be negations which must be replaced with $1 - X'_k$, if they found second time. In kind first time found not unique variable could be assigned to 0. This lead to value of 1 for negation. It could be done in $O\,(2n)$. Now we start solving process for last clause. If $X'_n = \neg X'_{n-1}$, $X'_n = 1$. If X'_n, X'_{n-1} are unique, let start to solve Eq. (21) from a system of LP equations

$$\begin{cases} \max \sum\limits_{i=n-1}^{n} X'_i \\ X'_n + X'_{n-1} = 1 \\ 0 \le X'_{n-1} \le 1 \\ 0 \le X'_n \le 1 \end{cases} . \tag{22}$$

Now we know two or one unique variables.

Aspvall et al. [2] found the linear time procedure for solving 2-satisfiability instances, based on the notion of strongly connected components from graph theory. There are several efficient linear time algorithms for finding the strongly connected components of a graph, based on depth first search: Tarjan's strongly connected components algorithm [14] and the path-based strong component algorithm [3] each perform a single depth first search.

Let choose the clause walking path as depth first search algorithm of the strongly connected components of a graph, where graph is formed as interconnected clauses with the not unique variables. So, other sub-equations could be solved as follow: if solving sub-equations have one variable with earlier found value, LP equations

$$\begin{cases} \max \sum\limits_{i=k-1}^{k} X_i \\ 1 \le X_k + X_{k-1} \le 2 \\ 0 \le X_{k-1} \le 1 \\ 0 \le X_k \le 1 \end{cases} \tag{23}$$

could be solved by reducing of Eq. (23) within inserting of earlier found variable value reassigned to 1 (lower and upper bound could be increased in case the broken plane result to 1 earlier); if all variables of solving sub-equations aren't unique and are negations of earlier found values in a different clauses, they values could be assigned to 0 and values of residual variables leads to 1; if all variables of solving sub-equations aren't unique and are negations of earlier found values in the same clause, one of variables must be assigned to 1 and other to 0; if at least two clauses $\exists$, where $X_i \wedge \neg X_i$, CNF is not satisfiable ;if solving sub-equations have two unique variables, LP equations

$$\begin{cases} \max \sum\limits_{i=k-1}^{k} X_i \\ X_k + X_{k-1} = 1 \\ 0 \le X_{k-1} \le 1 \\ 0 \le X_k \le 1 \end{cases} \tag{24}$$

could be solved. Each sub-system of equations is solved in $O(2^{3.5})$ to reach global maximum 1. Finally, all clauses should be optimized at most $v + e$ (v—amount of vertexes, e— amount of edges) times or in $O\left(2^{3.5}(v + e)\right)$.

4 3SAT is in P

Theorem 5 *If all variables are unique, equation*

$$\max \beta_3 \left(X_1, X_2, \ldots, X_{n-1}, X_n\right) , \tag{25}$$

where

$$\begin{gathered}\beta_3 \left(X_1, X_2, \ldots, X_{n-1}, X_n\right) = \\ \mu \left(X_1 + X_2 + X_3, \mu \left(X_4 + X_5 + X_6, \ldots, \right.\right. \\ \left.\left. \mu \left(X_{n-5} + X_{n-4} + X_{n-3}, X_{n-2} + X_{n-1} + X_n\right)\right)\right)\end{gathered} \tag{26}$$

$$\mu \left(a, b\right) = \begin{array}{c|cccc} a \backslash b & 0 & 1 & 2 & 3 \\ \hline 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 2 & 0 & 1 & 1 & 1 \\ 3 & 0 & 1 & 1 & 1 \end{array} \tag{27}$$

and $+$ *denotes algebraic summation could be solved for* $\forall X_i \in \{0, 1\}$ *in* $O\left(m\right)$.

Proof Let start to investigate Eq. (25) when $\forall X_i \in \mathbb{R}, i \in \{1, 2, \ldots, n\}$. Function β_3 could be calculated within $O\left(m\right)$ [11]. According to Eqs. (26), (25) could be rewritten as follow

$$\begin{gathered}\max \mu \left(X_1 + X_2 + X_3, \mu \left(X_4 + X_5 + X_6, \ldots, \right.\right. \\ \left.\left. \mu \left(X_{n-5} + X_{n-4} + X_{n-3}, X_{n-2} + X_{n-1} + X_n\right)\right)\right) .\end{gathered} \tag{28}$$

So, m local partial maximums must satisfy equalities

$$\max \mu(X_{k-2} + X_{k-1} + X_k, 1) = 1 \tag{29}$$

to avoid 0 result of global maximum. This leads to inequalities

$$1 \leq \left(X_{k-2} + X_{k-1} + X_k\right) \leq 3 , \tag{30}$$

but if we want to satisfy Eq. (12) we must leave $\left(X_{k-2} + X_{k-1} + X_k\right) = 1$.

If we have all variables unique, Eq. (25) could be solved repeating solving of system of LP equations

$$\begin{cases} \max \sum\limits_{i=k-2}^{k} X_i \\ X_k + X_{k-1} + X_{k-2} = 1 \\ 0 \le X_{k-2} \le 1 \\ 0 \le X_{k-1} \le 1 \\ 0 \le X_k \le 1 \end{cases}. \tag{31}$$

Each of them has not zero max and could be solved in $O(3^{3.5})$. So all clauses should be optimized in $O\left(3^{3.5}\mathrm{m}\right)$.

4.1 Special Cases

Theorem 6 *If some of unique variables are negations* $\neg X_i$*, equation*

$$\max \beta_3 (X_1, X_2, \ldots, X_{n-1}, X_n) \tag{32}$$

could be solved for $\forall X_i \in \{0, 1\}$ *in* $O(m)$.

Proof If we have all variables unique, let replace all negations $\neg X_k$ with X'_k and all others just renamed by X'_i. Now Eq. (32) could be solved repeating solving of system of LP equations

$$\begin{cases} \max \sum\limits_{i=k-2}^{k} X'_i \\ X'_k + X'_{k-1} + X'_{k-1} = 1 \\ 0 \le X'_{k-2} \le 1 \\ 0 \le X'_{k-1} \le 1 \\ 0 \le X'_k \le 1 \end{cases}. \tag{33}$$

Each of them has not zero max and could be solved in $O(3^{3.5})$. Going back to old variables do not change complexity of each solution. So all clauses should be optimized in $O\left(3^{3.5}\mathrm{m}\right)$.

Theorem 7 *If some of variables in different clauses are not unique, equation*

$$\max \beta_3 (X_1, X_2, \ldots, X_{n-1}, X_n) \tag{34}$$

could be solved for $\forall X_i \in \{0, 1\}$ *in* $O(m)$.

Proof If X_n, X_{n-1}, X_{n-2} are unique, let start to solve Eq. (34) starting from a system of LP equations

$$\begin{cases} \max \sum\limits_{i=n-2}^{n} X_i \\ X_n + X_{n-1} + X_{n-2} = 1 \\ 0 \le X_{n-2} \le 1 \\ 0 \le X_{n-1} \le 1 \\ 0 \le X_n \le 1 \end{cases}. \tag{35}$$

Now we know three unique variables. The partial maximum is reached and system of equations Eq. (35) is solved in $O(3^{3.5})$. If $X_n = X_{n-1}, X_n = 1 \wedge X_{n-1} = 1 \wedge X_{n-2} = 1$. All other sub-equations could be solved as follow: if solving sub-equations have one or two variables with earlier found value, LP equations

$$\begin{cases} \max \sum\limits_{i=k-2}^{k} X_i \\ 1 \le X_k + X_{k-1} + X_{k-2} \le 3 \\ 0 \le X_{k-2} \le 1 \\ 0 \le X_{k-1} \le 1 \\ 0 \le X_k \le 1 \end{cases} \tag{36}$$

could be solved by reducing of Eq. (36) with inserting of earlier found variables value reassigned to 1; solving each sub-equations three unique variables of LP equations

$$\begin{cases} \max \sum\limits_{i=k-2}^{k} X_i \\ X_k + X_{k-1} + X_{k-2} = 1 \\ 0 \le X_{k-2} \le 1 \\ 0 \le X_{k-1} \le 1 \\ 0 \le X_k \le 1 \end{cases} \tag{37}$$

could be found. So all clauses should be optimized in $O\left(3^{3.5}\mathrm{m}\right)$.

Theorem 8 *If some of variables in different clauses are not unique and are negation each other, equation*

$$\max \beta_3 (X_1, X_2, \ldots, X_{n-1}, X_n) \tag{38}$$

could be solved for $\forall X_i \in \{0, 1\}$ *in* $O(v + e)$.

Proof Let mark all variables they are unique or not starting from the end of CNF. Now cycle through n variables must be repeated to rename them to X'_k so that new replaced variables marked as not unique will be negations which must be replaced with $1 - X'_k$, if they found second time. In kind first time found not unique variable could be assigned to 0. This lead to value of 1 for negation. It could be done in $O(2n)$. Now we start solving process for last clause. If not all variables are unique in the first clause, after sorting of variables so that third one will be negation and second one will be unique, LP equations

$$\begin{cases} \max X'_n + X'_{n-1} + 1 - X'_n \\ X'_n + X'_{n-1} + 1 - X'_{n-2} = 1 \\ 0 \le X'_{n-2} \le 1 \\ 0 \le X'_{n-1} \le 1 \\ 0 \le X'_n \le 1 \end{cases} \tag{39}$$

could be solved. If X'_n, X'_{n-1}, X'_{n-2} are unique, let start to solve Eq. (26) from a system of LP equations

$$\begin{cases} \max \sum_{i=n-2}^{n} X'_i \\ X'_n + X'_{n-1} + X'_{n-2} = 1 \\ 0 \le X'_{n-2} \le 1 \\ 0 \le X'_{n-1} \le 1 \\ 0 \le X'_n \le 1 \end{cases}. \tag{40}$$

Now we know three or two unique variables.

Let choose the clause walking path as depth first search algorithm of the strongly connected components of a graph, where graph is formed as interconnected clauses with the not unique variables. So, other sub-equations could be solved as follow: if solving sub-equations have one or two variable with earlier found value, LP equations

$$\begin{cases} \max \sum_{i=k-2}^{k} X'_i \\ 1 \le X'_k + X'_{k-1} + X'_{k-2} \le 3 \\ 0 \le X'_{k-2} \le 1 \\ 0 \le X'_{k-1} \le 1 \\ 0 \le X'_k \le 1 \end{cases}, \tag{41}$$

where one or two of three variables X'_{k-2}, X'_{k-1}, X'_k are equal $1 - X'_q$ could be solved by reducing of Eq. (41) within inserting of earlier found variable value reassigned to 1; if all variables of solving sub-equations aren't unique and are negations of earlier found values in a different clauses, they values could be assigned to 0 and values of residual variables leads to 1; if all variables of solving sub-equations aren't unique and are negations of earlier found values in the same clause, one of variables must be assigned to 1 and other to 0; if solving sub-equations have three unique variables, LP equations

$$\begin{cases} \max \sum_{i=k-2}^{k} X_i \\ X_k + X_{k-1} + X_{k-2} = 1 \\ 0 \le X_{k-2} \le 1 \\ 0 \le X_{k-1} \le 1 \\ 0 \le X_k \le 1 \end{cases} \tag{42}$$

could be solved. Each sub-system of equations is solved in $O(3^{3.5})$ to reach global maximum 1. Finally, all clauses should be optimized at most $v + e$ (v—amount of vertexes, e— amount of edges) times or in $O\left(3^{3.5}\,(v+e)\right)$.

5 kSAT is in P

Theorem 9 *If all variables are unique, equation*

$$\max \beta_k\,(X_1, X_2, \ldots X_{n-1}, X_n)\,, \tag{43}$$

where

$$\beta_k\,(X_1, X_2, \ldots, X_{n-1}, X_n) = \mu\left(\sum_{i=1}^{k} X_i, \mu\left(\sum_{i=k+1}^{2k} X_i, \ldots, \quad \mu\left(\sum_{i=n-2k+1}^{n-k} X_i, \sum_{i=n-k+1}^{n} X_i\right)\right)\right) \tag{44}$$

$$\mu\,(a,b) = \begin{array}{c|l} a\backslash b & 0\ 1\ 2\ \ldots\ n-1 \\ \hline 0 & 0\ 0\ 0\ \ldots\ \ 0 \\ 1 & 0\ 1\ 1\ \ldots\ \ 1 \\ 2 & 0\ 1\ 1\ \ldots\ \ 1 \\ \ldots & \\ n-1 & 0\ 1\ 1\ \ldots\ \ 1 \end{array} \tag{45}$$

and $\sum$ *denotes algebraic summation could be solved for* $\forall X_i \in \{0,1\}$ *in* $O\left(k^{3.5}m\right)$.

Proof Let start to investigate Eq. (43) when $\forall X_i \in \mathbb{R}$, $i \in \{1, 2, \ldots, n\}$. Function β_k could be calculated within $O\,(m)$ [11]. According to Eqs. (8), (43) could be rewritten as follow

$$\max \mu\left(\sum_{i=1}^{k} X_i, \mu\left(\sum_{i=k+1}^{2k} X_i, \ldots, \quad \mu\left(\sum_{i=n-2k+1}^{n-k} X_i, \sum_{i=n-k+1}^{n} X_i\right)\right)\right). \tag{46}$$

If we have all variables unique, Eq. (44) could be solved repeating solving of system of LP equations

$$\begin{cases} \max \sum\limits_{i=k+1}^{2k} X_i \\ \sum\limits_{i=k+1}^{2k} X_i = 1 \\ 0 \le X_i \le 1 \ \wedge \forall i \in \{1, 2, \ldots, n\} \end{cases}. \tag{47}$$

Each of them has not zero max and could be solved in $O(k^{3.5})$. So all clauses should be optimized in $O\left(k^{3.5}\mathrm{m}\right)$.

5.1 Special Cases

Theorem 10 *If some of unique variables are negations* $\neg X_i$*, equation*

$$\max \beta_k \left(X_1, X_2, \ldots, X_{n-1}, X_n\right) \tag{48}$$

could be solved for $\forall X_i \in \{0, 1\}$ *in* $O\left(k^{3.5}m\right)$.

Proof If we have all variables unique, let replace all negations $\neg X_k$ with Y_k and all others just renamed by Y_i. Now Eq. (48) could be solved repeating solving of system of LP equations

$$\begin{cases} \max \sum\limits_{i=k+1}^{2k} Y_i \\ \sum\limits_{i=k+1}^{2k} Y_i = 1 \\ 0 \leq Y_i \leq 1 \ \wedge \forall i \in \{1, 2, \ldots, n\} \end{cases}. \tag{49}$$

Each of them has not zero max and could be solved in $O(k^{3.5})$. Going back to old variables do not change complexity of each solution. So all clauses should be optimized in $O\left(k^{3.5}\mathrm{m}\right)$.

Theorem 11 *If some of variables in different clauses are not unique, equation*

$$\max \beta_k \left(X_1, X_2, \ldots, X_{n-1}, X_n\right) \tag{50}$$

could be solved for $\forall X_i \in \{0, 1\}$ *in* $O\left(k^{3.5}m\right)$.

Proof If $X_n, X_{n-1}, \ldots, X_{n-k+1}$ are unique, let start to solve Eq. (50) starting from a system of LP equations

$$\begin{cases} \max \sum\limits_{i=n-k+1}^{n} X_i \\ \sum\limits_{i=n-k+1}^{n} X_i = 1 \\ 0 \leq X_i \leq 1 \wedge \forall i \in \{1, 2, \ldots, n\} \end{cases}. \tag{51}$$

Now we know k unique variables. The partial maximum is reached and system of equations is solved in $O(k^{3.5})$. If $X_n = X_{n-1}$, $X_n = 1 \wedge X_{n-2} = 1, \ldots$. All other sub-equations could be solved as follow: if solving sub-equations have at least one variable with earlier found value, LP equations

$$\begin{cases} \max \sum\limits_{i=n-k+1}^{n} X_i \\ 1 \leq \sum\limits_{i=n-k+1}^{k} X_i \leq k \\ 0 \leq X_i \leq 1 \wedge \forall i \in \{1, 2, \ldots, n\} \end{cases} \tag{52}$$

could be solved by reducing of Eq. (52) with inserting of earlier found variables value reassigned to 1; if solving sub-equations have k unique variables, LP equations

$$\begin{cases} \max \sum\limits_{i=n-k+1}^{k} X_i \\ \sum\limits_{i=n-k+1}^{k} X_i = 1 \\ 0 \leq X_i \leq 1 \wedge \forall i \in \{1, 2, \ldots, n\} \end{cases} \tag{53}$$

could be solved. So all clauses should be optimized in $O\left(k^{3.5}m\right)$.

Theorem 12 *If some of variables in different clauses are not unique and are negation each other, equation*

$$\max \beta_k \left(X_1, X_2, \ldots, X_{n-1}, X_n\right) \tag{54}$$

could be solved for $\forall X_i \in \{0, 1\}$ *in* $O\left(k^{3.5}\left(v + e\right)\right)$.

Proof Let mark all variables they are unique or not starting from the end of CNF. Now cycle through n variables must be repeated to rename them to X'_k so that new replaced variables marked as not unique will be negations which must be replaced with $1 - X'_k$, if they found second time. In kind first time found not unique variable could be assigned to 0. This lead to value of 1 for negation. It could be done in $O\left(2n\right)$. Now we start solving process for last clause. If not all variables are unique in the first clause, after sorting of variables so that negations occurs at the end of the list, LP equations

$$\begin{cases} \max \sum\limits_{i=n-k+1}^{l} X'_i + \sum\limits_{i=l+1}^{n} \left(1 - X'_i\right) \\ \sum\limits_{i=n-k+1}^{l} X'_i + \sum\limits_{i=l+1}^{n} \left(1 - X'_i\right) = 1 \\ 0 \leq X'_i \leq 1 \wedge \forall i \in \{1, 2, \ldots, n\} \end{cases} \tag{55}$$

could be solved. If $X'_n, X'_{n-1}, \ldots, X'_{n-k+1}$ are unique, let start to solve Eq. (54) from a system of LP equations

$$\begin{cases} \max \sum\limits_{i=n-k+1}^{n} X'_i \\ \sum\limits_{i=n-k+1}^{n} X'_i = 1 \\ 0 \le X'_i \le 1 \wedge \forall i \in \{1, 2, \ldots, n\} \end{cases}. \tag{56}$$

Now we know at most k unique variables.

Let choose the clause walking path as depth first search algorithm of the strongly connected components of a graph, where graph is formed as interconnected clauses with the not unique variables. So, other sub-equations could be solved as follow: if solving sub-equations have at least one variable with earlier found value, LP equations

$$\begin{cases} \max \sum\limits_{i=k+1}^{l} X'_i + \sum\limits_{i=l+1}^{2k} \left(1 - X'_i\right) \\ 1 \le \sum\limits_{i=k+1}^{l} X'_i + \sum\limits_{i=l+1}^{2k} \left(1 - X'_i\right) \le k \\ 0 \le X'_i \le 1 \wedge \forall i \in \{1, 2, \ldots, n\} \end{cases} \tag{57}$$

could be solved by reducing of Eq. (57) within inserting of earlier found variables values reassigned to 1; if all variables of solving sub-equations aren't unique and are negations of earlier found values in a different clauses, they values could be assigned to 0 and values of residual variables leads to 1; if all variables of solving sub-equations aren't unique and are negations of earlier found values in the same clause, one of variables must be assigned to 1 and other to 0; if solving sub-equations have k unique variables, LP equations

$$\begin{cases} \max \sum\limits_{i=k+1}^{2k} X'_i \\ \sum\limits_{i=k+1}^{2k} X'_i = 1 \\ 0 \le X'_i \le 1 \wedge \forall i \in \{1, 2, \ldots, n\} \end{cases} \tag{58}$$

could be solved. Each sub-system of equations is solved in $O(k^{3.5})$ to reach global maximum 1. Finally, all clauses should be optimized at most $v + e$ (v—amount of vertexes, e— amount of edges) times or in $O\left(k^{3.5}\,(v + e)\right)$.

6 Conclusions

Every 2SAT problem is solvable using linear programming formulation in linear time. Every 3SAT problem is solvable using linear programming formulation in linear time. Every kSAT problem is solvable using linear programming formulation

in linear time for $k > 3$. Every NP mathematical problem is reducible in polynomial time to kSAT problem and solvable in linear time if exist full, appropriate and correct knowledge basis for it and the time to get each item of knowledge basis is match less than calculation time on this items.

List of Symbols and Abbreviations

$\neg$	Logical NOT
$\vee$	Logical OR
$\wedge$	Logical AND
$\in$	in
$\mathbb{Z}^+$	Set of positive integer numbers
$g_k^n(a)$	Integer function of argument a
$\forall$	for all
$\mathbb{R}$	Set of real numbers
$\lfloor a \rfloor$	Floor round of a
$(\text{mod } n)$	Modulus of n function
ρ	Multi-valued logic function of one argument
$i_0, i_1, i_2, \ldots$	Corresponding integer indexes
μ	Multi-valued logic function of two arguments
$i_{0,0}, i_{0,1}, i_{0,2}, \ldots$	Corresponding integer indexes
$\prod$	Product
$\times$	Multiplication
$\sum$	Summation
$+$	Summation of two numbers
$\leq$	Less equal
$<$	Less
$\geq$	Greater equal
$>$	Greater
$=$	Equal
X_i	Integer variables
i	Integer index
v	Amount of vertexes
e	Amount of edges
$O(m)$	Big O notation
max	Maximum function
k	Number of variables in clause
n	Total number of variables
m	Total number of clauses
$\exists$	Exist
LP	Linear programming
CNF	Conjunctive normal form
SAT	Boolean satisfiability

References

1. Adler I, Karmarkar N, Resende MGC, Veiga G (1989) An implementation of Karmarkar's algorithm for linear programming. Math Program 44:297–335
2. Aspvall B, Plass MF, Tarjan RE (1979) A linear-time algorithm for testing the truth of certain quantified boolean formulas. Inf Proces Lett 8(3):121–123
3. Cheriyan J, Mehlhorn K (1996) Algorithms for dense graphs and networks on the random access computer. Algorithmica 15(6):521–549
4. Cook S (1971) The complexity of theorem proving procedures. In: Proceedings of the third annual ACM symposium on theory of, computing, pp 151–158
5. Diaby M (2010) Linear programming formulation of the set partitioning problem. Int J Oper Res 8(4):399–427
6. Diaby M (2010) Linear programming formulation of the vertex colouring problem. Int J Math Oper Res 2(3):259–289
7. Even S, Itai A, Shamir A (1976) On the complexity of time table and multi-commodity flow problems. SIAM J Comput 5(4):691–703
8. Krom MR (1967) The decision problem for a class of first-order formulas in which all disjunctions are binary. Zeitschrift für Mathematische Logik und Grundlagen der Mathematik 13:15–20
9. Levin L (1973) Universal search problems. Probl Inf Trans 9(3), 265–266. (Russian)
10. Maknickas AA (2010) Finding of k in Fagin's R. Theorem 24. arXiv:1012.5804v1 (2010).
11. Maknickas AA (2012) How to solve kSAT in polinomial time. arXiv:1203.6020v1
12. Maknickas AA (2012) How to solve kSAT in polinomial time. arXiv:1203.6020v2
13. Maknickas AA (2013) Programming formulation of kSAT. In: Lecture notes in engineering and computer science: Proceedings of the international multiConference of engineers and computer scientists 2013, pp. 1066–1070
14. Tarjan RE (1972) Depth-first search and linear graph algorithms. SIAM J Comput 1(2):146–160

Optimization of Path for Water Transmission and Distribution Systems

Ioan Sarbu and Emilian Stefan Valea

Abstract The efficient design of water adduction mains and distribution systems involves several optimization processes among which an important place is held by their path optimization. A principal application of the branched network optimal path is to evaluate the hydraulic performance of the distribution system based on selected schemes for many types of network analysis (e.g., design, operation, calibration). Already known algorithms for solving this problem usually determine a sole solution which in some cases proves to be suboptimal. This chapter is a mathematical approach of the adduction mains and branched pipe networks path optimization for water transmission and distribution. That is why there are developed two deterministic mathematical models for optimization of water adduction main path, based on techniques of sequential operational calculus, implemented in a computer program. Using these optimization models could be obtained an optimal solution for selection of source location and of water adduction main path based on graph-theory and dynamic programming. Also, it is developed an algorithm based on graph-theory which generates all minimal trees of the graph comprising nodes where consumers are placed and links (pipes) between them and is implemented in a computer program. Numerical example will be presented to demonstrate the accuracy and efficiency of the proposed optimization models. These show a good performance of the new models.

Keywords Adduction main · Branched network · Computer programs · Dynamic programming · Graph-theory · Optimal path · Optimization models · Water supply

I. Sarbu (✉) · E. S. Valea
Department of Building Services Engineering, "Politehnica" University of Timisoara,
Piata Bisericii, no. 4A, 300233 Timisoara, Romania
e-mail: ioan.sarbu@ct.upt.ro

E. S. Valea
e-mail: emilian.valea@ct.upt.ro

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 275, DOI: 10.1007/978-94-007-7684-5_23,

1 Introduction

Adduction mains and distribution network are an essential part of all water supply systems and its cost may be equal to or greater than 60 % of the entire cost of the project [1, 2].

Attempts should be made to reduce the cost and energy consumption of the transmission and distribution system through optimization in analysis and design. In some cases a distribution system is initially built to supply with low water flow rates a little territory and is extended over time as water consumption increases. In this case is realized originally a branched network which is then transformed into a looped network with increased capacity and security. As a consequence, the efficient design of a water adduction main and a branched network involves several optimization processes among which an important place is held by their path optimization. That problem is approached also on the looped network design for to determine an independent loop system (the virtual branched network) [2].

Traditionally, the choice of the optimal solution is made through analytical study of two or three versions selected from the possible set by predicted decisions [3]. The errors of these decisions are inverse proportional to the designer experience.

The modern mathematical disciplines as operational research give to the designer a vast apparatus of scientific analysis in optimal decisions establishing [4–9].

The mathematical theory and planning of multistage decision processes, the term was introduced by Richard Bellman in 1957 [10–12]. It may be regarded as a branch of mathematical programming or optimization problems formulated as a sequence of decision.

Traditional optimization algorithms have been applied to the minimum cost optimal design problem, such as linear programming [13], first introduced by Labye [14] for open networks. Dynamic programming is useful however for optimizing temporal processes [15] such as those typical in system operation problems. Sterling and Coulbeck [16], Coulbeck [17], Sabel and Helwig [18], and Lansey and Awumah [19] applied dynamic programming to determine optimal pumping operation for minimization of costs in a water system. Dynamic programming is also used primarily to solve tree–shaped networks [20, 21] and could be extended to solve to looped systems [22]. Also, Sarbu et al. [23, 24] has been applied graph theory to establish optimal path for water adduction mains and for branched water supply networks.

In pressurized pipe networks are first selected the nodes with water consumers and pressure device (pumping station, reservoir), then the network path that connects these nodes is optimized in several steps. This optimization problem can have one or more solutions. Already knows algorithms (Sollin, Kruskal, etc.) [3, 6, 25] to solve the problem determine usually a single solution that in certain situations [26] proves to be quasi optimal.

In this context, this chapter develops two deterministic mathematical models for optimization of water adduction main path, based on techniques of sequential operational calculus, implemented in a computer program. Using these optimization models could be obtained an optimal solution for selection of source location and of

water adduction main path based on graph-theory and dynamic programming. Also, it is developed an algorithm based on graph-theory which generates all minimal trees of the graph comprising nodes where consumers are placed and links (pipes) between them and is implemented in a computer program. Thus, can be determined all optimal solutions, for a given criterion. The results of a few numerical applications show the effectiveness and efficiency of the proposed optimization models.

2 Path Optimization Models for Water Adduction Mains

2.1 Optimization Model Based on Dynamic Programming

Sequential decision problem stated above is typical of dynamic programming, which is a procedure for optimizing a multistage decision at each stage. The technique is based on the simple Bellman's principle of optimality [10, 11], which states that "an optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision.

To solve a dynamic programming problem, it is necessary to evaluate both immediate and long-term consequence costs for each possible state at each stage. This evaluation is done through the development of the following recursive equation (for minimization):

$$\min Z = \sum_{i=1}^{N} V_i(X_{i-1}, X_i) \tag{1}$$

considering that each size X_i can vary in a field that depends on X_0 and X_{i+1}, $\forall \; i = 1, 2, \ldots, N$.

Optimization formula Eq. (1) is generalized for case if the terms X_i are vectors with n components:

$$X_i = \{x_{1i}\ x_{2i} \ldots x_{ni}\} \tag{2}$$

The form Eq. (1) imposed for function Z and the nature of variable variation domains allow use of a system of N phases for which $V_i\,(X_{i-1}, X_i); (i = 1, 2, \ldots, N)$ is the value function attached to each phase, and Z- value function attached for phase crowd.

Finding the minimum of the function Z requires solving at each step the functional equation:

$$f_{0,i}(X_0, X_i) = \min_{\{X_{i-1}\}} [V_i(X_{i-1}, X_i) + f_{0,i-1}(X_0, X_{i-1})] \tag{3}$$

where the notation $\{X_{i-1}\}$ means that X_{i-1} belongs to a values set which depend only on X_0 and X_i.

Calculation procedure consists in successively solving of minimization problems Eq. (3) for $i = 1, 2, \ldots, N$, with discrete variation of X_0 and storing intermediate results. Thus, it are calculated successive optimal sub-policies for phases 1 and 2 together, then phase 1, 2 and 3 together, …, for phases $\overline{1,\ N}$together, i.e. optimal policies:

$$\min Z = f_{0,N}(X_0, X_N) = \min_{\{X_{n-1}\}} [V_N(X_{N-1}, X_N) + f_{0,N-1}(X_0, X_{N-1}) \qquad (4)$$

in which $f_{0,N}(X_0,X_N)$ is optimal policy value from X_0 to X_N;$V_N(X_{N-1},X_N)$— sub-policy value from X_{N-1} to X_N; $f_{0,N-1}(X_0, X_{N-1})$— optimal policy value from X_0 to X_{N-1}.

A numerical example which illustrates the application of this optimization model shall be presented below.

2.2 Optimization Model Based on Graph-Theory

Mathematical model of dynamic processes of discrete and, determinist type can be simplified using graph theory.

The modelling of stated problem is realized by plotting oriented connected graph $G = (X, U)$ which consists of source as origin, paths as edges and critical points as vertices. Each edge $u^i_j \in U$ is assigned with a number $\lambda(u^i_j) \geq 0$, in conventional units, depending on the adopted optimization criterion. Optimal path is given by minimum value path in the graph, which is determined by applying the Bellman-Kalaba algorithm [23, 27].

The graph $G = (X, U)$ has attached a matrix $\mathbf{M}$ whose elements m_{ij} are:

$$m_{ij} = \begin{cases} \lambda(u^i_j) \ - \ \text{edge value from } x_i \ to \ x_j \\ \infty \ - \ \text{if vertices } x_i \text{ and } x_j \text{ are not adjacent} \\ 0 \ - \ \text{for } i = j \end{cases} \qquad (5)$$

Optimal path is given by path μ of graph, which has the total value:

$$\lambda(\mu) = \sum_{u^i_j \in \mu} \lambda(u^i_j) \ \rightarrow \ \min \qquad (6)$$

If V_i is the minimum value of existing path μ^i_n, $(i = \overline{0, n})$ from at vertex x_i to vertex x_n:

$$V_i = \lambda(\mu^i_n), \quad (i = \overline{0, n}) \qquad (7)$$

so

$$V_n = 0 \qquad (8)$$

then in accordance with the principle of optimality:

$$V_i = \min_{j \neq i}(V_j + m_{ij}), \quad (i = \overline{0, n-1}; \quad j = \overline{0, n}) \quad \text{and} \quad V_n = 0 \tag{9}$$

The system Eq. (9) is solved iteratively, noting with V_i^k the value of V_i obtained from iteration k, namely:

$$V_i^o = \min \quad (i = \overline{0, n-1}); \quad V_n^o = 0 \tag{10}$$

It is calculated:

$$V_i^1 = \min_{j \neq i}(V_j^{\mathrm{o}} + m_{ij}), \quad (i = \overline{0, n-1}; \quad j = \overline{0, n}); \quad V_n^1 = 0 \tag{11}$$

and then:

$$V_i^k = \min_{j \neq i}(V_j^{k-1} + m_{ij}), \quad (i = \overline{0, n-1}; \quad j = \overline{0, n}); \quad V_n^k = 0 \tag{12}$$

Order k of iteration expressed by Eq. (12) gives finite values only for paths with length at most $k - 1$, arriving at x_n, choosing between them minimal ones. From ones iteration to the next:

$$V_i^k \leq V_{i-1}^k, \forall j \tag{13}$$

Numbers V_i^k $(i \neq n; k = 0, 1, \ldots)$ form monotone decreasing strings that necessarily reach a minimum after a finite number of iterations that not exceed $n - 1$.

So, the algorithm stops when it comes to an iteration k such that $V_i^k = V_i^{k+1}$, $(i = \overline{0, n})$ and the value of the shorted path between vertex x_0 and x_n is $V_o^k = V_o^{k+1}$.

In order to identify the paths that have found minimum values, shall be deducted from Eq. (12) that along them, at the last iteration:

$$V_i^k = m_{ij} + V_j^{k-1} = m_{ij} + V_j^k \tag{14}$$

Based on the described optimization model a computer program OPTRAD was performed in FORTRAN programming language for PC compatible Microsystems, with the flow chart shown in Fig. 1, where: N is the graph order; $\mathbf{M}(I, J)$—matrix associated to the graph; $\mathbf{V}(I, J)$—column vector built for each iteration k; $X(I)$—vertices succession of path with minimum value; *VAL* – value of the shorted path in graph.

Bellman-Kalaba algorithm is not but the expression in another language of the optimality principle and analytical procedure for applying functional Eqs. (3) and (4) is only a special case, more elementary of this algorithm.

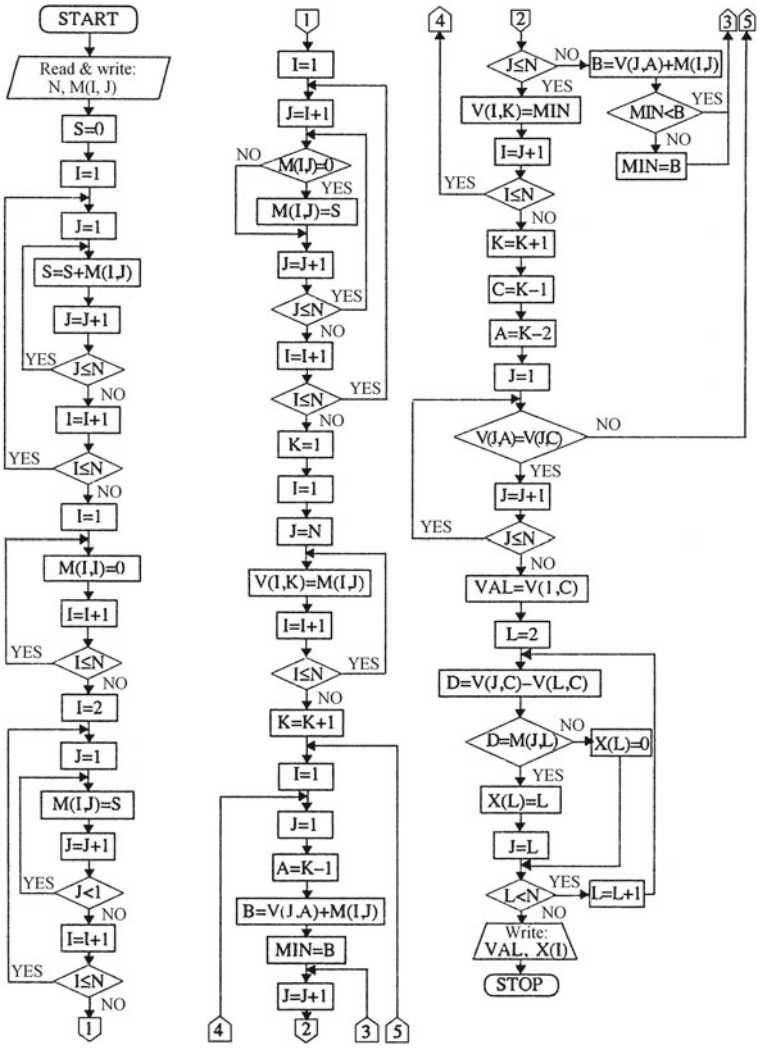


Fig. 1 Flow chart of computer program OPTRAD

2.3 Sensitivity of Optimal Solution

At a complex analysis that involve several optimization criteria in a sequentially program, is necessary to determine the optimal solution sensitivity. Exploration of optimal solution neighbourhoods leads to the problem of k-optimal solutions, i.e. at the determination of the solutions very closed of optimal one (quasi optimally).

It is sometimes necessary to optimize a single criterion when multiple optimal solutions are obtained. The choice of optimal solution can be done but calling another criteria or even several optimization criteria.

2.4 Numerical Applications

2.4.1 Application of Optimization Model Based on Dynamic Programming

For example, a water adduction main for a village L starting from two source locations S_1 and S_2 is considered (Fig. 2). Possible paths pass through critical points A, B, C, D, E, forming three sectors. Could be applied dynamic programming model to solve the selection problem of source location and of the path for this water adduction main.

Adopting as optimization criterion the minimum total investment cost, partial investment is determined for each path and sequentially graph is plotted in Fig. 3, in which each edge have associated a cost in conventional units. They noted with X_0, X_1, X_2 and X_3 decision variables related to the each sector. These variables will not take numerical values, but will be vertices in that graph which are on the same alignment.

The cost of sector 1 is noted $V_{\mathrm{I}}(X_0, X_1)$ and depend on values of X_0 and X_1 (in this case, X_0 could be only L). Identically are noted $V_{\mathrm{II}}(X_1, X_2)$ and $V_{\mathrm{III}}(X_2, X_3)$.

The total value of adduction system is expressed by:

$$Z = V_{\mathrm{I}}(X_0, X_1) + V_{\mathrm{II}}(X_1, X_2) + V_{\mathrm{III}}(X_2 + X_3) \tag{15}$$

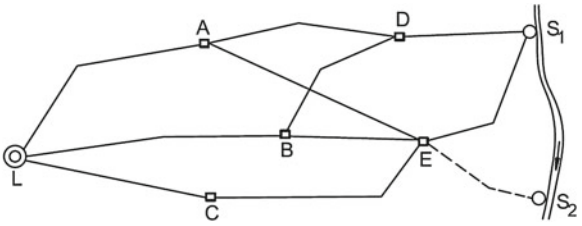


Fig. 2 Variants of adduction path

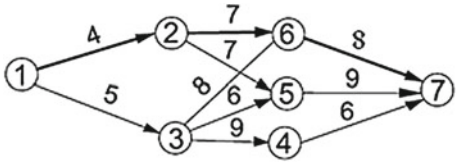


Fig. 3 Sequential graph of possible adduction paths

Initially, the adduction main is considered developed on a single sector (I). Noting with $f_I(X_1)$ the minimum cost of sector I, for each of critical points A, B, C, Eq. (3) becomes:

$$f_{\mathrm{I}}(A) = V_{\mathrm{I}}(L, A) = 8; \quad f_{\mathrm{I}}(B) = V_{\mathrm{I}}(L, B) = 9; \quad f_{\mathrm{I}}(C) = V_{\mathrm{I}}(L, C) = 6 \quad (16)$$

The operator "min" is missing because step zero does not exist. It is considered that the adduction main is formed of two sections (I, II), is denoted by $f_{\mathrm{I,\,II}}(X_2)$ the minimum cost for sectors I and II together, for different values of X_2 Eq. (3) becomes:

$$\begin{aligned} f_{\mathrm{I,\,II}}(D) &= \min_{X_1=A,B,C} [V_{\mathrm{II}}(X_1, D) + f_{\mathrm{I}}(X_1)] \\ f_{\mathrm{I,\,II}}(E) &= \min_{X_1=A,B,C} [V_{\mathrm{II}}(X_1, E) + f_{\mathrm{I}}(X_1)] \end{aligned} \quad (17)$$

Giving successively to X_1 the values A, B, C and taking into account that for inexistent links are considered ∞ value, from previous relationships results:

$$\begin{aligned} f_{\mathrm{I,\,II}}(D) &= \min[\underset{X_1=A}{7+8}, \underset{X_1=B}{7+9}, \underset{X_1=C}{\infty+6}] = 15, \quad \text{for } X_1 = A; \\ f_{\mathrm{I,\,II}}(E) &= \min[\underset{X_1=A}{8+8}, \underset{X_1=B}{6+9}, \underset{X_1=C}{9+6}] = 15, \quad \text{for } X_1 = C \end{aligned} \quad (18)$$

At the last step is considered that the adduction main is developed on sectors I, II and III. According with optimality principle the minimum cost for sectors I, II and III together, corresponding to values S_1 and S_2 for X_3, is written as:

$$\begin{aligned} f_{\mathrm{I,\,II,\,III}}(S_1) &= \min_{X_2=D,\,E} [V_{\mathrm{III}}(X_2, S_1) + f_{\mathrm{I,\,II}}(X_2)]; \\ f_{\mathrm{I,\,II,\,III}}(S_2) &= \min_{X_2=D,\,E} [V_{\mathrm{III}}(X_2, S_2) + f_{\mathrm{I,\,II}}(X_2) \end{aligned} \quad (19)$$

obtaining for the given example:

$$\begin{aligned} f_{\mathrm{I,\,II,\,III}}(S_1) &= \min[\underset{X_2=D}{4+15}, \underset{X_2=E}{5+15}] = 19, \quad \text{for } X_2 = D; \\ f_{\mathrm{I,\,II,\,III}}(S_2) &= \min[\underset{X_2=D}{\infty+15}, \underset{X_2=E}{6+15}] = 21, \quad \text{for } X_2 = E \end{aligned} \quad (20)$$

It results as optimal solution the source location in S_1 and the water adduction main path: S_1, D, A, L.

Because of the analytical computation procedure is less intuitive a table calculation procedure is presented. So, in Table 1 are shown in column 0 the sector, in column 1 the vertex which marks the start of sector, in column 2 the vertex which marks the end of sector, in column 3 the value of corresponding edge, in column 4 the value of previous minimum sub-policy, and in column 5 the sum of columns 3 and 4.

Table 1 Calculation procedure for optimal path

Sector	X_n^i	X_{n+1}^i	$\lambda\left(X_n^i, X_{n+1}^i\right)$	Value of previous minimum sub-policy	(3 + 4)
0	1	2	3	4	5
I	*A*	L	8	0	8
	B	L	9	0	9
	C	L	6	0	6
II	*D*	A	7	8	15
	D	B	7	9	16
	E	A	8	8	16
	E	B	6	9	15
	E	C	9	6	15
III	S_1	D	4	15	**19**
	S_1	E	5	15	20
	S_2	E	6	15	21

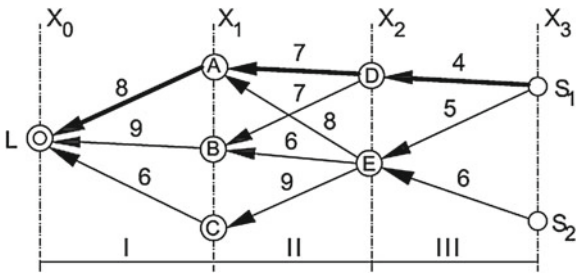


Fig. 4 Graph of adduction paths

After fulfilling the entire table, is read the minimum value from column 5 of the sector III, which is 19, the bolded number. From this, it starts with a horizontal arrow to column 4 and is obtained 15; from here it starts again back in column 5 and up, to the position where he was transferred this number, and so on. The arrows show the optimal solution. The path is read from the column 1 with the first vertex S_1, continue in column 2 on the position of horizontal arrow: S_1, D, A, L (vertices that were underlined in table).

2.4.2 Application of Optimization Model Based on Graph-Theory

It is presented a numerical example of application of optimizing model based on the graph theory to determine the optimal path of water adduction main for village L, starting from source S_1.

It is plotted in Fig. 4, the oriented connected graph of order $n = 7$, formed of source as origin, paths as edges and critical points as vertices.

The matrix **M** assigned to graph has the elements m_{ij} defined with the Eqs. (5):

$$\mathbf{M} = \begin{bmatrix} 0 & 4 & 5 & \infty & \infty & \infty & \infty \\ \infty & 0 & \infty & \infty & 7 & 7 & \infty \\ \infty & \infty & 0 & 9 & 6 & 8 & \infty \\ \infty & \infty & \infty & 0 & \infty & \infty & 6 \\ \infty & \infty & \infty & \infty & 0 & \infty & 9 \\ \infty & \infty & \infty & \infty & \infty & 0 & 8 \\ \infty & \infty & \infty & \infty & \infty & \infty & 0 \end{bmatrix} \tag{21}$$

Using computer program OPTRAD it was determined minimum path in graph: (1, 2, 6, 7), having the value 19. It results that the optimal path of water adduction main is S_1, D, A, L.

3 Path Optimization Model for Water Branched Networks

3.1 Computational Algorithm Based on Graph-Theory

A distribution network may be represented by a directed connected graph G comprising a finite number of edges (pipes) connected to one another by vertices. At the end of each edge are vertices with known energy grade (fixed-grade nodes) or external water consumption (junction nodes). Water flow through the edges and can enter or exit the graph at any vertex.

The main optimization criteria that can be used are:

- Minimum capital cost ($\Sigma c_{ij} L_{ij} \rightarrow \min$);
- Minimum length of the path ($\Sigma L_{ij} \rightarrow \min$);
- Minimum transport work ($\Sigma L_{ij} Q_{ij} \rightarrow \min$);
- Minimum head loss ($\Sigma R_{ij} Q_{ij}^2 \rightarrow \min$),

in which L_{ij}, Q_{ij} are the length and the flow rate of the pipe ij (between nodes i and j); c_{ij} is the specific capital cost; R_{ij} is the hydraulic resistance of pipe ij.

It is considered undirected connected graph $G = (X, U)$, where $X = \{1, 2, \ldots, n\}$ represent the set of vertices indexes and U—the set of edges. Each edge $u_j^i \in U$ has an associated value $\lambda\ (u_j^i) > 0$, in conventional units, according to the adopted optimization criterion. This graph has attached a matrix **C** of order n whose elements are:

$$c_{ij} = \begin{cases} \lambda\ (u_j^i), & \text{if } u_j^i \in U \\ \infty, & \text{if } u_j^i \notin U, \text{ or } i = j \end{cases} \tag{22}$$

where u_j^i is the edge u with one end in vertex i and the other one in vertex j.

If all $\lambda\ (u_j^i)$ values for $u_j^i \in U$ are distinct, the problem has a single solution. If there are values $\lambda\ (u_j^i)$ equal to more edges, the problem may have several solutions.

3.1.1 Partial Graph Determination of the Minimal Trees

The indexes of edge vertices u_j^i, belonging to the partial graph of the minimal trees, and the corresponding values c_{ij} are retained in a matrix **M** with 3 columns and $n-1$ rows, because a minimal tree that has $n-1$ edges. This matrix is performed in the following steps:

(S1) Minimal elements from each row of the matrix **C** are determined.
(S2) It is chosen one of the rows for which minimal element is unique and is denoted by r, s its indexes.
(S3) Values r, s, c_{rs} are stored on the first row of matrix **M**, and c_{rs}, c_{sr} elements from matrix **C** are marked.
(S4) It is determined the minimum unmarked element of **C** matrix rows on which exist at least a marked element, and is denoted by *r,s* the indexes of the minimum element or of one of them if there are more.
(S5) If on row s of the matrix **C** there are marked elements, are marked also elements c_{rs}, c_{sr} and proceed to step S6. If on row s of the matrix **C** there are not marked elements, for each c_{is} element located on a row marked and equal with c_{rs} is added to matrix **M** a new row which consists of the values i, s, c_{is}. Shall be marked elements c_{is}, c_{si} from matrix **C**, then proceed to step S6.
(S6) Proceed to step S4 or stop calculations as still exist or not in the matrix **C** rows with no unmarked element.

3.1.2 Generation of Minimal Trees

If for the matrix **M** result a row number bigger than considered graph order, the minimal tree problem has several solutions.

In this case first are permuted the rows of the matrix **M** so that the second column elements are ordered in ascending order. In the second column will find only indexes of $n-1$ vertices of the graph. If are denoted absolute frequencies of these indexes on the second column of the matrix **M** with $F_1, F_2, \ldots, F_n$, and added frequencies with $F_1^*, F_2^*, \ldots, F_n^*$ $(i = 1, 2, \ldots, n-1)$, the number n_a of minimal trees of the graph is given by:

$$n_a = \prod_{i=1}^{n-1} F_i \tag{23}$$

To identify the n_a minimal trees are performed a matrix **A** with n-1 rows and n_a columns in the following steps:

(S7) It is associated to each absolute frequency F_i a variable V_i whose values are $N_i = \left(F^*_{i-1}, F_{i-1}\right] \cap \{N\}$, $(i = 1, 2, \ldots, n-1)$ set elements, where $\{N\}$ is the natural number system, and $F^*_0 = 0$.

(S8) On each row i of the matrix $\mathbf{A}$ are inscribed n_a / F_i times the elements V_i of the set N_i, such as matrix $\mathbf{A}$ finally obtained have the columns lexicographically ordered.

On each column of the matrix $\mathbf{A}$ are the indexes of $\mathbf{M}$ matrix rows containing edge characteristics of one optimal tree.

If obtained solution is multiple, the choosing of optimal solution is performed taking into account other criteria.

3.2 Computer Program OTREDIRA

Based on above developed algorithm was elaborated OTREDIRA computer program in FORTRAN programming language for PC Microsystems. This has flow chart in Fig. 5, where: N is the graph order attached to the network; $\mathbf{C}(I, J)$ is the criterion matrix associated to the topological graph of the network; *MIN(I)* is the minimal element from row I; *NEM(I)* is the number of minimal elements from the row I; *MAR(I)* is the number of marked elements from the row I; *MINLM* is the minimum of the marked elements from the rows containing marked elements $\mathbf{A}(I, J)$ is the helpful matrix in identifying minimal trees; $\mathbf{M}(K, L)$ is the matrix of optimal edge vertices and of corresponding values.

3.3 Numerical Application

It is exemplified application of proposed optimization model to determine the optimal path of a pipe network with possible path graph of order $n = 9$ (Fig. 6). This graph has attached $\mathbf{C}$ capital cost matrix with the elements expressed in conventional units:

$$\mathbf{C} = \begin{bmatrix} \infty & 98 & 83 & 88 & \infty & \infty & \infty & \infty & \infty \\ 98 & \infty & 85 & \infty & \infty & 73 & 68 & \infty & \infty \\ 83 & 85 & \infty & 33 & 33 & 48 & 90 & \infty & \infty \\ 88 & \infty & 33 & \infty & 33 & \infty & \infty & 93 & \infty \\ \infty & \infty & 33 & 33 & \infty & 33 & \infty & 43 & \infty \\ \infty & 73 & 48 & \infty & 33 & \infty & 38 & 43 & \infty \\ \infty & 68 & 90 & \infty & \infty & 38 & \infty & 98 & 40 \\ \infty & \infty & \infty & 93 & 43 & 43 & 98 & \infty & 95 \\ \infty & \infty & \infty & \infty & \infty & \infty & 40 & 95 & \infty \end{bmatrix} \tag{24}$$

Using OTREDIRA computer program, the following results were obtained:

- The number of minimal trees n_a=2.

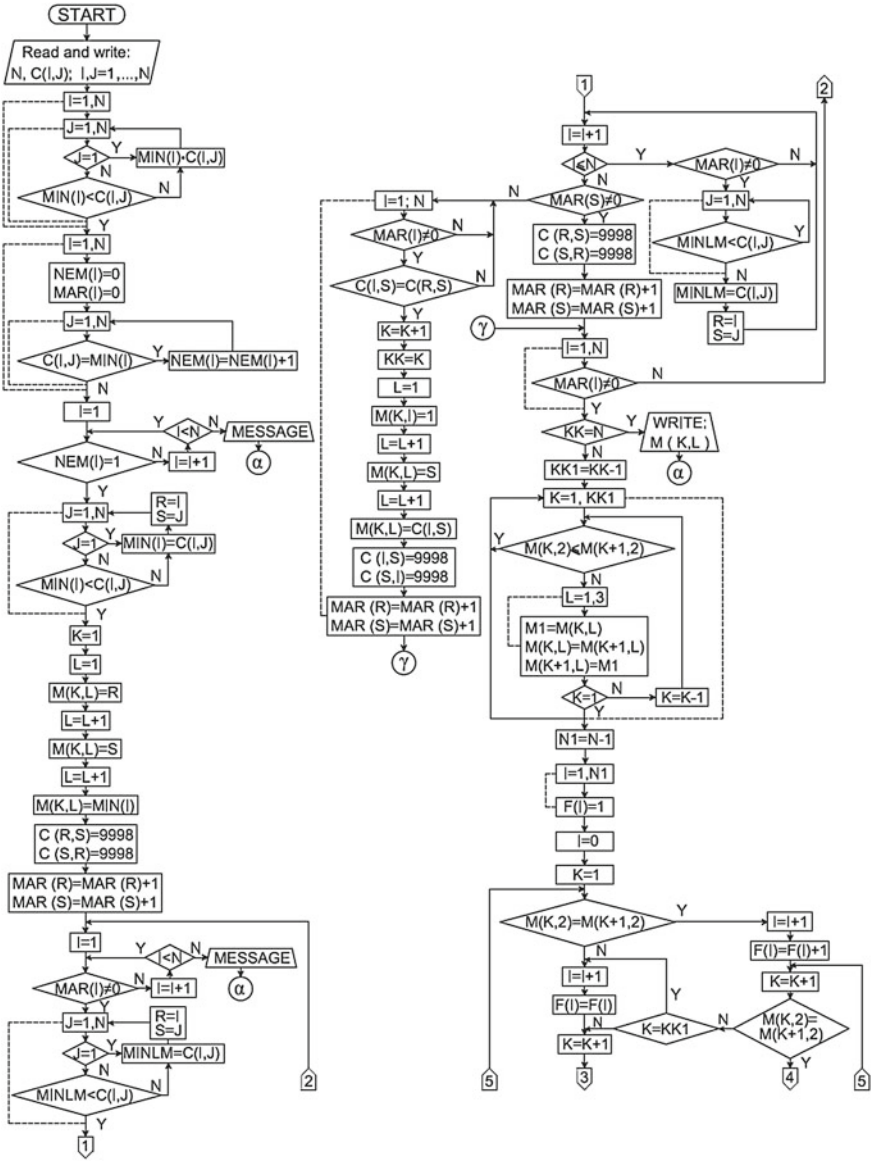


Fig. 5 Flow chart of computer program OTREDIRA

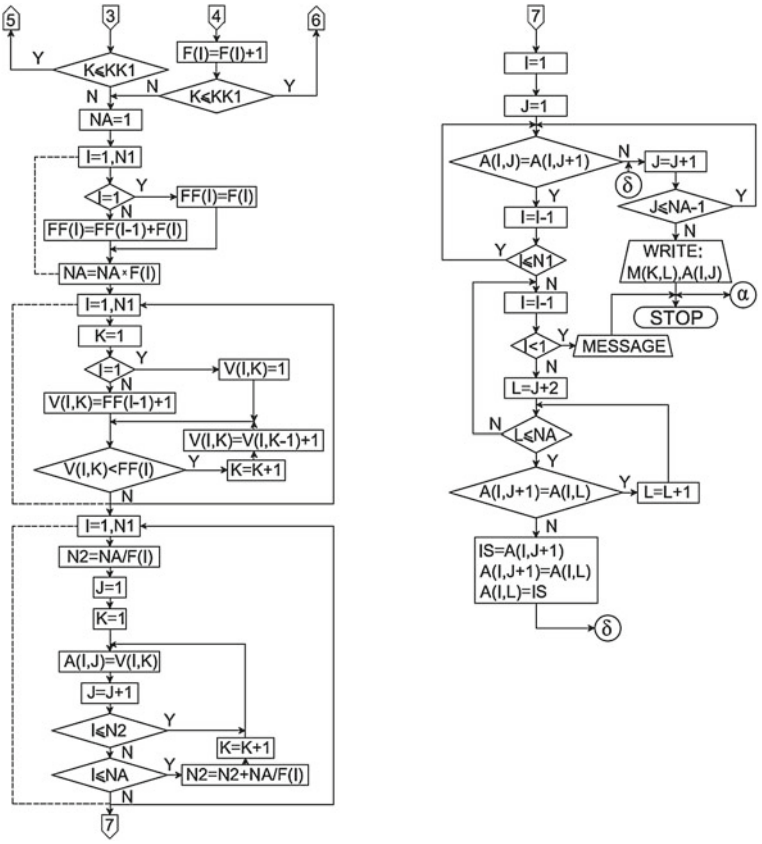


Fig. 5 (continued.)

- The matrix **M** of optimal edge vertices and the corresponding values:

$$\mathbf{M} = \begin{bmatrix} 7\ 2\ 68 \\ 1\ 3\ 83 \\ 3\ 4\ 33 \\ 5\ 4\ 33 \\ 3\ 5\ 33 \\ 5\ 6\ 33 \\ 6\ 7\ 38 \\ 5\ 8\ 33 \\ 7\ 9\ 40 \end{bmatrix} \tag{25}$$

- The matrix **A** on whose columns are found the indexes for the rows of the ordered matrix **M** containing the edges of one of the minimal trees:

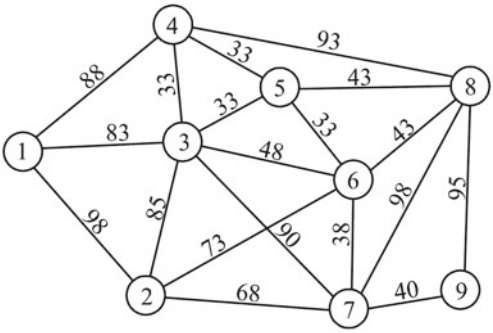


Fig. 6 Graph of network paths

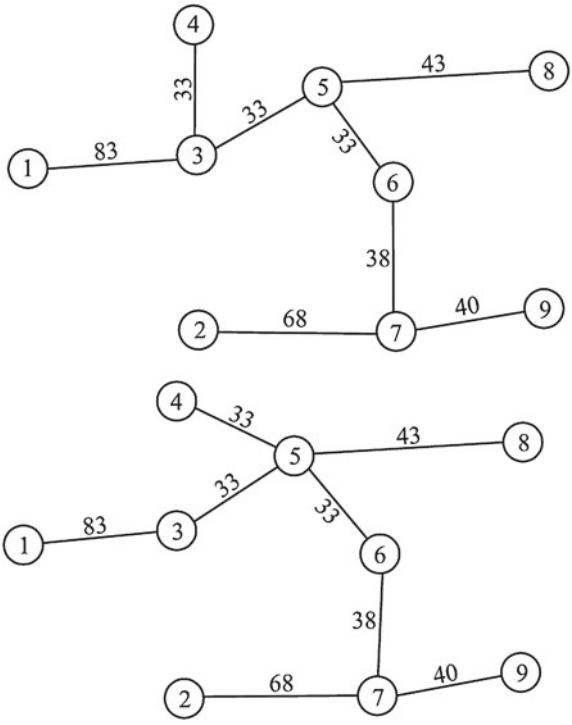


Fig. 7 Optimal paths of distribution network

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 2 & 2 \\ 3 & 4 \\ 5 & 5 \\ 6 & 6 \\ 7 & 7 \\ 8 & 8 \\ 9 & 9 \end{bmatrix} \tag{26}$$

Optimal path solutions, applying the minimum capital cost criterion, are illustrated in Fig. 7. Since obtained optimal solution is multiple, decision to adopt one of these alternatives is made considering other criteria too.

4 Conclusions

In the presented optimization models were described two techniques of operational research for solving sequential programs. Applying the dynamic programming in determinist case at built of an adduction main could be established optimal path for more possible variants.

The choice of the optimal path for branched pipe networks must be performs taking into account several criteria. The optimization model is applicable to design of the distribution networks for hydro-urban and hydro-amelioration systems.

Using the proposed optimization models leads to savings of the pipelines, earthworks, and electricity due to shortening the path and a more even distribution of the pressures in the system.

The elaborated computer programs OPTRAD and OTREDIRA based on the graph-theory allows performing a quick and efficient computation.

References

1. Walski TM, Chase DV, Savic DA, Grayman W, Beckwith S, Koelle E (2003) Advanced water distribution modeling and management. Haestad Press, Waterbury
2. Sarbu I (1997) Energetically optimization of water distribution systems. Publishing House of Romanian Academy, Bucharest
3. Sarbu I (2010) Numerical modelling and optimization in building services. Politehnica Publishing House, Timisoara
4. Dantzig GB (1967) All shortest routes in graph. In: Rosenthier P (ed) Theory of graph. Gordan and Breach, New York
5. Polak E (1971) Computational methods in optimization. Academic Press, New York
6. Stefanescu A, Zidaroiu C (1981) Operational researches. Teaching and Pedagogical Publishing House, Bucharest
7. Foulds LR (1992) Graph theory applications. Springer, New York
8. Gross JL, Tucker TW (2001) Topological graph theory. Wiley, Portland

9. Diestel R (2005) Graph theory. Springer, New York
10. Bellman R (1951) Dynamic programming. Princeton University Press, Princeton
11. Bellman RE (2003) Dynamic programming. Dover Publications, New York
12. Dreyfus S (2002) Richard Bellman on the birth of dynamic programming. Oper Res 50(1):48–51
13. Bazaraa MS, Jarvis JJ, Sherali HD (1990) Linear programming and network flows. Wiley, New York
14. Labye Y (1966) Étude des procédés de calcul ayant pour but le rendre minimal la coût d'un réseaux de distribution d'eau sous pression. La Houille Blanche 5:577–583
15. Bertsekas DP (2005) Dynamic programming and optimal control. Athena Scientific, Belmont
16. Sterling MJ, Coulbeck B (1973) A dynamic programming solution to optimization of pumping costs. Proc Inst Civil Eng 59(2):813
17. Coulbeck B (1984) Optimization of water networks. Trans Inst Meas Control 6(5):271
18. Sabel MH, Helwing OJ (1985) Cost effective operation of urban water supply system using dynamic programming. Water Res Bull 21(1):75
19. Lansey KE, Awumah K (1994) Optimal pump operation considering pump switches. J Water Res Plan Manage, ASCE 120(1):17
20. Liang T (1971) Design conduit system by dynamic programming. J Hydraul Div, ASCE 97(HY3):383–393
21. Yang P, Liang T, Wu IP (1975) Design of conduit system with diverging branches. J Hydraul Div, ASCE 101(HY1):167–188
22. Martin QW (1980) Optimal design of water conveyance systems. J Hydraul Div, ASCE, 106(HY9):1415–1433
23. Sarbu I, Ostafe G, Valea ES (2013) Application of operational research to determine optimal path for a water transmission main, Lecture Notes in Engineering and Computer Science. In: Proceedings of the international multiconference of engineers and computer scientists 2013, 13–15 March, 2013, Hong Kong, pp 1146–1150
24. Sarbu I, Valea E (2013) Optimization of water distribution networks path. J Eng Appl Sci 8(5):333–337
25. Kaufmann A (1963) Methods and models of operations research. Prentice Hall Inc, Englewood Cliffs
26. Rafiroiu M (1980) Models of operational researches applied in civil engineering. Technical Publishing House, Bucharest
27. Berge C (1962) Theory of graphs and its applications. Wiley, New York

Hierarchical Equilibrium and Generalized Variational Inequality Problems

Nopparat Wairojjana and Poom Kumam

Abstract In this paper, we study the convex feasibility problem (CFP) in the case that each is a solution set of the generalized variational inequality and the equilibrium problem and introduce a new approach method to find a common element in the intersection of the set of the solutions of a finite family of equilibrium problems and the intersection of the set of the solutions of a finite family of generalized variational inequality problems in a real Hilbert space which is a unique solution of the hierarchical equilibrium and generalized variational inequality problems(HEGVIP). Under appropriate conditions, some strong convergence theorems are established. Our results generalize and improve the corresponding results of Wairojjana and Kumam (2013) [27] and some authors.

Keywords Convex feasibility problem · Equilibrium problem · Generalized variational inequality problem · Hierarchical fixed point problem · Hierarchical generalized variational inequality problem · Hilbert space · Strong convergence

1 Introduction

Let H be a real Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$. Let C be a nonempty closed convex subset of H and let P_C be the *metric projection* of H onto C. We denote weak convergence and strong convergence by notations $\rightharpoonup$ and $\to$, respectively.

N. Wairojjana (✉) · P. Kumam
Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi (KMUTT), 126 Pracha-Uthit Road, Bang Mod, Thung Khru, Bangkok 10140, Thailand
e-mail: poom.kum@kmutt.ac.th

N. Wairojjana
e-mail: noparatw@windowslive.com

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 275, DOI: 10.1007/978-94-007-7684-5_24,

A mapping $f : C \to C$ is called k-contraction if there exists a positive real number $k \in (0, 1)$ such that $\|f(x) - f(y)\| \leq k\|x - y\|, \forall x, y \in C$. A mapping $T : C \to C$ is called *nonexpansive* if $\|Tx - Ty\| \leq \|x - y\|, \forall x, y \in C$. We use $F(T)$ to denote the set of *fixed points* of T, that is, $F(T) = \{x \in C : Tx = x\}$. It is well known that $F(T)$ is a closed convex set, if T is *nonexpansive*.

Consider the set of solutions of the following *generalized variational inequality:* given nonlinear mappings $A, B : C \to H$ find $x \in C$ such that

$$\langle x - \hat{\lambda}Bx + \lambda Ax, x - y\rangle \geq 0, \quad \forall y \in C, \tag{1}$$

where $\hat{\lambda}$ and λ are two positive constants. We use $GVI(C, B, A)$ to denote *the set of solutions of the generalized variational inequality* (1). It is easy to see that an element $x \in C$ is a solution to the variational inequality (1) if and only if x is a fixed point of the mapping $P_C(\hat{\lambda}B - \lambda A)$, that is $x = P_C(\hat{\lambda}B - \lambda A)x \Leftrightarrow x \in GVI(C, B, A)$. Therefore, fixed point algorithms can be applied to solve $GVI(C, B, A)$. Next, we consider a special case of (1). If $B = I$, the identity mapping and $\hat{\lambda} = 1$, then the generalized variational inequality (1) is reduced to *the variational inequality* as follow: find $x \in C$ such that

$$\langle Ax, x - y\rangle \geq 0, \quad \forall y \in C. \tag{2}$$

We use $VI(C, A)$ to denote the set of solutions of the variational inequality (2). It is well known that the variational inequality theory has emerged as an important tool in studying a wide class of obstacle, unilateral, and equilibrium problems; which arise in several branches of pure and applied sciences in a unified and general framework. Several numerical methods have been developed for solving variational inequalities and related optimization problems, see [2–13] and the references therein.

A hierarchical generalized variational inequality problems (HGVIP) is the problem of finding a point $\tilde{x} \in GVI(C, B, A)$ such that $\langle F\tilde{x}, x - \tilde{x}\rangle \leq 0, \quad \forall x \in GVI(C, B, A)$, where $GVI(C, B, A)$ is the solution set of the generalized variational inequality. If If the set $GVI(C, B, A)$ is replaced by the set $VI(C, A)$, the solution set of the variational inequality, then the HGVIP is called a hierarchical variational inequality problems (HVIP). Many problems in mathematics, for example the signal recovery [14], the power control problem [15] and the beamforming problem [16] can be modeled as HGVIP.

Let $F : C \times C \longrightarrow \Re$ be a bifunction. The *equilibrium problem* for finding $x \in C$ such that

$$F(x, y) \geq 0, \quad \forall y \in C. \tag{3}$$

The set of solutions of (3) is denoted by $EP(F)$, that is,

$$EP(F) = \big\{ x \in C : F(x, y) \geq 0, \quad \forall y \in C \big\}. \tag{4}$$

Given a mapping $A : C \to H$, let $F(x, y) = \langle Ax, y - x\rangle$ for all $x, y \in C$. Then, $z \in EP(F)$ if and only if $\langle Az, y - z\rangle \geq 0$ for all $y \in C$, that is, z is a solution of the

variational inequality. Numerous problems in physics, optimization, and economics reduce to find a solution of (3); see, for example [17–21] and the references therein.

A convex feasibility problem (CFP) is the problem of finding a point in the intersection of finitely many closed convex sets in a real Hilbert spaces H. That is, finding an $x \in \cap_{m=1}^{r} C_m$, where $r \geq 1$ is an integer and each C_m is a nonempty closed and convex subset of H. Many problems in mathematics, for example in physical sciences, in engineering and in real-world applications of various technological innovations can be modeled as CFP. There is a considerable investigation on CFP in the setting of Hilbert spaces which captures applications in various disciplines such as image restoration [22, 23] computer tomography [24] and radiation therapy treatment planning [25].

In 2011, He and Du [26] introduced the iteration as follows: a sequence $\{x_n\}$ defined by

$$\begin{cases} x_1 \in C, \\ u_n^i = T_{r_n}^i x_n, \qquad \forall i = 1, 2, \ldots, l, \\ z_n = \frac{u_n^1 + u_n^2 + \cdots + u_n^l}{l}, \\ y_n = (1-\lambda)x_n + \lambda T z_n, \\ x_{n+1} = \alpha_n f(x_n) + (1-\alpha_n) y_n, \end{cases} \tag{5}$$

where $T_{r_n}^i x = \left\{ z \in C : F_i(z, y) + \frac{1}{r_n} < y - z, z - x > \geq 0, \forall y \in C \right\}$, $\alpha_n \subset (0, 1)$ and $r_n \subset (0, +\infty)$. They proved that if the sequence $\{\alpha_n\}$ and $\{r_n\}$ of parameters satisfies appropriate conditions and $\Omega = (\cap_{i=1}^{l} EP(F_i)) \cap F(T) \neq \emptyset$, then the sequence $\{x_n\}$ generated by (5) strongly converges to the unique solution $c = P_\Omega f(c) \in \Omega$.

In 2013, Wairojjana and Kumam [27] introduced the iteration as follows: a sequence $\{x_n\}$ defined by

$$x_{n+1} = \alpha_n \gamma f(x_n) + (I - \alpha_n \mu G) \sum_{m=1}^{r} \beta_{(m,n)} P_C(\hat{\lambda}_m B_m x_n - \lambda_m A_m x_n), \quad \forall n \geq 1, \tag{6}$$

where $\{\alpha_n\}$ and $\{\beta_{(i,n)}\}$ are sequences in $(0, 1)$ for all $i = 12, \ldots r$. They proved that if the sequence $\{\alpha_n\}$ and $\{\beta_{(i,n)}\}$ of parameters satisfies appropriate conditions and $\Omega = \cap_{m=1}^{r} GVI(C, B_m, A_m) \neq \emptyset$, then the sequence $\{x_n\}$ generated by (6) strongly converges to the unique solution $\tilde{x} \in \Omega$, which is the unique solution of the HGVIP:

$$\langle (\gamma f - \mu G)\tilde{x}, x - \tilde{x} \rangle \leq 0, \quad \forall x \in \Omega. \tag{7}$$

Motivated and inspired by He and Du 's results and Wairojjana and Kumam's results, we consider and study the CFP in the case that each C_m is a solution set of the generalized variational inequality $GVI(C, B_m, A_m)$ and the equilibrium problem $EP(F_i)$ and are devoted to solve the following the hierarchical equilibrium and generalized variational inequality problems (HEGVIP): find $c \in (\cap_{i=1}^{l} EP(F_i)) \cap (\cap_{m=1}^{r} GVI(C, B_m, A_m)) = \Omega$ such that

$$\langle (\gamma f - \mu G)c, x - c \rangle \leq 0, \quad \forall x \in \Omega. \tag{8}$$

The HEGVIP is general than the HGVIP. Consequently, we prove a strong convergence theorem for finding a point c which is a unique solution of the HEGVIP.

2 Preliminaries

This section collects some definitions and lemmas which be use in the proofs for the main results in the next section. Some of them are known; others are not hard to derive.

Definition 1 Let $A : C \to H$ and $G : C \to C$ be a nonlinear mappings. Recall the following definitions:

(a) G is said to be L-Lipschitzian on C if there exists a positive real number $L > 0$ such that $\|G(x) - G(y)\| \leq L\|x - y\|, \forall x, y \in C$.
(b) A is said to be monotone if $\langle Ax - Ay, x - y\rangle \geq 0, \quad \forall x, y \in C$.
(c) A is said to be ρ-strongly monotone if there exists a positive real number $\rho > 0$ such that $\langle Ax - Ay, x - y\rangle \geq \rho\|x - y\|^2, \quad \forall x, y \in C$.
(d) A is said to be η-cocoercive if there exists a positive real number $\eta > 0$ such that $\langle Ax - Ay, x - y\rangle \geq \eta\|Ax - Ay\|^2, \quad \forall x, y \in C$.
(e) A is said to be relaxed η-cocoercive if there exists a positive real number $\eta > 0$ such that $\langle Ax - Ay, x - y\rangle \geq (-\eta)\|Ax - Ay\|^2, \quad \forall x, y \in C$.
(f) A is said to be relaxed (η, ρ)-cocoercive if there exists a positive real number $\eta, \rho > 0$ such that $\langle Ax - Ay, x - y\rangle \geq (-\eta)\|Ax - Ay\|^2 + \rho\|x - y\|^2, \quad \forall x, y \in C$.

Lemma 1 [28]. *Let H be a Hilbert space, C be a closed convex subset of H and $T : C \to C$ be a nonexpansive mapping with $F(T) \neq \emptyset$. If $\{x_n\}$ is a sequence in C weakly converging to x and if $\{(I-T)x_n\}$ converges strongly to y, then $(I-T)x = y$; in particular, if $y = 0$ then $x \in F(T)$.*

Lemma 2 [29]. *Let C be a nonempty closed and convex subset of a real Hilbert space H. Let $S_1 : C \to C$ and $S_2 : C \to C$ be nonexpansive mappings on C. Suppose that $F(S_1) \cap F(S_2)$ is nonempty. Define a mapping $S : C \to C$ by $Sx = aS_1 + (1 - a)S_2, \quad \forall x \in C$, where a is a constant $\in (0, 1)$. Then S is nonexpansive with $F(S) = F(S_1) \cap F(S_2)$.*

Lemma 3 [30]. *Let $F : C \to C$ be a η-strongly monotone and L-Lipschitzian operator with $L > 0, \eta > 0$. Assume that $0 < \mu < 2\eta/L^2, \tau = \mu(\eta - \mu L^2/2)$ and $0 < t < 1$. Then $\|(I - \mu t F)x - (I - \mu t F)y\| \leq (1 - t\tau)\|x - y\|$.*

Lemma 4 [17]. *In a real Hilbert space H, we have the equations hold:*

1. $\|x + y\|^2 \leq \|x\|^2 + 2\langle y, x + y\rangle, \ \forall x, y \in H$;
2. $\|x + y\|^2 \geq \|x\|^2 + 2\langle y, x\rangle, \ \forall x, y \in H$.

Lemma 5 [31]. *Assume that $\{a_n\}$ is a sequence of nonnegative numbers such that*

$$a_{n+1} \leq (1-\gamma_n)a_n + \delta_n, \quad \forall n \geq 0,$$

where $\{\gamma_n\}$ is a sequence in $(0,1)$ and $\{\delta_n\}$ is a sequence in $\Re$ such that

1. $\sum_{n=1}^{\infty} \gamma_n = \infty$,
2. $\limsup_{n\to\infty} \frac{\delta_n}{\gamma_n} \leq 0$ *or* $\sum_{n=1}^{\infty} |\delta_n| < \infty$.

Then $\lim_{n\to\infty} a_n = 0$.

Lemma 6 [30]. *Let H be a real Hilbert space, $f : H \to H$ a contraction with coefficient $0 < k < 1$, and $G : H \to H$ a L-Lipschitzian continuous operator and ξ-strongly monotone operator with $L > 0$, $\xi > 0$. Then for $0 < \gamma < \mu\xi/k$,*

$$\langle x-y, (\mu G - \gamma f)x - (\mu G - \gamma f)y\rangle \geq (\mu\xi - \gamma k)\|x-y\|^2, \quad \forall x, y \in H. \quad (9)$$

That is, $\mu G - \gamma f$ is strongly monotone with coefficient $\mu\xi - \gamma k$.

Lemma 7 [32]. *Let C be a closed convex subset of H. Let $\{x_n\}$ be a bounded sequence in H. Assume that*

1. *The weak ω-limit set $\omega_w(x_n) \subset C$,*
2. *For each $z \in C$, $\lim_{n\to\infty} \|x_n - z\|$ exists.*

Then $\{x_n\}$ is weakly convergent to a point in C.

Lemma 8 [33]. *Let C be a nonempty closed convex subset of H and let $r > 0$ and $x \in H$. Let $F : C \times C \longrightarrow \Re$ satisfying*

(A1) $F(x,x) = 0$ for all $x \in C$;
(A2) F is monotone, i.e., $F(x,y) + F(y,x) \leq 0$ for all $x, y \in C$;
(A3) for each $x, y, z \in C$, $\lim_{t\downarrow 0} F(tz + (1-t)x, y) \leq F(x,y)$;
(A4) for each $x \in C$, $y \mapsto F(x,y)$ is convex and lower semicontinuous.

Then, there exists $z \in C$ such that

$$F(z,y) + \frac{1}{r}\langle y - z, z - x\rangle \geq 0, \quad \forall y \in C.$$

Lemma 9 [34]. *Assume that $F : C \times C \to \Re$ satisfies (A1)–(A4). For $r > 0$ and $x \in H$, define a mapping $T_r^F : H \to C$ as follows:*

$$T_r^F(x) = \left\{ z \in C : F(z,y) + \frac{1}{r}\langle y - z, z - x\rangle \geq 0, \quad \forall y \in C \right\}$$

for all $x \in H$. Then, the following hold:

1. T_r^F *is single-valued;*
2. T_r^F *is firmly nonexpansive, i.e.,* $\forall x, y \in H$, $\left\|T_r^F x - T_r^F y\right\|^2 \leq \left\langle T_r^F x - T_r^F y, x - y\right\rangle$;
3. $F(T_r^F) = EP(F)$; *and*
4. $EP(F)$ *is closed and convex.*

Lemma 10 [26]. *Let H be a real Hilbert space. Then for any $x_1, x_2, ..., x_k \in H$ and $a_1, a_2, ..., a_k \in [0, 1]$ with $\sum_{i=1}^k a_i = 1, k \in \aleph$, we have*

$$\left\|\sum_{i=1}^k a_i x_i\right\|^2 = \sum_{i=1}^k a_i \|x_i\|^2 - \sum_{i=1}^{k-1} \sum_{j=i+1}^k a_i a_j \|x_i - x_j\|^2.$$

3 Main Results

Let $I = \{1, 2, \ldots, l\}$ be a finite index set. For each $i \in I$, let F_i be a bi-function from $C \times C$ into $\Re$ satisfying (A1)–(A4). Denote $T_{r_n}^i : H \to C$ by

$$T_{r_n}^i(x) = \left\{z \in C : F_i(z, y) + \frac{1}{r_n}\langle y - z, z - x\rangle \geq 0, \forall y \in C\right\}.$$

Theorem 1 *Let C be a nonempty closed and convex subset of a real Hilbert space H such that $C \pm C \subset C$. For each $i \in I$, let F_i be a bi-function from $C \times C$ into $\Re$ satisfying (A1)–(A4). Let $f : C \to C$ be a contraction with coefficient $k \in (0, 1)$. Let $G : C \to C$ be a ξ-strongly monotone and L-Lipschitz continuous mapping. let $A_m : C \to H$ be a relaxed (η_m, ρ_m)-cocoercive and ν_m-Lipschitz continuous mapping and $B_m : C \to H$ be a relaxed $(\hat{\eta}_m, \hat{\rho}_m)$-cocoercive and $\hat{\nu}_m$-Lipschitz continuous mapping for each $1 \leq m \leq r$. Let $p_m = \sqrt{1 - 2\lambda_m\rho_m + \lambda_m^2\nu_m^2 + 2\lambda_m\eta_m\nu_m^2}$ and $q_m = \sqrt{1 - 2\hat{\lambda}_m\hat{\rho}_m + \hat{\lambda}_m^2\hat{\nu}_m^2 + 2\hat{\lambda}_m\hat{\eta}_m\hat{\nu}_m^2}$, where $\{\lambda_m\}$ and $\{\hat{\lambda}_m\}$ are two positive sequences for each $1 \leq m \leq r$. Assume that $\Omega = (\cap_{i=1}^l EP(F_i)) \cap (\cap_{m=1}^r GVI(C, B_m, A_m)) \neq \emptyset$, $\xi > 0, L > 0$, $0 < \mu < 2\xi/L^2$, $0 < \gamma < \mu(\xi - \mu L^2/2)/k = \pi/k$ and $p_m, q_m \in [0, \frac{1}{2})$, for each $1 \leq m \leq r$. Given $\{x_n\}$ is a sequence generated by*

$$\begin{cases} x_1 \in C, \\ u_n^i = T_{r_n}^i x_n, \quad \forall i \in I, \\ v_n = \frac{u_n^1 + u_n^2 + \cdots + u_n^l}{l}, \\ y_n^m = P_C(\hat{\lambda}_m B_m v_n - \lambda_m A_m v_n), \quad \forall m = 1, 2, \ldots, r, \\ x_{n+1} = \alpha_n \gamma f(x_n) + (I - \alpha_n \mu G)\sum_{m=1}^r \beta_{(m,n)} y_n^m, \quad \forall n \geq 1, \end{cases} \tag{10}$$

where $\{\alpha_n\}, \{\beta_{(m,n)}\} \subset (0,1), \forall 1 \leq m \leq r$ *and* $\{r_n\} \subset (0, +\infty)$ *satisfying the following conditions:*

(C1) $\lim_{n\to\infty} \alpha_n = 0, \sum_{n=1}^{\infty} \alpha_n = \infty$ *and* $\sum_{n=1}^{\infty} |\alpha_{n+1} - \alpha_n| < \infty$;
(C2) $\sum_{m=1}^{r} \beta_{(m,n)} = 1, \forall n \geq 1$, $\sum_{n=1}^{\infty} |\beta_{(m,n+1)} - \beta_{(m,n)}| < \infty$ *and* $\lim_{n\to\infty} \beta_{(m,n)} = \beta_m \in (0,1), \forall 1 \leq m \leq r$.
(C3) $\liminf_{n\to\infty} r_n > 0$ *and* $\sum_{n=1}^{\infty} |r_{n+1} - r_n| < \infty$;

Then the sequence $\{x_n\}$ *converges strongly to a common element* $c \in \Omega$, *which is the unique solution of the HEGVIP:*

$$\langle (\gamma f - \mu G)c, x - c\rangle \leq 0, \quad \forall x \in \Omega. \tag{11}$$

Proof We will proceed with the following steps.

Step 1: We prove that $P_C(\hat{\lambda}_m B_m - \lambda_m A_m)$ is a nonexpansive mapping.

Put $T_m = P_C(\hat{\lambda}_m B_m - \lambda_m A_m), \forall 1 \leq m \leq r$. $\forall x, y \in C$ and $1 \leq m \leq r$, we have

$$\|T_m x - T_m y\| \leq \|(x-y) - \lambda_m(A_m x - A_m y)\| + \|(x-y) - \hat{\lambda}_m(B_m x - B_m y)\|. \tag{12}$$

It follows from the assumption that each A_m is relaxed (η_m, ρ_m)-cocoercive and ν_m-Lipschitz continuous that $\|(x-y) - \lambda_m(A_m x - A_m y)\|^2 \leq p_m^2 \|x-y\|^2$. This shows that $\|(x-y) - \lambda_m(A_m x - A_m y)\| \leq p_m \|x-y\|$. In a similar way, we can obtain that $\|(x-y) - \hat{\lambda}_m(B_m x - B_m y)\| \leq q_m \|x-y\|$. So, we have

$$\|T_m x - T_m y\| \leq (p_m + q_m)\|x-y\| \leq \|x-y\|. \tag{13}$$

Hence T_m is a nonexpansive mapping and

$$F(T_m) = F(P_C(\hat{\lambda}_m B_m - \lambda_m A_m)) = GVI(C, B_m, A_m), \quad \forall 1 \leq m \leq r.$$

Put $S_n = \sum_{m=1}^{r} \beta_{(m,n)} T_m$. By Lemma 2, we conclude that S_n is a nonexpansive mapping and $F(S_n) = \cap_{m=1}^{r} GVI(C, B_m, A_m), \forall n \geq 1$. We can rewrite the algorithm as

$$x_{n+1} = \alpha_n \gamma f(x_n) + (I - \alpha_n \mu G) S_n v_n. \tag{14}$$

Step 2: We prove that the sequence $\{x_n\}$, $\{y_n^m\}$, $\{v_n\}$ and $\{u_n^i\}$ are bounded. Take $c \in \Omega$. For each $i \in I$, we have

$$\|u_n^i - c\| = \|T_{r_n}^i x_n - T_{r_n}^i c\| \leq \|x_n - c\|, \quad \forall n \geq 1. \tag{15}$$

From (10) and (15) we have

$$\|v_n - c\| \leq \|x_n - c\|, \quad \forall n \geq 1 \tag{16}$$

For each $1 \leq m \leq r$, we have

$$\|y_n^m - c\| \leq (p_m + q_n)\|v_n - c\| \leq \|x_n - c\| \tag{17}$$

From (10), (17) and Lemma 3, we have

$$\begin{aligned}\|x_{n+1} - c\| &\leq \alpha_n\|\gamma(f(x_n) - f(c)) + \gamma f(c) - \mu Gc\| \\ &\quad + (1 - \alpha_n\pi)\|\sum_{m=1}^{r}\beta_{(m,n)}y_n^m - c\| \\ &\leq max\left\{\|x_n - c\|, \frac{\|\gamma f(c) - \mu Gc\|}{\pi - \gamma k}\right\}, \quad \forall n \geq 1.\end{aligned}$$

By induction, we obtain $\|x_n - c\| \leq max\left\{\|x_1 - c\|, \frac{\|\gamma f(c)-\mu Gc\|}{\pi-\gamma k}\right\}, \forall n \geq 1$. Hence $\{x_n\}$ is bounded. Also, we know that $\{y_n^m\}$, $\{v_n\}$ and $\{u_n^i\}$ are all $\forall 1 \leq m \leq r, \forall 1 \leq i \leq l$. Since S_n is nonexpansive mappings for $n \geq 1$, we see that

$$\|S_n x_n - c\| \leq \|x_n - c\| \leq \max\left\{\|x_1 - c\|, \frac{\|\gamma f(c) - \mu Gc\|}{\pi - \gamma k}\right\}.$$

Therefore, $\{S_n x_n\}$ is bounded. Since G is a L-Lipschitz continuous mapping, we have $\|GS_n x_n - Gc\| \leq L\|x_n - c\| \leq \max\left\{L\|x_1 - c\|, L\frac{\|\gamma f(c)-\mu Gc\|}{\pi-\gamma k}\right\}$. Hence $\{GS_n x_n\}$ is bounded. Since f is contraction and $\{x_n\}$ is bounded, so $\{f(x_n)\}$ is bounded.

Step 3: We prove that $\lim_{n\to\infty}\|x_{n+1} - x_n\| = 0$.

From (10), we consider

$$\begin{aligned}x_{n+1} - x_n &= [(I - \alpha_n\mu G)\sum_{m=1}^{r}\beta_{(m,n)}y_n^m - (I - \alpha_n\mu G)\sum_{m=1}^{r}\beta_{(m,n-1)}y_{n-1}^m] \\ &\quad + (\alpha_n - \alpha_{n-1})\gamma f(x_{n-1}) + (\alpha_{n-1} - \alpha_n)\mu G\sum_{m=1}^{r}\beta_{(m,n-1)}y_{n-1}^m \\ &\quad + \alpha_n\gamma(f(x_n) - f(x_{n-1})),\end{aligned}$$

it follows that

$$\begin{aligned}\|x_{n+1} - x_n\| &\leq (1 - \alpha_n\pi)\|\sum_{m=1}^{r}\beta_{(m,n)}y_n^m - \sum_{m=1}^{r}\beta_{(m,n-1)}y_{n-1}^m\| \\ &\quad + \alpha_n\gamma k\|x_n - x_{n-1}\| + |\alpha_n - \alpha_{n-1}|M_1,\end{aligned} \tag{18}$$

where $M_1 = \sup_{n\geq 1}\{\gamma\|f(x_n)\| + \mu\|G\sum_{m=1}^{r}\beta_{(m,n)}y_n^m\|\}$, and

$$\|\sum_{m=1}^{r}\beta_{(m,n)}y_n^m-\sum_{m=1}^{r}\beta_{(m,n-1)}y_{n-1}^m\|\le M_2\sum_{m=1}^{r}|\beta_{(m,n)}-\beta_{(m,n-1)}|, \tag{19}$$

where $M_2=\max\{\sup_{n\ge1}\|y_n^m-y_{n-1}^m\|,\forall 1\le m\le r\}$. $\forall\in I$, since $u_{n-1}^i,u_n^i\in C$, we have

$$F_i(u_n^i,u_{n-1}^i)+\frac{1}{r_n}\langle u_{n-1}^i-u_n^i,u_n^i-x_n\rangle\ge0, \tag{20}$$

and

$$F_i(u_{n-1}^i,u_n^i)+\frac{1}{r_{n-1}}\langle u_n^i-u_{n-1}^i,u_{n-1}^i-x_{n-1}\rangle\ge0. \tag{21}$$

From (20), (21) and (A2), we see that

$$\begin{aligned}0&\le r_n[F_i(u_n^i,u_{n-1}^i)+F_i(u_{n-1}^i,u_n^i)]+\langle u_{n-1}^i-u_n^i,u_n^i-x_n-\frac{r_n}{r_{n-1}}(u_{n-1}^i-x_{n-1})\rangle\\&\le\langle u_{n-1}^i-u_n^i,u_n^i-x_n-\frac{r_n}{r_{n-1}}(u_{n-1}^i-x_{n-1})\rangle\end{aligned}$$

which implies

$$\langle u_{n-1}^i-u_n^i,u_{n-1}^i-u_n^i+x_n-x_{n-1}+x_{n-1}-u_{n-1}^i+\frac{r_n}{r_{n-1}}(u_{n-1}^i-x_{n-1})\rangle\le0. \tag{22}$$

It follows from (22) that

$$\|u_n^i-u_{n-1}^i\|\le\|x_n-x_{n-1}\|+|\frac{r_n-r_{n-1}}{r_{n-1}}|\|x_{n-1}-u_{n-1}^i\|,\quad\forall n\ge1. \tag{23}$$

Without loss of generality, let us assume that there exists a real number d such that $r_n>d>0$ for all $n\ge1$. Since $v_n=\frac{1}{l}(u_n^1+u_n^2+\ldots+u_n^l)$, by (23), we have

$$\|v_n-v_{n-1}\|\le\frac{1}{l}\sum_{i=1}^{l}\|u_n^i-u_{n-1}^i\|\le\|x_n-x_{n-1}\|+\frac{|r_n-r_{n-1}|}{d}M_3,\quad\forall n\ge1, \tag{24}$$

where $M_3=\max\{\sup_{n\ge1}\frac{1}{l}\sum_{i=1}^{l}\|x_{n-1}-u_{n-1}^i\|,\forall1\le i\le l\}$. From (18), (19) and (24), we have

$$\begin{aligned}\|x_{n+1}-x_n\|&\le(1-\alpha_n[\pi-\gamma k])\|x_n-x_{n-1}\|\\&\quad+\left[|\alpha_n-\alpha_{n-1}|+\sum_{m=1}^{r}|\beta_{(m,n)}-\beta_{(m,n-1)}|+\frac{|r_n-r_{n-1}|}{d}\right]M,\end{aligned} \tag{25}$$

where M is appropriate constant such that $M \geq \max\{M_1, M_2, M_3\}$. By conditions $(C1)$, $(C2)$ and $(C3)$ and Lemma 5, we obtain that

$$\lim_{n\to\infty} \|x_{n+1} - x_n\| = 0. \tag{26}$$

Define a mapping $S : C \to C$ by

$$Sx = \sum_{m=1}^{r} \beta_m P_C(\hat{\lambda}_m B_m x - \lambda_m A_m x), \quad \forall x \in C, \tag{27}$$

where $\beta_m = \lim_{n\to\infty} \beta_{(m,n)}$. From Lemma 2, we see that S is a nonexpansive mapping and $F(S) = \cap_{m=1}^{r} F(T_m) = \cap_{m=1}^{r} GVI(C, B_m, A_m), \forall n \geq 1$.

Step 4: We will show that $\lim_{n\to\infty} \|Sx_n - x_n\| = 0$.

By Lemma 9, we see that

$$\begin{aligned} \|u_n^i - c\|^2 &\leq \langle T_{r_n}^i x_n - T_{r_n}^i c, x_n - c\rangle \\ &= \|x_n - c\|^2 - \|u_n^i - x_n\|^2. \end{aligned} \tag{28}$$

From (28) and Lemma 10,

$$\|v_n - c\|^2 \leq \frac{1}{l}\sum_{i=1}^{l} \|u_n^i - c\|^2 \leq \|x_n - c\|^2 - \frac{1}{l}\sum_{i=1}^{l} \|u_n^i - x_n\|^2. \tag{29}$$

From (29) and Lemma 4, we have

$$\begin{aligned} \|x_{n+1} - c\|^2 &\leq (1 - \alpha_n\pi)^2\|v_n - c\|^2 + 2\alpha_n\langle \gamma f(x_n) - \mu Gc, x_{n+1} - c\rangle \\ &\leq \|v_n - c\|^2 + 2\alpha_n\|\gamma f(x_n) - \mu Gc\|\|x_{n+1} - c\| \\ &\leq \|x_n - c\|^2 - \frac{1}{l}\sum_{i=1}^{l} \|u_n^i - x_n\|^2 + 2\alpha_n\|\gamma f(x_n) - \mu Gc\|\|x_{n+1} - c\|. \end{aligned}$$

It follows that

$$\begin{aligned} \frac{1}{l}\sum_{i=1}^{l} \|u_n^i - x_n\|^2 \leq \big[\|x_n - c\| + \|x_{n+1} - c\|\big]\|x_{n+1} - x_n\| \\ + 2\alpha_n\|\gamma f(x_n) - \mu Gc\|\|x_{n+1} - c\|. \end{aligned} \tag{30}$$

Letting $n \to \infty$ in the equality (30), we obtain

$$\lim_{n\to\infty} \|u_n^i - x_n\| = 0, \quad \forall i \in I. \tag{31}$$

By Lemma 10, we get

$$\|v_n - x_n\|^2 = \left\| \sum_{i=1}^{l} \frac{1}{l}[u_n^i - x_n] \right\|^2 \leq \frac{1}{l} \sum_{i=1}^{l} \|u_n^i - x_n\|^2. \tag{32}$$

Hence

$$\lim_{n\to\infty} \|v_n - x_n\| = 0. \tag{33}$$

Furthermore, it is easy to prove that

$$\lim_{n\to\infty} \|v_n - u_n^i\| = 0, \quad \forall i \in I. \tag{34}$$

From (14), we observe that

$$\|x_{n+1} - S_n v_n\| = \alpha_n \|\gamma f(x_n) + \mu G S_n v_n\|. \tag{35}$$

Hence, $\lim_{n\to\infty} \|x_{n+1} - S_n v_n\| = 0$. Since $\|x_n - S_n v_n\| \leq \|x_{n+1} - S_n v_n\| + \|x_n - x_{n+1}\|$, it follows that

$$\lim_{n\to\infty} \|x_n - S_n v_n\| = 0. \tag{36}$$

From (31), (33), (34), (36) and S_n is nonexapnsive, we have

$$\lim_{n\to\infty} \|x_n - S_n x_n\| = \lim_{n\to\infty} \|v_n - S_n v_n\| = \lim_{n\to\infty} \|u_n^i - S_n u_n^i\| = 0, \quad \forall i \in I. \tag{37}$$

Now, we show that $Sx_n - x_n \to 0$ as $n \to \infty$. Note that

$$\begin{aligned} \|Sx_n - x_n\| &\leq \|\sum_{m=1}^{r} \beta_m T_m x_n - \sum_{m=1}^{r} \beta_{(m,n)} T_m x_n\| + \|S_n x_n - x_n\| \\ &\leq M_2 (\sum_{m=1}^{r} |\beta_m - \beta_{(m,n)}|) + \|S_n x_n - x_n\|. \end{aligned}$$

By the condition $(C2)$ and (37), we have

$$\lim_{n\to\infty} \|x_n - Sx_n\| = 0. \tag{38}$$

From the boundedness of x_n, there exists a subsequence $\{x_{n_j}\} \subset \{x_n\}$ such that $x_{n_j} \rightharpoonup z$ as $j \to \infty$, by Lemma 1 and (38), we obtain $z = Sz$. So, we have

$$z \in F(S) = \cap_{m=1}^{r} GVI(C, B_m, A_m), \quad \forall n \geq 1. \tag{39}$$

From (31), we also have $u^i_{n_j} \rightharpoonup z$ as $j \to \infty, \forall i \in I$, since $F_i(u^i_{n_j}, y) + \frac{1}{r_{n_j}}\langle y - u^i_{n_j}, u^i_{n_j} - x_{n_j}\rangle \geq 0, \forall y \in C$, it follows from (A2) that

$$\frac{1}{r_{n_j}}\langle y - u^i_{n_j}, u^i_{n_j} - x_{n_j}\rangle \geq F_i(y, u^i_{n_j}) + F_i(u^i_{n_j}, y) + \frac{1}{r_{n_j}}\langle y - u^i_{n_j}, u^i_{n_j} - x_{n_j}\rangle$$

$$\langle y - u^i_{n_j}, \frac{u^i_{n_j} - x_{n_j}}{r_{n_j}}\rangle \geq F_i(y, u^i_{n_j}), \quad \forall y \in C. \tag{40}$$

From (40) and (A4), we have

$$F_i(y, z) \leq 0, \quad \forall y \in C. \tag{41}$$

Put $y_t = ty + (1-t)z, t \in (0,1)$. Then $y_t \in C$ and $F_i(y_t, z) \leq 0$ for all $i \in I$. By (A1) and (A4), we obtain $0 = F_i(y_t, y_t) \leq tF_i(y_t, y) + (1-t)F_i(y_t, z) \leq tF_i(y_t, y), \forall i \in I$. By (A3), we get

$$F_i(z, y) \geq \lim_{t \downarrow 0} F_i(ty + (1-t)z, y) = \lim_{t \downarrow 0} F_i(y_t, y) \geq 0, \forall i \in I.$$

It follows that $z \in \bigcap_{i=i}^{l} EP(F_i)$. Hence, $z \in \Omega$. So, we have $\omega_w(x_n) \subset \Omega$. By Lemma 6, $\mu G - \gamma f$ is strongly monotone, so the variational inequality (11) has a unique solution $c \in \Omega$.

Step 5: We will show that $\limsup_{n\to\infty}\langle(\gamma f - \mu G)c, x_n - c\rangle \leq 0$.

Indeed, since $\{x_n\}$ is bounded, then there exists a subsequence $\{x_{n_i}\} \subset \{x_n\}$ such that

$$\limsup_{n\to\infty}\langle(\gamma f - \mu G)c, x_n - c\rangle = \lim_{i\to\infty}\langle(\gamma f - \mu G)c, x_{n_i} - c\rangle. \tag{42}$$

Without loss of generality, we may further assume that $x_{n_i} \rightharpoonup z$. It follows from (42) that $z \in \Omega$. Since z is the unique solution of (11), we obtain

$$\begin{aligned}\limsup_{n\to\infty}\langle(\gamma f - \mu G)c, x_n - c\rangle &= \lim_{i\to\infty}\langle(\gamma f - \mu G)c, x_{n_i} - c\rangle \\ &= \langle(\gamma f - \mu G)c, z - c\rangle \leq 0.\end{aligned} \tag{43}$$

Step 6: Finally, we will show that $x_n \to c$ as $n \to \infty$.

From Lemma 3, Lemma 4 and (10) we have

$$\begin{aligned}\|x_{n+1} - c\|^2 &\leq (1 - \alpha_n\pi)^2\|\frac{1}{l}\sum_{i=1}^{l}(u^i_n - c)\|^2 + 2\alpha_n\gamma\langle f(x_n) - f(c), x_{n+1} - c\rangle \\ &\quad + 2\alpha_n\langle\gamma f(c) - \mu Gc, x_{n+1} - c\rangle \\ &\leq \left[1 - \frac{2\alpha_n(\pi - \gamma k)}{1 - \alpha_n\gamma k}\right]\|x_n - c\|^2 + \frac{(\alpha_n\pi)^2}{1 - \alpha_n\gamma k}\|x_n - c\|^2\end{aligned}$$

$$+\frac{2\alpha_n}{1-\alpha_n\gamma k}\langle\gamma f(c)-\mu Gc, x_{n+1}-c\rangle$$
$$=(1-\theta_n)\|x_n-c\|^2+\delta_n,$$

where $\theta_n := \frac{2\alpha_n(\pi-\gamma k)}{1-\alpha_n\gamma k}$ and $\delta_n := \frac{\alpha_n}{1-\alpha_n\gamma k}[\alpha_n\pi^2\|x_n-c\|^2+2\langle\gamma f(c)-\mu Gc, x_{n+1}-c\rangle]$. Note that,

$$\theta_n := \frac{2\alpha_n(\pi-\gamma k)}{1-\alpha_n\gamma k} \le \frac{2(\pi-\gamma k)}{1-\gamma k}\alpha_n. \tag{44}$$

By (C1), we obtain that $\lim_{n\to\infty}\theta_n = 0$. On the other hand, we have

$$\theta_n := \frac{2\alpha_n(\pi-\gamma k)}{1-\alpha_n\gamma k} \ge 2\alpha_n(\pi-\gamma k). \tag{45}$$

From (C1), we have $\sum_{n=1}^{\infty}\theta_n = \infty$. Put $M = \sup_{n\in N}\{\|x_n-c\|\}$, we have

$$\frac{\delta_n}{\theta_n} = \frac{1}{2(\pi-\gamma k)}[\alpha_n\pi^2 M + 2\langle\gamma f(c)-\mu Gc, x_{n+1}-c\rangle]. \tag{46}$$

It follows that $\limsup_{n\to\infty}\frac{\delta_n}{\theta_n} \le 0$. Hence, by Lemma 5, we conclude that

$$\lim_{n\to\infty}\|x_n-c\| = 0. \tag{47}$$

Therefore $x_n \to c$ as $n\to\infty$. This completes the proof. □

As direct consequences of Theorem 1, we obtain corollaries.

Corollary 1 *Let C be a nonempty closed and convex subset of a real Hilbert space H such that $C \pm C \subset C$. Let F be a bi-function from $C\times C$ into $\Re$ satisfying (A1)–(A4). Let $f : C\to C$ be a contraction with coefficient $k\in(0,1)$. Let $G : C\to C$ be a ξ-strongly monotone and L-Lipschitz continuous mapping. Let $A_m : C\to H$ be a relaxed (η_m,ρ_m)-cocoercive and v_m-Lipschitz continuous mapping and $B_m : C\to H$ be a relaxed $(\hat{\eta}_m,\hat{\rho}_m)$-cocoercive and $\hat{v}_m$-Lipschitz continuous mapping for each $1\le m\le r$. Let $p_m = \sqrt{1-2\lambda_m\rho_m+\lambda_m^2v_m^2+2\lambda_m\eta_mv_m^2}$ and $q_m = \sqrt{1-2\hat{\lambda}_m\hat{\rho}_m+\hat{\lambda}_m^2\hat{v}_m^2+2\hat{\lambda}_m\hat{\eta}_m\hat{v}_m^2}$, where $\{\lambda_m\}$ and $\{\hat{\lambda}_m\}$ are two positive sequences for each $1\le m\le r$. Assume that $\Delta = EP(F)\cap(\cap_{m=1}^{r}GVI(C,B_m,A_m)) \ne \emptyset$, $\xi>0$, $L>0$, $0<\mu<2\xi/L^2$, $0<\gamma<\mu(\xi-\mu L^2/2)/k = \pi/k$ and $p_m, q_m\in[0,\frac{1}{2})$, for each $1\le m\le r$. Given $\{x_n\}$ is a sequence generated by*

$$\begin{cases} x_1 \in C, \\ u_n = T_{r_n} x_n, \\ y_n^m = P_C(\hat{\lambda}_m B_m u_n - \lambda_m A_m u_n), \quad \forall m = 1, 2, \ldots, r, \\ x_{n+1} = \alpha_n \gamma f(x_n) + (I - \alpha_n \mu G) \sum_{m=1}^{r} \beta_{(m,n)} y_n^m, \quad \forall n \geq 1, \end{cases} \tag{48}$$

where $\{\alpha_n\}, \{\beta_{(m,n)}\} \subset (0,1), \forall 1 \leq m \leq r$ *and* $\{r_n\} \subset (0, +\infty)$ *satisfying the following conditions:*

(C1) $\lim_{n\to\infty} \alpha_n = 0, \sum_{n=1}^{\infty} \alpha_n = \infty$ *and* $\sum_{n=1}^{\infty} |\alpha_{n+1} - \alpha_n| < \infty$;
(C2) $\sum_{m=1}^{r} \beta_{(m,n)} = 1, \forall n \geq 1, \sum_{n=1}^{\infty} |\beta_{(m,n+1)} - \beta_{(m,n)}| < \infty$ *and* $\lim_{n\to\infty} \beta_{(m,n)} = \beta_m \in (0,1), \forall 1 \leq m \leq r$.
(C3) $\liminf_{n\to\infty} r_n > 0$ *and* $\sum_{n=1}^{\infty} |r_{n+1} - r_n| < \infty$;

Then the sequence $\{x_n\}$ *converges strongly to a common element* $c \in \Delta$, *which is the unique solution of the HEGVIP:* $\langle (\gamma f - \mu G)c, x - c\rangle \leq 0, \quad \forall x \in \Delta$.

Corollary 2 *Let* C *be a nonempty closed and convex subset of a real Hilbert space* H *such that* $C \pm C \subset C$. *Let* F *be a bi-function from* $C \times C$ *into* $\Re$ *satisfying (A1)–(A4). Let* $f : C \to C$ *be a contraction with coefficient* $k \in (0,1)$. *Let* $G : C \to C$ *be a* ξ*-strongly monotone and* L*-Lipschitz continuous mapping. Let* $A : C \to H$ *be a relaxed* (η, ρ)*-cocoercive and* ν*-Lipschitz continuous mapping. Let* $B : C \to H$ *be a relaxed* $(\hat{\eta}, \hat{\rho})$*-cocoercive and* $\hat{\nu}$*-Lipschitz continuous mapping. Let* $p = \sqrt{1 - 2\lambda\rho + \lambda^2\nu^2 + 2\lambda\eta\nu^2}$ *and* $q = \sqrt{1 - 2\hat{\lambda}\hat{\rho} + \hat{\lambda}^2\hat{\nu}^2 + 2\hat{\lambda}\hat{\eta}\hat{\nu}^2}$, *where* λ *and* $\hat{\lambda}$ *are two positive real numbers. Assume that* $\Lambda = EP(F) \cap GVI(C, B, A) \neq \emptyset$, $\xi > 0, L > 0, 0 < \mu < 2\xi/L^2, 0 < \gamma < \mu(\xi - \mu L^2/2)/k = \pi/k$ *and* $p, q \in [0, \frac{1}{2})$. *Given the initial guess* $x_1 \in C$ *and* $\{x_n\}$ *is a sequence generated by*

$$\begin{cases} x_1 \in C, \\ u_n = T_{r_n} x_n, \\ y_n = P_C(\hat{\lambda} B u_n - \lambda A u_n), \\ x_{n+1} = \alpha_n \gamma f(x_n) + (I - \alpha_n \mu G) y_n, \quad \forall n \geq 1, \end{cases} \tag{49}$$

where $\{\alpha_n\} \subset (0,1)$ *and* $\{r_n\} \subset (0, +\infty)$, *satisfying the following conditions:*

(C1) $\lim_{n\to\infty} \alpha_n = 0, \sum_{n=1}^{\infty} \alpha_n = \infty$ *and* $\sum_{n=1}^{\infty} |\alpha_{n+1} - \alpha_n| < \infty$;
(C2) $\liminf_{n\to\infty} r_n > 0$ *and* $\sum_{n=1}^{\infty} |r_{n+1} - r_n| < \infty$;

Then the sequence $\{x_n\}$ *converges strongly to a common element* $c \in \Lambda$, *which is the unique solution of the HEGVIP:* $\langle (\gamma f - \mu G)c, x - c\rangle \leq 0, \quad \forall x \in \Lambda$.

If $F_i(x, y) \equiv 0, \forall (x, y) \in C \times C$ in Theorem 1, for all $i \in I$. Then, from the algorithm (10), we have $u_n^i \equiv P_C x_n$, for all $i \in I$. So we have the following result.

Corollary 3 *Let* C *be a nonempty closed and convex subset of a real Hilbert space* H *such that* $C \pm C \subset C$. *Let* $f : C \to C$ *be a contraction with coefficient* $k \in (0,1)$. *Let* $G : C \to C$ *be a* ξ*-strongly monotone and* L*-Lipschitz continuous mapping. let* A_m :

$C \to H$ be a relaxed (η_m, ρ_m)-cocoercive and ν_m-Lipschitz continuous mapping and $B_m : C \to H$ be a relaxed $(\hat{\eta}_m, \hat{\rho}_m)$-cocoercive and $\hat{\nu}_m$-Lipschitz continuous mapping for each $1 \leq m \leq r$. Let $p_m = \sqrt{1 - 2\lambda_m\rho_m + \lambda_m^2\nu_m^2 + 2\lambda_m\eta_m\nu_m^2}$ and $q_m = \sqrt{1 - 2\hat{\lambda}_m\hat{\rho}_m + \hat{\lambda}_m^2\hat{\nu}_m^2 + 2\hat{\lambda}_m\hat{\eta}_m\hat{\nu}_m^2}$, where $\{\lambda_m\}$ and $\{\hat{\lambda}_m\}$ are two positive sequences for each $1 \leq m \leq r$. Assume that $\Theta = \cap_{m=1}^{r} GVI(C, B_m, A_m) \neq \emptyset$, $\xi > 0, L > 0, 0 < \mu < 2\xi/L^2, 0 < \gamma < \mu(\xi - \mu L^2/2)/k = \pi/k$ and $p_m, q_m \in [0, \frac{1}{2})$, for each $1 \leq m \leq r$. Given the initial guess $x_1 \in C$ and $\{x_n\}$ is a sequence generated by

$$x_{n+1} = \alpha_n\gamma f(x_n) + (I - \alpha_n\mu G)\Sigma_{m=1}^{r}\beta_{(m,n)} P_C(\hat{\lambda}_m B_m x_n - \lambda_m A_m x_n), \quad \forall n \geq 1, \tag{50}$$

where $\{\alpha_n\}$ and $\{\beta_{(m,n)}\} \subset (0, 1), \forall 1 \leq m \leq r$ satisfying the following conditions:

(C1) $\lim_{n\to\infty} \alpha_n = 0, \sum_{n=1}^{\infty} \alpha_n = \infty$ *and* $\sum_{n=1}^{\infty} |\alpha_{n+1} - \alpha_n| < \infty$;
(C2) $\sum_{m=1}^{r} \beta_{(m,n)} = 1, \forall n \geq 1, \sum_{n=1}^{\infty} |\beta_{(m,n+1)} - \beta_{(m,n)}| < \infty$ *and* $\lim_{n\to\infty} \beta_{(m,n)} = \beta_m \in (0, 1), \forall 1 \leq m \leq r$.

Then the sequence $\{x_n\}$ converges strongly to a common element $c \in \Theta$, which is the unique solution of the HGVIP: $\langle(\gamma f - \mu G)c, x - c\rangle \leq 0, \quad \forall x \in \Theta$.

Acknowledgments The authors were supported by the Higher Education Research Promotion and National Research University Project of Thailand, Office of the Higher Education Commission (Grant no. NRU56000508).

References

1. Slifka MK, Whitton JL (2000) Clinical implications of dysregulated cytokine production. J Mol Med. doi:10.1007/s001090000086
2. Stampacchia G (1964) Formes bilineaires coercitives sur les ensembles convexes. Comptes rendus de I'Academie des Sci 258:4413–4416
3. Lions J-L, Stampacchia G (1967) Variational inequalities. Commun Pure Appl Math 20:493–519
4. Liu F, Nashed MZ (1998) Regularization of nonlinear ill-posed variational inequalities and convergence rates. Set-Valued Anal 6:313–344
5. Korpelevic GM (1976) An extragradient method for finding saddle points and for other problems. E'konomika i Matematicheskie Metody 12:747–756
6. Iusem AN, Svaiter BF (1997) A variant of Korpelevich's method for variational inequalities with a new search strategy. Optimization 42:309–321
7. Khobotov EN (1989) Modification of the extra-gradient method for solving variational inequalities and certain optimization problems. USSR Comput Math Math Phys 27:120–127
8. Solodov MV, Svaiter BF (1999) A new projectionmethod for variational inequality problems. SIAM J Control Optim 37:765–776
9. Noor Aslam M (2004) Some developments in general variational inequalities. Appl Math Comput 152:199–277
10. Yao Y, Liou Y C, Yao J C (2008) A new hybrid iterative algorithm for fixed-point problems, variational inequality problems and mixed equilibrium problems. Fixed point theory and applications

11. Wang S, Marino G, Wang F (2010) Strong convergence theorems for a generalized equilibrium problem with a relaxed monotone mapping and a countable family of nonexpansive mappings in a Hilbert space. Fixed point theory and applications
12. Cianciaruso F, Marino G, Muglia L, Yao Y (2010) A hybrid projection algorithm for finding solutions of mixed equilibrium problem and variational inequality problem. Fixed point theory and applications
13. Peng JW, Wu SY, Yao JC (2010) A new iterativemethod for finding common solutions of a system of equilibrium problems fixed-point problems and variational inequalities. Abstract and applied analysis
14. Combettes PL (2003) A block-iterative surrogate constraint splitting method for quadratic signal recovery. IEEE Trans Signal Process 51:1771–1782
15. Iiduka H (2012) Fixed point optimization algorithmand its application to power control in CDMA data networks. Math Program 133:227–242
16. Slavakis K, Yamada I (2007) Robust wideband beamforming by the hybrid steepest descent method. IEEE Trans Signal Process 55:4511–4522
17. Chang S-S, Joseph Lee HW, Chan CK (2009) A new method for solving equilibrium problem fixed point problem and variational inequality problem with application to optimization. Nonlinear Anal Theor Methods Appl 70:3307–3319
18. Colao V, Marino G, Xu H-K (2008) An iterative method for finding common solutions of equilibrium and fixed point problems. J Math Anal Appl 344:340–352
19. Combettes PL, Hirstoaga SA (1997) Equilibrium programming using proximal-like algorithms. Math Program 78:29–41
20. Plubtieng S, Punpaeng R (2008) A new iterative method for equilibrium problems and fixed point problems of nonexpansive mappings and monotone mappings. Appl Math Comput 197:548–558
21. Takahashi S, Takahashi W (2008) Strong convergence theorem for a generalized equilibrium problem and a nonexpansive mapping in a Hilbert space. Nonlinear Anal Theor Methods Appl 69:1025–1033
22. Combettes PL (1996) The convex feasibility problem: in image recovery. In: Hawkes P (ed) Advances in imaging and electron physics, vol 95. Academic Press, Orlando, pp 155–270 (1996).
23. Kotzer T, Cohen N, Shamir J (1995) Images to ration by a novel method of parallel projection onto constraint sets. Opt Lett 20:1172–1174
24. Sezan MI, Stark H (1987) Application of convex projection theory to image recovery in tomograph and related areas. In: Stark H (ed) Image recovery: theory and application. Academic Press, Orlando, pp 155–270
25. Censor Y, Zenios SA (1997) Parallel Optimization. Theory Algorithms and Applications Numerical Mathematics and Scientific Computation. Oxford University Press, New York
26. He Z, Du W-S (2011) Strong convergence theorems for equilibrium problems and fixed point problems: a new iterative method, some comments and applications. Fixed Point Theory Appl. doi:10.1186/1687-1812-2011-33
27. Wairojjana N, Kumam P (2013) General iterative method for convex feasibility problem via the hierarchical generalized variational inequality problems. Lecture Notes in Engineering and Computer Science: Proceeding of the International MultiConference of Engineers and Computer Scientists 2013. 2:1129–1134
28. Browder FE (1976) Nonlinear operators and nonlinear equations of evolution in Banach spaces. Proc Symp Pure Math 18:78–81
29. Bruck RE (1973) Properties of fixed point sets of nonexpansive mappings in Banach spaces. Trans Am Math Soc 179:251–262
30. Tian M (2010) A general iterative algorithm for nonexpansive mappings in Hilbert spaces. Nonlinear Anal 73:689–694
31. Xu HK (2002) Iterative algorithms for nonlinear operators. J London Math Soc 66:240–256
32. Acedoa GL, Xu HK (2007) Iterative methods for strict pseudo-contractions in Hilbert spaces. Nonlinear Anal 66:2258–2271

33. Blum E, Oettli W (1994) From optimization and variational inequalities to equilibrium problems. Math Student 63:123–145
34. Combettes PL, Hirstoaga SA (2005) Equilibrium programming in Hilbert spaces. J Nonlinear Convex Anal 6:29–41

Workforce Scheduling Using the PEAST Algorithm

Nico R. M. Kyngäs, Kimmo J. Nurmi and Jari R. Kyngäs

Abstract Workforce scheduling has become increasingly important for both the public sector and private companies. Good rosters have many benefits for an organization, such as lower costs, more effective utilization of resources and fairer workloads and distribution of shifts. This paper presents a framework and an algorithm that have been successfully used to model and solve workforce scheduling problems in Finnish companies. The algorithm has been integrated into market-leading workforce management software in Finland.

Keywords Computational intelligence · Local search · Metaheuristics · PEAST algorithm · Population-based methods · Staff scheduling · Workforce scheduling

1 Introduction

Workforce scheduling, also called staff scheduling and labor scheduling, is a difficult and time consuming problem that every company or institution that has employees working on shifts or on irregular working days must solve. The workforce scheduling problem has a fairly broad definition. Most of the studies focus on assigning employees to shifts, determining working days and rest days or constructing flexible shifts and their starting times. Different variations of the problem and subproblems are NP-hard and NP-complete [1–5], and thus extremely hard to solve. The first mathematical formulation of the problem based on a generalized set covering model was

N. R. M. Kyngäs (✉) · K. J. Nurmi · J. R. Kyngäs
Satakunta University of Applied Sciences, Tiedepuisto 3, 28600 Pori, Finland
e-mail: nico.kyngas@samk.fi

K. J. Nurmi
e-mail: cimmo.nurmi@samk.fi

J. R. Kyngäs
e-mail: jari.kyngas@samk.fi

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 275, DOI: 10.1007/978-94-007-7684-5_25,

proposed by Dantzig [6]. Good overviews of workforce scheduling are published by Alfares [7], Ernst et al. [8] and Meisels and Schaerf [9].

Section 2 briefly introduces the necessary terminology and the workforce scheduling process as we have encountered it in various real-world cases. In Sect. 3 we describe the preprocessing phase of the workforce scheduling process. Section 4 presents the staff scheduling phase. Along with the problems and definitions of the subphases themselves we introduce some new real-world cases. Section 5 gives an outline of our computational intelligence algorithm.

We have used the PEAST algorithm, as described in Sect. 5, to solve numerous real-world staff scheduling problems for different Finnish companies. The algorithm has been integrated into the workforce management system of our business partner, and it is in constant real-world use.

2 Terminology and the Workforce Scheduling Process in Brief

The *planning horizon* is the time interval over which the employees have to be scheduled. Each employee has a total working time that he/she has to work during the planning horizon. Furthermore, each employee has *competences* (qualifications and skills) that enable him/her to carry out certain tasks. Days are divided into *working days* (days-on) and *rest days* (days-off). Each day is divided into periods or timeslots. A *timeslot* is the smallest unit of time and the length of a timeslot determines the granularity of the schedule. A *shift* is a contiguous set of working hours and is defined by a day and a starting period on that day along with a *shift length* (the number of occupied timeslots) or *shift time*. Shifts are sometimes grouped into *shift types*, such as morning, day and night shifts. Each shift is composed of *tasks* and *breaks*. The sum of the length of a shift's tasks is called *working time*, whereas sometimes the sum of the length of a shift's breaks is called *linkage time*. A timeslot-long piece of a task or break is called an *activity*. A consecutive sequence of activities dedicated to a single task is called a *stretch*. A shift or a task may require the employee assigned to it to possess one or more *competences*. A work schedule over the planning horizon for an employee is called a *roster*. A roster is a combination of shifts and days-off assignments that covers a fixed period of time.

Workload prediction, also referred to as demand forecasting or demand modeling, is the process of determining the staffing levels—that is, how many employees are needed for each timeslot in the planning horizon. The staffing is preceded by actual workload prediction or *workload determination* based on static workload constraints given by the company, depending on the situation. In *preference scheduling*, each employee gives a list of preferences and attempts are made to fulfill them as well as possible. The employees' preferences are often considered in the days-off scheduling and staff rostering subphases, but may also be considered during shift generation. Together these two subphases form the *preprocessing phase*.

Shift generation is the process of determining the shift structure, along with the activities to be carried out in particular shifts and the competences required for

different shifts. *Days-off scheduling* deals with the assignment of rest days between working days over a given planning horizon. Days-off scheduling also includes the assignment of vacations and special days, such as union steward duties and training sessions. *Staff rostering*, also referred to as shift scheduling, deals with the assignment of employees to shifts. It can also specify the starting time and duration of shifts for a given day, even though in most cases they are pre-assigned during shift generation. This subphase may include both *resource analysis* to examine the compatibility between the available workforce and the shifts, and *partitioning* in the case of massive datasets (i.e. hundreds of employees) to speed up and improve on the results of the rostering. Together these five subphases form the *staff scheduling phase*.

Rescheduling deals with ad hoc changes that are necessary due to sick leaves or other no-shows. The changes are usually carried out manually. Finally, participation in *evaluation* ranges from the individual employee through personnel managers to executives. A reporting tool should provide performance measures in such a way that the personnel managers can easily evaluate both the realized staffing levels and the employee satisfaction. When necessary, parts of the whole workforce scheduling process may be restarted. *Workforce scheduling* consists roughly of everything from determining the needs of the customers to determining the exact schedule of each employee.

The workforce scheduling process presented in this paper is mostly concerned with short-term planning, as defined in [10]. We have chosen to split the problem into subphases as seen in Fig. 1. This may cause problems in extremely difficult cases, due to the search space at each subphase being constrained by the choices

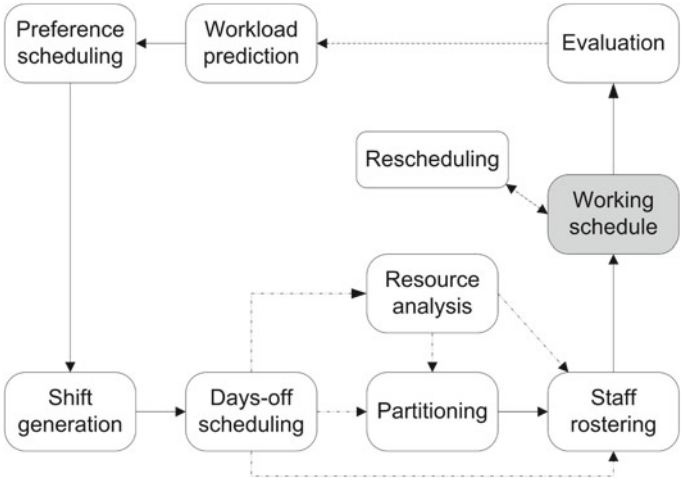


Fig. 1 The workforce scheduling process. The *upper boxes* represent the subphases that may occur in both mid-term and short-term planning (preprocessing phase). The subphases represented by the *lower boxes* (staff scheduling phase) only occur in short term planning, although information gathered from these may prove useful in future mid-term planning

made during previous subphases. This would be an untenable approach for finding the global optimum for most problems. However, our goal is to find a good enough solution for a broad range of problems. Our subphase-based approach is flexible enough to achieve this goal. Another benefit is decreased computational complexity due to the constantly narrowing search space.

The staff scheduling phase can be solved using computational intelligence. Computational workforce scheduling is key to increased productivity, quality of service, customer satisfaction and employee satisfaction. Other advantages include reduced planning time, reduced payroll expenses and ensured regulatory compliance.

3 The Preprocessing Phase

The preprocessing phase is the foundation upon which the actual staff scheduling phase is built. It may involve identifying both the needs of the customer(s) and the attributes (preferences, skills etc.) of the employees, and determining staffing requirements based on the former. This phase can be thought of as the transition between mid-term and short-term planning, since it touches both. This is the point in the workforce scheduling process where historical data and the schedules of previous planning horizons are most useful.

3.1 Workload Prediction/Determination

The nature of determining the amount and type of work to be done at any given time during the next planning horizon depends greatly on the nature of the job. If the workload is uncertain then some form of workload prediction is called for [8]. Some examples of this are the calls incoming to a call center or the customer influx to a hospital.

We define the *service level* SL(n) as the percentage of customers that need to wait for service for at most n seconds. Usually workload prediction aims to provide a certain service level (or above) for some fixed n. We simulate the randomly distributed workload based on historical data and statistical analysis, and find a suitable working employee structure (i.e. how many and what kinds of employees are needed) over time [11]. Computationally this approach is much more intensive than methods based on queuing theory. However, it has the benefit of being applicable to almost any real-world situation.

If the workload is static, no forecasting is necessary. For example, a local transport company might be under a strict contract to drive completely pre-assigned bus lines. In such a case shift generation may be necessary to combine the different bus lines into shifts, but the workload as such is static and thus calls for no forecasting.

3.2 Preference Scheduling

Research by Kellogg and Walczak [12] indicates that it is crucial for a workforce management system to allow the employees to affect their own schedules. In general it improves employee satisfaction. This in turn reduces sick leaves and improves the efficiency of the employees, which means more profit for the employer. Hence we use an easy-to-use user interface that allows the employees to input their preferences into the workforce management system. This eases the organizational workload of the personnel manager. A measure of fairness is incorporated via limiting the number and type of different wishes that can be expressed per employee. We are looking into incorporating more rigorous fairness measures and more complex preferences to the system. In our experience our current system is satisfactory, yet there is always room for improvement.

Preferences can be considered at every subphase of the staff scheduling phase [13–15]. The different types of preferences we consider are found under the respective subphases of the workforce scheduling process.

4 The Staff Scheduling Phase

4.1 Shift Generation

Shift generation transforms the determined workload into shifts. This includes deciding break times when applicable. Shift generation is essential especially in cases where the workload is not static. In other cases companies often want to hold on to their own established shift building methods.

A basic shift generation problem includes a variable number of activities for each task in each timeslot. Some tasks are not time-dependent; instead, there may be a daily quota to be fulfilled. Activities may require competences. The most important optimization target is to match the shifts to the workload as accurately as possible. In our solutions we create the shifts for each day separately, each shift corresponding to a single employee's competences and preferences. We do not minimize the number of different shifts. The choice between hard and soft constraints is given by the instances themselves.

We now present a real-world case from a Finnish haulage company. The problem as it was presented to us, related to a cargo terminal of theirs, is as follows. The planning horizon is five days, extending from a Sunday evening to a Friday evening. Each hour a number $M(d, h)$, where d = day and h = hour, of arrival manifests needs to be processed. The arrivals not handled immediately are queued, and the queue needs to be empty in the morning (6.30) and in the evening (20.30). The values of M for different days and hours can be found in [16].

It is assumed that the employees are identical in their processing capacity: each employee can handle 11 manifests per hour during the day (6–22) and 17.5 manifests

per hour during the night (22–6). There are 76 full-time employees and 13 part-time employees. A full-time employee's shifts must be 4–10 h long, and the total working time over a 6-week period must be 240 h. A part-time employee's shifts must be 4–6 h long. Additionally, a part-time employee should have 3 shifts per week, and the total number of part-timers' working hours must be at most 10 % of the total working hours of all the employees.

The length of a timeslot is 30 min. The number of full-time employees was restricted to 69 in order to keep the workload of the part-time employees suitable. Thus we end up with 82 employees in total. The following hard constraints were used. The notation for different constraints is from [16].

(SGS1): No shift should contain timeslots with multiple types of activities.
(SGS2): No shift should contain gaps.
(SGP3): Each shift must contain a 30-min lunch break.
(SGV5): Each shift's length must be within the allowed limits of the corresponding employee.

The following soft constraints were used.

(SGP4): For shift s, let d_s be the distance of its break from the midpoint of the shift in slots (it does not matter if the optimal positions are not integers), and let a_s be 0.25 $\times$ (length of s). If $d_s > a_s$, then a cost of round$(d_s - a_s)$ is incurred.

(SGC4) + (SGC5): A compound constraint is used in order to make sure that the queue of arrivals is empty at 6.30 and 20.30, and that no working time is lost due to having too much workforce at work too early. The *cumulative effective workload* CEW[*day, type, time*] represents the workload that has effectively been contributed to handling the manifests up to timeslot *time*. It is calculated as

$$\mathrm{CEW}\,[day,\ type,\ time] = \min \left\{ \begin{array}{l} \mathrm{MCS}\,[day,\ type,\ time]\,, \\ \mathrm{CEW}\,[day,\ type, time-1] \\ +\mathrm{w}\,[day,\ type,\ time] \end{array} \right\}, \qquad (1)$$

where w is the number of workers scheduled to do a certain task at a certain time and MCS is the *maximum cumulative workload* given in [16]. For each day, the penalty given is the sum of differences between total workload and effective workload for day and night tasks.

The results are briefly described in Table 1 along with comparative numbers from the company's own solution. Our solution has no violations in SGP4, so the total penalty (649) represents exactly the number of timeslots that the effective working time is short of the total time that the jobs require ($649 \times 30 = 178,170 - 158,700$). The numbers from the company's current scheduling method made us doubt whether all the assumptions were close enough to reality and if the data/model were precise enough, but based on our results a contract for the use of our optimization software was signed.

Table 1 Comparison between our solution and a manual solution

	Our solution	Manual solution
Total working minutes (actual)	175,020	201,300
Effective working minutes (actual)	158,700	144,330
Total job minutes (goal)	178,170	178,170
Job completion (%)	89	81
Percentage of wasted working time	9	28

We hope to get more precise data in the future in order to improve both our model and our solutions. In Sect. 4.3.3 we'll roster the staff using the generated shifts.

4.2 Days-Off Scheduling

Days-off scheduling decides the rest days and the working days of the employees. It is based on the result of the shift generation: for each day a set of suitable employees must be available to carry out the shifts. This is the first subphase where employees' preferences usually have a big emphasis. The choice between hard and soft constraints is highly dependent on the problem instance.

We have used the list of constraints given in [16] to successfully model and solve some real-world days-off scheduling problems [13] and some nurse rostering problems [17].

4.3 Staff Rostering

4.3.1 Resource Analysis (Optional)

To see if there will be any chance of succeeding at matching the workforce with the shifts while adhering to the given constraints, an analysis is run on the data. If we have already optimized the days-off, this subphase is not necessary but it may still be useful. In addition to helping the personnel manager see the problem with the data quickly and efficiently, it may help convince the management level that the current practices and processes of generating the schedules are simply untenable. We have developed a statistical tool for this.

4.3.2 Partitioning of Massive Data (Optional)

Some real-world datasets are huge. They may consist of hundreds of employees with a corresponding number of jobs. Such datasets are often computationally very

challenging. If there are no apparent "trivial" partitioning criteria (for example, a bus driver could be limited to driving buses starting from a specific bus depot, but if the constraint is not hard or there are drivers without a designated depot, it is not a trivial partitioning criterion) we can use the PEAST algorithm to partition the data, as in [14].

We now present a real-world case from a Finnish bus company. The problem consists of rostering 175 bus drivers over a planning horizon of 2 weeks. The days-off are invariant. There are 6 different kinds of days-off. The hard constraints of the problem are as follows. We have used the list of constraints given in [16] as the basis for modeling the case.

(SRR1): The working time of an employee must be strictly less than his/her goal working time. The shift time of an employee must be greater than his/her goal working time.
(SRR3): The rest time of 9 h must be respected between adjacent shifts.
(SRR5): There is 1 person with 6 working days during which he/she cannot work certain shifts.
(SRO3): The 3 most common kinds of days-off (90 % of all days-offs) must be whole, i.e. they cannot be immediately preceded by a shift that ends after midnight
(SRO4): There are in total 140 pre-assigned shifts.

The soft constraints of the problem are as follows.

(SRR1): The required number of working hours must be respected. The total working hours of the employees range from 1,200 to 4,815 min. The working minutes per day per employee range from 360 to 535. The difference (17,888 min) between employees' total working time goal (786,750 min) and the sum of the working time of all the shifts (768,862 min) should be evenly distributed among the shifts. There are 1,643 shifts, which results in approximately 10.9 min per shift. Each employee should thus be $SG(e) = 10.9\times$ (number of working days of employee e) minutes short of their personal workload goal. Define $S(e)$ as the actual shortage for employee e. If $|S(e) - SG(e)| > 0.1 \times SG(e)$, then the cost given is $|S(e) - SG(e)| - 0.1 \times SG(e)$. This ensures fairness in regard to the working time. The arbitrary threshold is used, since the goal is to have highly similar but not necessarily equal shortages.

Additionally, the linkage time (i.e. time spent having lunch or waiting for another vehicle, totalling 24,322 min in this instance) should be distributed evenly among the employees. This means approximately 14.8 min of linkage time per shift. Thus each employee should have $LG(e) = 14.8\times$ (number of working days of employee e) linkage minutes. Define $L(e)$ as the actual linkage time for employee e. If $|L(e) - LG(e)| > 0.1 \times LG(e)$, then the cost given is $|L(e) - LG(e)| - 0.1 \times LG(e)$. The linkage time is not nearly evenly distributed among different shift types. Almost 90 % of all linkage time belongs to the 60 % of shifts that start before 9 o'clock in the morning, which means that the early shifts have on average 6 times as much linkage

Table 2 Average and quartiles of 6 runs for total hard, total soft and preference constraint violations

	No partitions			Partitions		
	Hard	Soft	Pref	Hard	Soft	Pref
Average	11.7	383.8	284.5	2.5	499.0	224.7
Min	9.0	329.0	264.0	1.0	266.0	214.0
Q_1	10.3	343.8	268.5	2.0	373.5	223.5
Q_2	12.0	372.5	283.0	2.5	408.5	225.0
Q_3	13.0	423.8	297.5	3.0	586.0	228.0
Max	14.0	454.0	311.0	4.0	904.0	232.0

time as the later shifts. Since some employees only want morning shifts while others only want later shifts, compromises have to be made.

(SRR3): Each employee should have at least 11 h of rest time between two adjacent shifts. Each violation of this rule incurs a cost of 1.

(SRP2): There are 1,102 working days with a shift type preference defined. Each unfulfilled wish incurs a cost of 1.

Our results both using and not using partitioning are briefly described in Table 2. One hard constraint violation is unavoidable: there is an employee whose previous planning horizon ended with a late job, yet he has an early pre-assigned job on the first Monday of the new planning horizon, causing a rest time violation. This is a very challenging dataset and as such it shows that partitioning has its benefits. However, in order to eliminate the remaining hard constraint violations with consistency we need to either consider alternative methods or, as the preferred alternative, point out to the problem owner the inaccuracies in their current system and investigate what could be done to rectify the problems caused by their contradictory constraints.

4.3.3 Staff Rostering (Shift Scheduling)

The final optimization subphase of the workforce scheduling process is staff rostering, during which the shifts are assigned to the employees. The length of the planning horizon for this subphase is usually between two and six weeks. The preferences of the employees are usually given a relatively large weight but, as before, the choice between hard and soft constraints stems from the instances themselves. The most important constraints are usually resting times and certain competences, since these are often laid down by the collective labour agreements and government regulations. Working hours of the employees are also important. We have used the list of constraints given in [16] to successfully model and solve some real-world staff rostering cases [14, 18] along with some nurse rostering cases [17].

In Sect. 4.1 we generated the shifts for a haulage company. Next we will schedule those shifts in order to optimize working time and resting time for each employee. In

this case no separate days-off scheduling is necessary, since there are no constraints involving days-off directly.

We generated shifts for 69 full-time employees and 13 part-time employees. A full-time employee's shifts must be 4–10 h long, and the total working time over a 6-week period must be 240 h. However, our planning horizon is only 5 days (one week), so each full-timer should have approximately 40 h of work. A part-time employee's shifts must be 4–6 h long, and a part-time employee should have 3 shifts per week. The following hard constraints were used. The notation for different constraints is from [16].

(SRR3): Each employee must have at least 7 h of rest time between two adjacent shifts.
(SRO1): Part-timers only have competence to work shifts that are less than 6 h in length.

The following soft constraints were used.

(SRR1): Each full-time employee should have a total working time of 2,400 min. Each part-time employee should work 3 shifts.
(SRR3): Each employee should have at least 11 h of rest time between two adjacent shifts. Each violation of this rule incurs a cost of 1.

We scheduled 52 full-time employees with a total working time of 2,400 min and 17 full-time employees with a total working time of 2,370 min, which is optimal. There are 13 violations in the rest time constraint (SRR3). Every part-timer has 3 shifts. Thus the schedule is acceptable.

5 Our Solution Method

The PEAST algorithm [19, 20] is a population-based local search method. The acronym PEAST stems from the methods used: Population, Ejection, Annealing, Shuffling and Tabu. Aside from workforce scheduling, it has been used to solve real-world school timetabling problems [21] and real-world sports scheduling problems [22]. The PEAST algorithm uses GHCM, the Greedy Hill-Climbing Mutation heuristic introduced in [23] as its local search method. The pseudo-code of the algorithm is given in Fig. 2.

The reproduction phase of the algorithm is, to a certain extent, based on steady-state reproduction: the new schedule replaces the old one if it has a better or equal objective function value. Furthermore, the least fit is replaced with the best one when n better schedules have been found, where n is the size of the population. Marriage selection is used to select a schedule from the population of schedules for a single GHCM operation. In the marriage selection we randomly pick a schedule, S, and then attempt to randomly pick a better one at most $k - 1$ times. We choose the first better schedule, or, if none is found, we choose S.

```
Set the time limit t, no_change limit m and the population size n
Generate a random initial population of individuals
Set no_change = 0 and better_found = 0
WHILE elapsed_time < t
REPEAT n times
    Select an individual A by using a marriage selection with k = 3
    (explore promising areas in the search space)
    Apply GHCM to A to get a new individual A'
    Calculate the change Δ in objective function value
    IF Δ < = 0 THEN
        Replace A with A'
        IF Δ < 0 THEN
            better_found = better_found + 1
            no_change = 0
            END IF
        ELSE
            no_change = no_change + 1
    END IF
END REPEAT
IF better_found > n THEN
    Replace the worst individual with the best individual
    Set better_found = 0
END IF
IF no_change > m THEN
    (escape from the local optimum)
    Apply shuffling operators
    Set no_change = 0
END IF
 (avoid staying stuck in the promising search areas too long)
Update simulated annealing framework
Update the dynamic weights of the hard constraints (ADAGEN)
END WHILE
Choose the best individual from the population
```

Fig. 2 The pseudo-code of the PEAST algorithm

The heart of the GHCM heuristic is based on similar ideas to the Lin-Kernighan procedures [24] and ejection chains [25]. The basic hill-climbing step is extended to generate a sequence of *moves* in one step, leading from one solution candidate to another. The GHCM heuristic moves an *object*, o_1, from its old position in some *cell*, c_1, to a new cell, c_2, and then moves another object, o_2, from cell c_2 to a new cell, c_3, and so on, ending up with a sequence of moves. An object is a task-based activity or a whole break (in shift generation), a day-off (in days-off scheduling) or a shift (in shift scheduling). A cell is a shift (in shift generation) or an employee (in days-off scheduling and shift scheduling). A move involves removing an object from a certain position within a cell and inserting it either into a new cell (position is invariant) or a new position (cell is invariant).

The initial cell selection is random. The cell that receives an object is selected by considering all the possible cells and selecting the one that causes the least increase in the objective function when only considering the relocation cost. Then, another object from that cell is selected by considering all the objects in that cell and picking the one for which the removal causes the biggest decrease in the objective function when only considering the removal cost. Next, a new cell for that object is selected, and so on. The sequence of moves stops if the last move causes an increase in the objective function value and if the value is larger than that of the previous non-improving move. Then, a new sequence of moves is started. The initial solution is randomly generated.

The decision whether or not to commit to a sequence of moves in the GHCM heuristic is determined by a refinement [23] of the standard simulated annealing method [26]. Simulated annealing is useful to avoid staying stuck in the promising search areas for too long. The initial temperature T_0 is calculated by

$$T_0 = 1/\log(1/X_0) \tag{2}$$

where X_0 is the degree to which we want to accept an increase in the cost function (we use a value of 0.75). The exponential cooling scheme is used to decrement the temperature:

$$T_k = \alpha T_{k-1} \tag{3}$$

where α is usually chosen between 0.8 and 0.995. We stop the cooling at some predefined temperature. Therefore, after a certain number of iterations, m, we continue to accept an increase in the cost function with some constant probability, p. Using the initial temperature given above and the exponential cooling scheme, we can calculate the value

$$\alpha = (-1/(T_0 \log p))^{1/m}. \tag{4}$$

We choose m equal to the maximum number of iterations with no improvement to the cost function and p equal to 0.0015.

For most PEAST applications we introduce a number of shuffling operators—simple heuristics used to perturb a solution into a potentially worse solution in order to escape from local optima—that are called upon according to some rule. The most used

heuristics include moving a single random object from one cell to another random cell, or swapping two random objects between two random cells. For further details on the different shuffling operators used, see [13–15], [17, 18, 27]. The operator is called every $l/20$th iteration of the algorithm, where l equals the maximum number of iterations with no improvement to the cost function.

We use the weighted-sum approach for multi-objective optimization. We use the ADAGEN method [23] which assigns dynamic weights to the hard constraints. The weights are updated every kth generation using the formula given in [23]. The soft constraint weights are static yet instance-dependent.

References

1. Garey MR, Johnson DS (1979) Computers and intractability: a guide to the theory of NP-completeness. Freeman, New York
2. Tien J, Kamiyama A (1982) On manpower scheduling algorithms. SIAM Rev 24(3):275–287
3. Lau HC (1996) On the complexity of manpower shift scheduling. Comp Oper Res 23(1):93–102
4. Marx D (2004) Graph coloring problems and their applications in scheduling. Periodica Polytech Ser El Eng 48:5–10
5. Di Gaspero L, Gärtner J, Kortsarz G, Musliu N, Schaerf A, Slany W (2007) The minimum shift design problem. Ann Oper Res 155(1):79–105
6. Dantzig GB (1954) A comment on Edie's traffic delays at toll booths. Oper Res 2:339–341
7. Alfares HK (2004) Survey, categorization and comparison of recent tour scheduling literature. Ann Oper Res 127:145–175
8. Ernst AT, Jiang H, Krishnamoorthy M, Sier D (2004) Staff scheduling and rostering: a review of applications, methods and models. Eur J Oper Res 153(1):3–27
9. Meisels A, Schaerf A (2003) Modelling and solving employee timetabling problems. Ann Math Artif Intell 39:41–59
10. De Causmaecker P, Vanden Berghe G (2012) Towards a reference model for timetabling and rostering. Ann Oper Res 194(1):167–176
11. Nurmi K, Kyngäs N, Salli J (2012) A workload prediction, staffing and shift generation method for contact centers. In: Proceedings of the 4th EURO working group on stochastic modeling, Paris, France
12. Kellogg DL, Walczak S (2007) Nurse scheduling: from academia to implementation or not? Interfaces 37(4):355–369
13. Kyngäs J, Nurmi K (2011) Days-off scheduling for a bus transportation company. Int J Innovative Comput Appl 3(1):42–49
14. Kyngäs N, Nurmi K, Kyngäs J (2012) Optimizing large-scale staff rostering instances. In: Proceedings of the international multiconference of engineers and computer scientists, Hong Kong. Lecture notes in engineering and computer science, pp 1524–1531
15. Kyngäs N, Goossens D, Nurmi K, Kyngäs J (2012) Optimizing the unlimited shift generation problem. In: Proceedings of the international conference on the applications of evolutionary computation, Malaga, Spain, pp 508–518
16. Kyngäs N, Nurmi K, Kyngäs J (2013) The workforce scheduling process using the PEAST algorithm. In: Proceedings of the international multiconference of engineers and computer scientists, 13–15 Mar 2013, Hong Kong. Lecture notes in engineering and computer science, pp 1048–1056
17. Kyngäs N, Nurmi K, Ásgeirsson EI, Kyngäs J (2012) Using the PEAST algorithm to roster nurses in an intensive-care unit in a Finnish hospital. In: Proceedings of the 9th conference on the practice and theory of automated timetabling (PATAT), Son, Norway

18. Nurmi K, Kyngäs J, Post G (2012) Driver rostering for a Finnish transportation company. In: Ao Sio-Iong (ed) IAENG transactions on engineering technologies, vol 7. Springer, USA
19. Kyngäs J (2011) Solving challenging real-world scheduling problems. Ph.D. dissertation, Department of Information Technology, University of Turku, Finland. Available: http://urn.fi/URN:ISBN:978-952-12-2634-2
20. Kyngäs N, Nurmi K, Kyngäs J (2013) Crucial components of the PEAST algorithm in solving real-world scheduling problems. In: 2nd International conference on software and computer applications (ICSCA 2013) (submitted for publication)
21. Nurmi K, Kyngäs J (2007) A framework for school timetabling problem. In: Proceedings of the 3rd multidisciplinary international scheduling conference: theory and applications. France, Paris, pp 386–393
22. Kyngäs J, Nurmi K (2009) Scheduling the Finnish major ice hockey league. In: Proceedings of the IEEE symposium on computational intelligence in scheduling. Nashville, USA
23. Nurmi K (1998) Genetic algorithms for timetabling and traveling salesman problems, Ph.D. dissertation, Department of Applied Mathematics, University of Turku, Finland. Available: http://www.bit.spt.fi/cimmo.nurmi/dissertation/cimmodis.zip
24. Lin S, Kernighan BW (1973) An effective heuristic for the traveling salesman problem. Oper Res 21:498–516
25. Glover F (1992) New ejection chain and alternating path methods for traveling salesman problems. In: Sharda, Balci, Zenios (eds) Computer science and operations research: new developments in their interfaces. Elsevier, pp 449–509
26. van Laarhoven PJM, Aarts EHL (1987) Simulated annealing: theory and applications. Kluwer Academic Publishers
27. Kyngäs N, Nurmi K, Kyngäs J (2013) Solving the person-based multitask shift generation problem with breaks. In: 5th International conference on modeling, simulation and applied optimization (ICMSAO'13) (submitted for publication)

A New Hybrid Relaxed Extragradient Algorithm for Solving Equilibrium Problems, Variational Inequalities and Fixed Point Problems

Supak Phiangsungnoen and Poom Kumam

Abstract The purpose of this paper is to introduce a new hybrid relaxed extragradient iterative method for finding a common element of the solution of set a equilibrium, variational inequality and the set of fixed point of a ξ-strict pseudocontraction mapping in Hilbert spaces. We obtain a strong convergence theorem of the purposed iterative algorithm under some suitable conditions. The results presented in this paper generalize, improve and extend some well-known results in the literature.

Keywords Equilibrium problem · Fixed point problem · Hilbert spaces · Hybrid relaxed extragradient method · Strict pseudocontraction mapping · Strong convergence

1 Introduction

The equilibrium problems, found application in optimization problems, fixed point problems, convex minimization problems. In other words, equilibrium problems are a unified model for problems arising in physics, engineering, economics, and so on (see [1]).

Throughout in this paper, we always assume that H be a real Hilbert space whose inner product and norm denoted by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$, respectively. Let C be a nonempty closed convex subset of H and $F : C \times C \to R$, where R denotes the set of real number, a bifunction. We concider the Ky Fan inequality which was first introduced by Fan [2] as follows:

S. Phiangsungnoen · P. Kumam (✉)
Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi, 126 Pracha-Uthit Road, Bangmod, Thung khru, Bangkok 10140, Thailand
e-mail: poom.kum@kmutt.ac.th

S. Phiangsungnoen
e-mail: supuk_piang@hotmail.com

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 275, DOI: 10.1007/978-94-007-7684-5_26,

$$\text{find} \quad x \in C \quad \text{such that} \quad F(x, y) \geq 0, \quad \forall y \in C.$$

The set of $x^* \in C$ is denoted by $EP(F)$, i.e.,

$$EP(F) = \{x^* \in C : F(x^*, y) \geq 0, \quad \forall y \in C\}. \tag{1}$$

Numerous problems in physics, optimization, and economics reduce to find a solution of (1). Some methods have been proposed to solve the Ky Fan inequality (or some time called equilibrium problem, see [3–6]).

Given a mapping $A : C \to H$ is a nonlinear mapping, we can choose $F(x, y) = \langle Ax, y - x\rangle$, so an equilibrium point is a solution of variational inequality problem, denoted by $VI(A, C)$.

$$\text{Find}\, x^* \in C \,\text{such that}\, \langle Ax^*, y - x^*\rangle \geq 0, \quad \forall y \in C.$$

Recall that a nonself-mapping $S : C \to H$ is called a ξ-strict pseudocontraction if there exists a constant $\xi \in [0, 1)$ such that

$$\|Sx - Sy\|^2 \leq \|x - y\|^2 + \xi\|(x - Sx) - (y - Sy)\|^2 \tag{2}$$

for every $x, y \in C$.

We use $Fix(S)$ to denote the fixed point set of the mapping S, that is, $Fix(S) = \{x \in C : Sx = x\}$. As $\xi = 0$, S is said to be nonexpansive, that is, $\|Sx - Sy\| \leq \|x - y\|$, for all $x, y \in C$. S is said to be pseudocontractive if $\xi = 1$ and is also said to be strongly pseudocontractive if there exists a positive constant $\lambda \in (0, 1)$ such that $S + \lambda I$ is pseudocontractive. Clearly, the class of a ξ-strict pseudocontractions falls into the one between classes of nonexpansive mappings and pseudocontractions. Moreover, strict pseudocontractions have more powerful applications than nonexpansive mappings do in solving inverse problems (see, e.g., [7]). Therefore, it is interesting to develop the theory of iterative methods for equilibrium, variational inequality and fixed point problems which S is a ξ strict pseudocontraction mapping.

In 1953, Mann [8] introduced the iteration as follows: a sequence $\{x_n\}$ defined by

$$x_{n+1} = \alpha_n x_n + (1 - \alpha_n)Sx_n, \tag{3}$$

where the initial guess element $x_0 \in C$ is arbitrary and $\{\alpha_n\}$ is a real sequence in [0, 1]. The Mann iteration has been extensively investigated for nonexpansive mappings. One of the fundamental convergence results is proved by Reich [9].

In an infinite-dimensional Hilbert space, the Mann iteration can conclude only weak convergence [10]. Attempts to modify the Mann iteration method (3) so that strong convergence is guaranteed have recently been made. Nakajo and Takahashi [11] proposed the following modification of the Mann iteration method (3) as follows: $x_0 \in C$ is arbitrary,

$$\begin{cases} y_n = \alpha_n x_n + (1-\alpha_n) S x_n, \\ C_n = \{z \in C : \|y_n - z\| \le \|x_n - z\|\}, \\ Q_n = \{z \in C : \langle x_n - z,\ x_0 - x_n \rangle \ge 0\}, \\ x_{n+1} = P_{C_n \cap Q_n} x_0, \quad n = 0, 1, 2 \ldots, \end{cases} \tag{4}$$

where P_C is metric projection on C.

For finding an element of $Fix(S) \cap VI(A, C)$, Takahashi and Toyoda [12] introduced the following iterative scheme:

$$x_{n+1} = \alpha_n x_n + (1-\alpha_n) S P_C(x_n - \lambda_n A x_n) \tag{5}$$

for every $n = 0, 1, 2, \ldots$, where P_C is the metric projection on the set C, $x_0 = x \in C$, $\{\alpha_n\}$ is a sequence in $(0, 1)$ and $\{\lambda_n\}$ is a sequence in $(0, 2\alpha)$. They proved a weak convergence theorem in a Hilbert space.

In 2011, Ceng et al. [13], they proved the strong convergence of the sequences generated by the iterative scheme

$$\begin{cases} u_n = P_C[P_C(x_n - \mu_2 B_2 x_n) - \mu_1 B_1 P_C(x_n - \mu_2 B_2 x_n)], \\ \tilde{u} = P_C(u_n - \lambda_n A u_n), \\ y_n = \alpha Q x_n + (1-\alpha) P_C(u_n - \lambda_n A \tilde{u}), \\ x_{n+1} = \beta_n x_n + \gamma_n y_n + \delta_n S y_n, \end{cases} \tag{6}$$

where $\mu_i \in (0, \beta_i)$ for $i = 1, 2$, $\{\lambda\} \subset (0, \alpha]$ and $\{\alpha_n\}, \{\beta_n\}, \{\gamma_n\}, \{\delta_n\} \subset [0, 1]$ such that $\beta_n + \gamma_n + \delta_n = 1$, for all $n \ge 0$.

Recently, Vuong et al. [14] was introduced the iterative for finding a common element of the set of points satisfying a Ky Fan inequality, and the set of fixed points of a contraction mapping in a Hilbert space, they considered the sequences $\{x_n\}, \{y_n\}, \{z_n\}$, and $\{t_n\}$ generated by $x_0 \in C$ and

$$\begin{cases} y_n = \arg\min_{y \in C} \{\lambda_n F(x_n, y) + \frac{1}{2}\|y - x_n\|^2\}, \\ z_n = \arg\min_{y \in C} \{\lambda_n F(y_n, y) + \frac{1}{2}\|y - x_n\|^2\}, \\ t_n = \alpha_n x_n + (1-\alpha_n)[\beta_n z_n + (1-\beta_n) S z_n], \\ x_{n+1} = t_n, \end{cases} \tag{7}$$

for every $n \in N$, where $\{\alpha_n\} \subset [0, 1[$, $\{\beta_n\} \subset]0, 1[$, and $\{\lambda_n\} \subset]0, 1]$.

Very recently, Phiangsungnoen and Kumam [15] introduced a new hybrid extragradient iterative method for finding a common element of the set of points satisfying the Ky Fan inequalities, variational inequality and the set of fixed points of a strict pseudocontraction mapping in Hilbert spaces. They obtained the strong convergence of an iterative algorithm generated by the hybrid extragradient projection method,

under some suitable assumptions the function associated with Ky Fan inequality is pseudomonotone and weakly continuous.

In this paper, we introduce and study a hybrid relaxed extragradient method for solving equilibrium problem with a fixed-point method and variational inequality method. More precisely, we consider the sequences $\{x_n\}$, $\{y_n\}$, $\{z_n\}$, $\{u_n\}$, $\{w_n\}$ and $\{t_n\}$ generated by $x_0 \in C$ and

$$\begin{cases} y_n = \arg\min_{y\in C}\{\lambda_n F(x_n, y) + \frac{1}{2}\|y - x_n\|^2\}, \\ z_n = \arg\min_{y\in C}\{\lambda_n F(y_n, y) + \frac{1}{2}\|y - x_n\|^2\}, \\ w_n = P_C(z_n - \delta_n A z_n), \\ u_n = P_C(w_n - \varphi_n B w_n), \\ t_n = \alpha_n x_n + (1 - \alpha_n)[\beta_n S u_n + (1 - \beta_n)u_n], \\ C_n = \{z \in C : \|t_n - z\| \le \|x_n - z\|\}, \\ D_n = \{z \in C : \langle x_n - z, x_0 - x\rangle \ge 0\}, \\ x_{n+1} = P_{C_n \cap D_n} x_0, \end{cases} \tag{8}$$

for every $n \in N$, where $\{\alpha_n\} \subset [0, 1)$, $\{\lambda_n\} \subset (0, 1]$, $\{\delta_n\} \subset (0, 2\alpha)$ and $\{\varphi_n\} \subset (0, 2\beta)$. Then, we obtain the strong convergent theorem for fixed point problems, variational inequalities, and equilibrium problems.

2 Preliminaries

We need some facts and tools in a real Hilbert space H which are listed below.

Lemma 1 *Let H be a real Hilbert space. Then, the following identities hold.*

(1) $\|x - y\|^2 = \|x\|^2 - \|y\|^2 - 2\langle x - y, y\rangle$
(2) $\|\lambda x + (1 - \lambda)y\|^2 = \lambda\|x\|^2 + (1 - \lambda)\|y\|^2 - \lambda(1 - \lambda)\|x - y\|^2$

for all $x, y \in H$ and $\lambda \in [0, 1]$.

Recall that the (nearest point) projection P_C from H onto C assigns to each $x \in H$ the unique point in $P_C x \in C$ satisfying property

$$\|x - P_C x\| \le \|x - y\| \quad \text{for all } y \in C.$$

Lemma 2 *For a given $x \in H$, $z \in C$*

$$z = P_C x \Leftrightarrow \langle z - x, y - x\rangle \quad \textit{for all } y \in C.$$

It is well known that P_C is a firmly nonexpansive mapping of H onto C and satisfies

$$\|P_C x - P_C y\|^2 \leq \langle P_C x - P_C y, x - y\rangle \quad \text{for all } x, y \in H. \tag{9}$$

Moreover, $P_C x$ is characterized by the following properties: $P_C \in C$ and for all $x \in H$, $y \in C$,

$$\langle x - P_C x, y - P_C x\rangle \leq 0. \tag{10}$$

Lemma 3 *Let H be a Hilbert space, let C be a nonempty closed convex subset of H and let A be a mapping of C into H. Let $u \in C$. Then for $\lambda > 0$,*

$$u \in VI(A, C) \Leftrightarrow u = P_C(u - \lambda Au), \text{ for all } \lambda > 0. \tag{11}$$

where P_C is the metric projection of H onto C.

It is also known that set valued mapping $T : H \to 2^H$ is called monotone if for all $x, y \in H$, $f \in Tx$ and $g \in Ty$ imply $\langle x - y, f - g\rangle \geq 0$. A monotone mapping $T : H \to 2^H$ is maximal if the graph $G(T)$ is not properly contained in the graph of any other monotone mapping. It known that a monotone mapping T is maximal if and only if for $(x, f) \in H \times H$, $\langle x - y, f - g\rangle \geq 0$ for every $(y, g) \in G(T)$ implies $f \in Tx$. Let A be an inverse-strongly monotone mapping of C into H and let $N_C v$ be the normal cone to C at $v \in C$, i.e.,

$$N_C v = \{w \in H : \langle v - u, w\rangle \geq 0, \quad \forall u \in C\}$$

and define

$$Tv = \begin{cases} Av + N_C v, & v \in C, \\ \emptyset, & v \notin C. \end{cases} \tag{12}$$

Then T is maximal monotone and $0 \in Tv$ if and only if $v \in VI(A, C)$; see [16, 17].

Lemma 4 [18] *Each Hilbert space H satisfies the Opial's condition, i.e., for any sequence $\{x_n\}$ with $x_n \rightharpoonup x$, the inequality*

$$\liminf_{n\to\infty} \|x_n - x\| < \liminf_{n\to\infty} \|x_n - y\|$$

holds for every $y \in H$ with $y \neq x$.

Lemma 5 [19, 20] *Each Hilbert space H satisfies the Kadec-Klee property, that is, for any sequence $\{x_n\}$ with $x_n \rightharpoonup x$ and $\|x_n\| \to \|x\|$ together imply $\|x_n - x\| \to 0$.*

For solving the equilibrium problem, let us assume that the following conditions are satisfied on the bifunction $F : C \times C \to R$.

(A1) $F(x, x) = 0$ for every $x \in C$;
(A2) F is pseudomonotone on C, i.e., $F(x, y) \geq 0 \Rightarrow F(y, x) \leq 0$ $\forall x, y \in C$;

(A3) F is jointly weakly continuous on $C \times C$ in the sense that, if $x, y \in C$ and $\{x_n\}$ and $\{y_n\}$ are two sequences in C converging weakly to x and y, respectively, then $F(x_n, y_n) \to F(x, y)$;

(A4) $F(x, \cdot)$ is convex, lower semicontinuous, and subdifferentiable on C foe very $x \in C$;

(A5) F satisfies the Lipschitz-type condition, there exist positive integer c_1 and c_2, such that for every $x, y, z \in C$,

$$F(x, y) + F(y, z) \geq F(x, z) - c_1\|y - x\|^2 - c_2\|z - y\|^2.$$

If F satisfies the properties (A1)–(A4), then the set $EP(F)$ of solutions to the Ky Fan inequality is closed and convex. The example of function F satisfying assumption (A5), is given by

Example 6 Let $G : C \to H$ is Lipschitz continuous on C (with constant $L > 0$) (see [21]).

$$F(x, y) = \langle G(x), y - x\rangle \quad \forall x, y \in C,$$

where In that example, $c_1 = c_2 = L\backslash 2$. Another example, related to the Cournot-Nash equilibrium model, is described in [22].

Proposition 7 ([23], *Lemma* 3.1) *For every $x^* \in EP(F)$, and every $n \in N$, one has*

(1) $\langle x_n - y_n, y - y_n\rangle \leq \lambda_n f(x_n, y) - \lambda_n f(x_n, y_n), \quad \forall y \in C;$
(2) $\|z_n - x^*\|^2 \leq \|x_n - x^*\|^2 - (1 - 2\lambda_n c_1)\|y_n - x_n\|^2 - (1 - 2\lambda_n c_2)\|z_n - y_n\|^2.$

Proposition 8 [24] *Let K be a nonempty closed and convex subset of H. Let $u \in H$ and let $\{x_n\}$ be a sequence in H. If any weak limit point of $\{x_n\}$ belongs to K, and $\|x_n - u\| \leq \|u - P_K u\|$ for all $n \in N$, then $x_n \to P_K u$.*

In order to apply this Proposition in our context, we set $K = EP(F) \cap Fix(S) \cap VI(A, C) \cap VI(B, C)$ and $u = x_0$. Furthermore, we impose that the sequence $\{x_n\}$ generated by our algorithms satisfies, for all $n \in N$, in equality

$$\|x_n - x_0\| \leq \|x_{n+1} - x_0\| \leq \|\tilde{x_0} - x_0\|,$$

where $\tilde{x_0} = P_{EP(F)\cap Fix(S)\cap VI(A,C)\cap VI(B,C)}x_0$. In that case, the sequence $\{\|x_n - x_0\|\}$ is convergent and the sequence $\{x_n\}$ is bounded. These properties will be useful to prove that any weak limit point of $\{x_n\}$ belongs to $EP(F) \cap Fix(S) \cap VI(A, C) \cap VI(B, C)$.

3 Main results

In this section, we prove the strong convergence of the sequences generated by the iterative scheme (8), we obtain the following basic algorithm.

Algorithm 9 Choose the sequences $\{\alpha_n\} \subset [0, 1)$, $\{\beta_n\} \subset (0, 1)$, $\{\lambda_n\} \subset (0, 1]$ and ξ be a constant in $[0, 1)$.

Step (1) Let $x_0 \in C$. Set $n = 0$
Step (2) Solve successively the strongly convex programs $\arg\min_{y\in C}\{\lambda_n F(x_n, y) + \frac{1}{2}\|y - x_n\|^2\}$ and $\arg\min_{y\in C}\{\lambda_n F(y_n, y) + \frac{1}{2}\|y - x_n\|^2\}$ to obtain the unique optimal solution y_n and z_n, respectively,
Step (3) Compute $w_n = P_C(z_n - \delta_n A z_n)$,
Step (4) Compute $u_n = P_C(w_n - \varphi_n B w_n)$,
Step (5) Compute $t_n = \alpha_n x_n + (1 - \alpha_n)[\beta_n S u_n + (1 - \beta_n)u_n]$. If $y_n = x_n$ and $t_n = x_n$, then STOP: $x_n \in EP(F) \cap Fix(S) \cap VI(A, C) \cap VI(B, C)$. Otherwise, go to Step (6).
Step (6) Compute $x_{n+1} = P_{C_n \cap D_n} x_0$, where $C_n = \{z \in C : \|t_n - z\| \leq \|x_n - z\|\}$, $D_n = \{z \in C : \langle x_n - z, x_0 - x\rangle \geq 0\}$,
Step (7) Set $n := n + 1$, and go to Step (2).

Theorem 1 *Let C be a nonempty, closed and convex subset of a real Hilbert space H, let f be a bifunction $f : C \times C$ into R satisfying conditions (A1)–(A5), let A be an α-inverse-strongly monotone mapping of C into H and B be a β-inverse-strongly monotone mapping of C into H. Let S be a ξ-strict pseudocontraction mapping from C to C and such that $\Omega := EP(F) \cap Fix(S) \cap VI(A, C) \cap VI(B, C) \neq \emptyset$. Let $\{x_n\}$, $\{y_n\}$, $\{z_n\}$, $\{u_n\}$ and $\{w_n\}$ be sequences generated by $x_0 \in C$. For every $n \in N$, where $\{\alpha_n\} \subset [0, 1)$, $\{\lambda_n\} \subset (0, 1]$, $\{\beta_n\} \subset [0, 1)$, $\{\delta_n\} \subset (0, 2\alpha)$, and $\{\varphi_n\} \subset (0, 2\beta)$ satisfy the following conditions:*

(1) $\{\lambda_n\} \subset [\lambda_{\min}, \lambda_{\max}]$, *where* $0 < \lambda_{\min} \leq \lambda_{\max} < \min\left\{\frac{1}{2c_1}, \frac{1}{2c_2}\right\}$,
(2) $\{\delta_n\} \subset [a, b]$ *for some* a, b *with* $0 < a < b < 2\alpha$,
(3) $\{\varphi_n\} \subset [c, d]$ *for some* d, e *with* $0 < c < d < 2\beta$,
(4) $\{\beta_n\} \subset [e, f]$ *for some* $0 \leq \xi < e \leq f < 1$, *and* $\lim_{n\to\infty} \alpha_n = 0$,

Then the sequence $\{x_n\}$ generated by Algorithm 9 converges strongly to the projection of x_0 on to the set Ω.

Proof First, we show that the mapping $I - \delta_n A$ and $I - \lambda_n B$ are nonexpansive for each $n \geq 1$. Indeed, for any $x, y \in C$, since A is an α-inverse-strongly monotone mapping and B is a β-inverse-strongly monotone mapping, we have

$$\begin{aligned}\|(I - \delta_n A)x - (I - \delta_n A)y\|^2 &= \|(x - y) - \delta_n(Ax - Ay)\|^2 \\ &\leq \|x - y\|^2 + \delta_n(\delta_n - 2\alpha)\|Ax - Ay\|^2 \\ &\leq \|x - y\|^2.\end{aligned}$$

which implies that the mapping $I - \delta_n A$ is nonexpansive, and so is $I - \lambda_n B$. We will devide the proof into several steps.

Step 1. We show that sequence $\{x_n\}$ is well defined. Obviously, C_n and D_n are closed and D_n is closed and convex for every $n \in N$. We prove that C_n is convex. Since

$$C_n = \{z \in C : \|t_n - z\| \leq \|x_n - z\|\}$$

we can write C_n under the form

$$C_n = \{z \in C : \|t_n - x_n\|^2 + 2\langle t_n - x_n, x_n - z\rangle \leq 0\},$$

it follows that C_n is convex. So $C_n \cap D_n$ is closed and convex subset of H for any $n \in N$. This implies that $\{x_n\}$ is well defined.

Step 2. We show that $\Omega \subset C_n \cap D_n$ for each $n \geq 1$. Let $x^* \in \Omega$. Then $Sx^* = x^*$ and $y^* = P_C(x^* - \delta_n Ax^*)$ and $x^* = P_C[P_C(x^* - \delta_n Ax^*) - \varphi_n B P_C(x^* - \delta_n Ax^*)]$. Since A be an α-inverse-strongly monotone mapping of C into H and B be a β-inverse-strongly monotone mapping of C into H. By Proposition 7 (2), we have

$$\|z_n - x^*\|^2 \leq \|x_n - x^*\|^2 - (1 - 2\lambda_n c_1)\|y_n - x_n\|^2 - (1 - 2\lambda_n c_2)\|z_n - y_n\|^2 \quad (13)$$

that is, $\|z_n - x^*\| \leq \|x_n - x^*\|$.

Put $y^* = P_C(x^* - \delta_n Ax^*)$ and $w_n = P_C(z_n - \delta_n Az_n)$, we have

$$\begin{aligned} \|w_n - y^*\|^2 =& \|P_C(z_n - \delta_n Az_n) - P_C(x^* - \delta_n Ax^*)\|^2 \\ &\leq \|(z_n - \delta_n Az_n) - (x^* - \delta_n Ax^*)\|^2 \\ &\leq \|z_n - x^*\|^2 + \delta_n(\delta_n - 2\alpha)\|Az_n - Ax^*\|^2 \\ &\leq \|z_n - x^*\|^2. \end{aligned} \quad (14)$$

Since $\|z_n - x^*\| \leq \|x_n - x^*\|$, that is $\|w_n - y^*\| \leq \|x_n - x^*\|$.

For simplicity, we write $u_n = P_C(w_n - \varphi_n Bw_n)$ and $x^* = P_C(y^* - \varphi_n By^*)$, we have

$$\begin{aligned} \|u_n - x^*\|^2 =& \|P_C(w_n - \varphi_n Bw_n) - P_C(y^* - \varphi_n By^*)\|^2 \\ &\leq \|(w_n - \varphi_n Bw_n) - (y^* - \varphi_n By^*)\|^2 \\ &\leq \|w_n - y^*\|^2 + \varphi_n(\varphi_n - 2\beta)\|Bw_n - By^*\|^2 \\ &\leq \|w_n - y^*\|^2 \end{aligned} \quad (15)$$

that is, $\|u_n - x^*\| \leq \|w_n - y^*\|$ and $\|u_n - x^*\| \leq \|x_n - x^*\|$.

Define $t_n = \alpha_n x_n + (1 - \alpha_n)[\beta_n Su_n + (1 - \beta_n)u_n]$, for all $n \geq 0$. It follows that

$$
\begin{aligned}
&\|t_n - x^*\|^2 \\
&= \|\alpha_n x_n + (1-\alpha_n)[\beta_n S u_n + (1-\beta_n)u_n] - x^*\|^2 \\
&= \|\alpha_n(x_n - x^*) + (1-\alpha_n)[\beta_n S u_n + (1-\beta_n)u_n - x^*]\|^2 \\
&\quad \le \alpha_n\|x_n - x^*\|^2 + (1-\alpha_n)\|\beta_n(S u_n - x^*) + (1-\beta_n)(u_n - x^*)\|^2 \\
&= \alpha_n\|x_n - x^*\|^2 + (1-\alpha_n)[\beta_n\|S u_n - x^*\|^2 + (1-\beta_n)\|u_n - x^*\|^2 \\
&\quad - \beta_n(1-\beta_n)\|S u_n - u_n\|^2] \\
&\quad \le \alpha_n\|x_n - x^*\|^2 + (1-\alpha_n)\beta_n[\|u_n - x^*\|^2 + \xi\|u_n - S u_n\|^2] \\
&\quad + (1-\alpha_n)(1-\beta_n)\|u_n - x^*\|^2 - (1-\alpha_n)\beta_n(1-\beta_n)\|S u_n - u_n\|^2 \\
&\quad \le \|x_n - x^*\|^2 - (1-\alpha_n)(1-2\lambda_n c_1)\|y_n - x_n\|^2 - (1-\alpha_n)(1-2\lambda_n c_2)\|z_n - y_n\|^2 \\
&\quad + (1-\alpha_n)\delta_n(\delta_n - 2\alpha)\|A z_n - A x^*\|^2 + (1-\alpha_n)\varphi_n(\varphi_n - 2\beta)\|B w_n - B y^*\|^2 \\
&\quad - (1-\alpha_n)(1-\beta_n)(\beta_n - \xi)\|S u_n - u_n\|^2.
\end{aligned} \tag{16}
$$

Hence $\|t_n - x^*\| \le \|x_n - x^*\|$ for every $n \ge 0$ and $x^* \in C_n$. So, we have $\Omega \subset C_n$ for all $n \in N$. Next, we prove by induction that

$$\Omega \subset C_n \cap D_n, \quad \forall n \in N \cup \{0\}.$$

For $n = 0$, we have $x_0 = x \in C$, $\Omega \subset C_0$ and $D_0 = C$. So, we get $\Omega \subset C_0 \cap D_0$. Suppose that $\Omega \subset C_k \cap D_k$ for some $k \in N$. Since $C_k \cap D_k$ is closed and convex, we can define $x_{k+1} = P_{C_k \cap D_k}(x_0)$, we have

$$\langle x_{k+1} - p, x_0 - x_{k+1}\rangle \ge 0, \quad \forall p \in C_k \cap D_k.$$

Since $\Omega \subset C_k \cap D_k(x_0)$, we also have

$$\langle x_{k+1} - x^*, x_0 - x_{k+1}\rangle \ge 0, \quad \forall x^* \in \Omega.$$

So, we get $\Omega \subset D_{k+1}$. Then we obtain $\Omega \subset C_{k+1} \cap D_{k+1}$.

Step 3. Next, we show that $\{x_n\}$ is bounded and $\lim_{n\to\infty}\|x_{n+1} - x_n\| = 0$ and $\lim_{n\to\infty}\|x_n - t_n\| = 0$. Put $z_0 = P_\Omega(x_0)$. From $x_{n+1} = P_{C_n \cap D_n}$, we get

$$\|x_{n+1} - x_0\| \le \|z - x_0\|, \quad \forall z \in C_n \cap D_n.$$

From $z_0 \in \Omega \subset C_n \cap D_n$, we also have

$$\|x_{n+1} - x_0\| \le \|z_0 - x_0\|$$

for all $n \in N \cup \{0\}$. Hence $\{x_n\}$ is bounded.

Since $x_{n+1} \in C_n \cap D_n \subset D_n$ and $x_n = P_{D_n}(x_0)$, we get $\|x_n - x_0\| \le \|x_{n+1} - x_0\|$. Since $\{x_n\}$ is bounded and nondecreasing sequence, we get $\lim_{n\to\infty}\|x_n - x_0\|$ exists. So, we obtain $(\|x_n - x_0\|^2 - \|x_{n+1} - x_0\|^2) \to 0$. On the other hand, from $x_{n+1} \in D_n$, we have

$$\langle x_n - x_{n+1}, x_0 - x_n \rangle \geq 0.$$

So, for $n \in N \cup \{0\}$, we get

$$\begin{aligned}\|x_n - x_{n+1}\|^2 = & \|x_n - x_0 + x_0 - x_{n+1}\|^2 \\ & \leq \|x_{n+1} - x_0\|^2 - \|x_n - x_0\|^2 - 2\langle x_n - x_{n+1}, x_0 - x_n \rangle \\ & \leq \|x_{n+1} - x_0\|^2 - \|x_n - x_0\|^2.\end{aligned}$$

This implies

$$\|x_{n+1} - x_n\| \to 0. \tag{17}$$

Since $x_{n+1} \in C_n$, we have

$$\begin{aligned}C_n = \{z \in C : \|t_n - z\| \leq \|x_n - z\|\}; \\ \|t_n - x_{n+1}\| \leq \|x_n - x_{n+1}\|\end{aligned}$$

and

$$\|x_n - t_n\| \leq \|x_n - x_{n+1}\| + \|x_{n+1} - t_n\| \leq 2\|x_n - x_{n+1}\|.$$

By (17), we obtain

$$\lim_{n\to\infty} \|x_n - t_n\| = 0. \tag{18}$$

Step 4. We show that $\lim_{n\to\infty} \|Su_n - u_n\| = 0$, $\lim_{n\to\infty} \|Bw_n - By^*\| = 0$, $\lim_{n\to\infty} \|Az_n - Ax^*\| = 0$, $\lim_{n\to\infty} \|y_n - x_n\| = 0$, and $\lim_{n\to\infty} \|z_n - y_n\| = 0$. Since $x^* \in \Omega$, from ([16]) and the sequence $\{x_n\}$, $\{t_n\}$ are bounded, we have

$$\begin{aligned}&\|t_n - x^*\|^2 \\ &\leq \|x_n - x^*\|^2 - (1-\alpha_n)(1-2\lambda_n c_1)\|y_n - x_n\|^2 - (1-\alpha_n)(1-2\lambda_n c_2)\|z_n - y_n\|^2 \\ &+ (1-\alpha_n)\delta_n(\delta_n - 2\alpha)\|Az_n - Ax^*\|^2 + (1-\alpha_n)\varphi_n(\varphi_n - 2\beta)\|Bw_n - By^*\|^2 \\ &- (1-\alpha_n)(1-\beta_n)(\beta_n - \xi)\|Su_n - u_n\|^2.\end{aligned} \tag{19}$$

It follows that

$$\begin{aligned}(1-\alpha_n)(1-\beta_n)(\beta_n - \xi)\|Su_n - u_n\|^2 \\ \leq \|x_n - x^*\|^2 - \|t_n - x^*\|^2 \\ \leq [\|x_n - x^*\| + \|t_n - x^*\|]\|x_n - t_n\|.\end{aligned} \tag{20}$$

Since $\lim_{n\to\infty} \|x_n - t_n\| = 0$ and by condition (1)–(4), we obtain $\lim_{n\to\infty} \|Su_n - u_n\| = 0$, $\lim_{n\to\infty} \|Bw_n - By^*\| = 0$, $\lim_{n\to\infty} \|Az_n - Ax^*\| = 0$, $\lim_{n\to\infty} \|y_n - x_n\| = 0$ and $\lim_{n\to\infty} \|z_n - y_n\| = 0$. It easy to see that $\lim_{n\to\infty} \|z_n - x_n\| = 0$.

Step 5. We show that, $\lim_{n\to\infty}\|u_n - w_n\| = 0, \lim_{n\to\infty}\|u_n - x_n\| = 0$ and $\lim_{n\to\infty}\|Su_n - x_n\| = 0$.

$$\begin{aligned}
&\|u_n - x^*\|^2\\
&= \| P_C(w_n - \varphi_n Bw_n) - P_C(y^* - \varphi_n By^*)\|^2 \le \big\langle (w_n - \varphi_n Bw_n) - (y^* - \varphi_n By^*), u_n - x^*\big\rangle\\
&\quad \le \frac{1}{2}\big[\|w_n - y^*\|^2 + \|u_n - x^*\|^2 - \|(w_n - u_n) - \varphi_n(Bw_n - By^*) + (x^* - y^*)\|^2\big]\\
&\quad \le \frac{1}{2}\big[\|w_n - y^*\|^2 + \|u_n - x^*\|^2 - \|w_n - u_n\|^2 + 2\varphi_n\langle w_n - u_n, Bw_n - By^*\rangle\\
&\quad - \varphi_n^2\|Bw_n - By^*\|^2\big]\\
&\quad \le \frac{1}{2}\big[\|x_n - x^*\|^2 - \|w_n - u_n\|^2 + 2\varphi_n\langle w_n - u_n, Bw_n - By^*\rangle - \varphi_n^2\|Bw_n - By^*\|^2\big].
\end{aligned}$$

It follows that

$$\|u_n - x^*\|^2 \le \|x_n - x^*\|^2 - \|w_n - u_n\|^2 + 2\varphi_n\langle w_n - u_n, Bw_n - By^*\rangle - \varphi_n^2\|Bw_n - By^*\|^2.$$

Consequently, we obtain

$$\begin{aligned}
&\|t_n - x^*\|^2\\
&\le \alpha_n\|x_n - x^*\|^2 + (1-\alpha_n)[\|x_n - x^*\|^2 - \|w_n - u_n\|^2\\
&\quad + 2\varphi_n\langle w_n - u_n, Bw_n - By^*\rangle - \varphi_n^2\|Bw_n - By^*\|^2]\\
&\quad = \|x_n - x^*\|^2 - (1-\alpha_n)\|w_n - u_n\|^2 + (1-\alpha_n)2\varphi_n\langle w_n - u_n, Bw_n - By^*\rangle\\
&\quad - (1-\alpha_n)\varphi_n^2\|Bw_n - By^*\|^2]
\end{aligned}$$

which implies

$$\begin{aligned}
&(1-\alpha_n)\|w_n - u_n\|^2\\
&\quad \le \|x_n - x^*\|^2 - \|t_n - x^*\|^2 + (1-\alpha_n)2\varphi_n\langle w_n - u_n, Bw_n - By^*\rangle\\
&-(1-\alpha_n)\varphi_n^2\|Bw_n - By^*\|^2]\\
&\quad \le [\|x_n - x^*\| + \|t_n - x^*\|]\|x_n - t_n\| + 2\varphi_n(1-\alpha_n)\|w_n - u_n\|\|Bw_n - By^*\|.
\end{aligned}$$

Since $1-\alpha_n > 0$, $\lim_{n\to\infty}\|x_n - t_n\| = 0$ and $\lim_{n\to\infty}\|Bw_n - By^*\| = 0$. From $\{t_n\}$, $\{w_n\}$ and $\{u_n\}$ are bounded, we have $\lim_{n\to\infty}\|w_n - u_n\| = 0$, which implies that $\lim_{n\to\infty}\|w_n - x_n\| = 0$. We consider

$$\begin{aligned}
\|w_n - y^*\|^2 = \|P_C(z_n - \delta_n Az_n) - P_C(x^* - \delta_n Ax^*)\|^2&\\
\le \big\langle (z_n - \delta_n Az_n) - (x^* - \delta_n Ax^*), w_n - y^*\big\rangle&\\
\le \frac{1}{2}\big[\|z_n - x^*\|^2 + \|w_n - y^*\|^2 - \|(z_n - w_n) - (x^* - y^*)\|^2&\\
+ 2\delta_n\|(z_n - w_n) - (x^* - y^*)\|\|Az_n - Ax^*\|\big]&
\end{aligned}$$

that is,

$$\begin{aligned}\|w_n - y^*\|^2 \leq \|z_n - x^*\|^2 - \|(z_n - w_n) - (x^* - y^*)\|^2 \\ + 2\delta_n\|(z_n - w_n) - (x^* - y^*)\|\|Az_n - Ax^*\|.\end{aligned} \tag{21}$$

Similarly to the above argument, we derive

$$\begin{aligned}\|u_n - x^*\|^2 &= \| P_C(w_n - \varphi_n Bw_n) - P_C(y^* - \varphi_n By^*)\|^2 \\ &\leq \langle (w_n - \varphi_n Bw_n) - (y^* - \varphi_n By^*), u_n - x^*\rangle \\ &\leq \frac{1}{2}\big[\|w_n - y^*\|^2 + \|u_n - x^*\|^2 - \|(w_n - u_n) + (x^* - y^*)\|^2 \\ &+ 2\varphi_n\|(w_n - u_n) + (x^* - y^*)\|\|Bw_n - By^*\|\big]\end{aligned}$$

that is,

$$\begin{aligned}\|\|u_n - x^*\|^2 \leq \|w_n - y^*\|^2 - \|(w_n - u_n) + (x^* - y^*)\|^2 \\ + 2\varphi_n\|(w_n - u_n) + (x^* - y^*)\|\|Bw_n - By^*\|.\end{aligned} \tag{22}$$

From (13), (21) and ([22]), it follows that

$$\begin{aligned}\|u_n - x^*\|^2 \leq \|x_n - x^*\|^2 - (1 - 2\lambda_n c_1)\|y_n - x_n\|^2 - (1 - 2\lambda_n c_2)\|z_n - y_n\|^2 \\ -\|(z_n - w_n) - (x^* - y^*)\|^2 + 2\delta_n\|(z_n - w_n) - (x^* - y^*)\| \\ \|Az_n - Ax^*\| - \|(w_n - u_n) + (x^* - y^*)\|^2 + 2\varphi_n\|(w_n - u_n) \\ + (x^* - y^*)\|\|Bw_n - By^*\|\end{aligned} \tag{23}$$

Hence from ([16]) and (23)

$$\begin{aligned}&\|t_n - x^*\|^2 \\ &\leq \|x_n - x^*\|^2 - (1 - \alpha_n)(1 - 2\lambda_n c_1)\|y_n - x_n\|^2 - (1 - \alpha_n)(1 - 2\lambda_n c_2)\|z_n - y_n\|^2 \\ &- (1 - \alpha_n)\|(z_n - w_n) - (x^* - y^*)\|^2 + (1 - \alpha_n)2\delta_n\|(z_n - w_n) \\ &- (x^* - y^*)\|\|Az_n - Ax^*\| - (1 - \alpha_n)\|(w_n - u_n) + (x^* - y^*)\|^2 \\ &+ (1 - \alpha_n)2\varphi_n\|(w_n - u_n) - (x^* - y^*)\|\|Bw_n - By^*\| \\ &- (1 - \alpha_n)(1 - \beta_n)(\beta_n - \xi)\|Su_n - u_n\|^2\end{aligned} \tag{24}$$

which implies that

$$\begin{aligned}&(1 - \alpha_n)[\|(z_n - w_n) - (x^* - y^*)\|^2 + \|(w_n - u_n) + (x^* - y^*)\|^2] \\ &\leq [\|x_n - x^*\| + \|t_n - x^*\|]\|x_n - t_n\| - (1 - 2\lambda_n c_1)\|y_n - x_n\|^2 \\ &\quad - (1 - 2\lambda_n c_2)\|z_n - y_n\|^2 + 2\delta_n\|(z_n - w_n) - (x^* - y^*)\|\|Az_n - Ax^*\| \\ &\quad + 2\varphi_n\|(w_n - u_n) - (x^* - y^*)\|\|Bw_n - By^*\| \\ &\quad - (1 - \alpha_n)(1 - \beta_n)(\beta_n - \xi)\|Su_n - u_n\|^2.\end{aligned} \tag{25}$$

Since $\lim_{n\to\infty}\alpha_n = 0, 1-\alpha_n > 0, \|x_n - t_n\| \to 0, \|y_n - x_n\| \to 0, \|z_n - y_n\| \to 0, \|Az_n - Ax^*\| \to 0, \|Bw_n - By^*\| \to 0$ and $\|Su_n - u_n\| \to 0$, as $n \to \infty$. From $\{z_n\}$, $\{w_n\}$ and $\{u_n\}$ are bounded, we have

$$\lim_{n\to\infty} \|(z_n - w_n) - (x^* - y^*)\| = 0 \quad \text{and} \quad \lim_{n\to\infty} \|(w_n - u_n) + (x^* - y^*)\| = 0.$$

Consequently, it follows that $\lim_{n\to\infty}\|z_n - u_n\| = 0$ and $\lim_{n\to\infty}\|u_n - x_n\| = 0$. Hence implies that $\lim_{n\to\infty}\|Su_n - x_n\| = 0$ and $\lim_{n\to\infty}\|Sx_n - x_n\| = 0$

Step 6. We will show that $\tilde{x} \in \Omega$. First, we show that $\tilde{x} \in EP(F)$. Since $\{x_n\}$ is bounded, there exists a subsequence $\{x_{n_i}\}$ of $\{x_n\}$ which converges weakly to $\tilde{x}$ we can assume that $x_{n_i} \rightharpoonup \tilde{x}$ and $\lim_{n\to\infty}\|x_n - y_n\| = 0$, we have that $y_{n_i} \rightharpoonup \tilde{x}$. On the other hand, by using Proposition 7, we have, for every $y \in C$ and for every $i \in N$, that

$$\langle x_{n_i} - y_{n_i}, y - y_{n_i}\rangle \le \lambda_{n_i} F(x_{n_i}, y) - \lambda_{n_i} F(x_{n_i}, y_{n_i}). \tag{26}$$

Since $\|x_{n_i} - y_{n_i}\| \to 0$ and $y - y_{n_i} \rightharpoonup y - \tilde{x}$ as $i \to \infty$ and since $\forall i \in N, \quad 0 < \lambda_{\min} \le \lambda_{n_i} \le \lambda_{\max}$. As $i \to \infty$, we get

$$F(\tilde{x}, y) \ge 0, \quad \forall y \in C.$$

It means that $\tilde{x} \in EP(F)$.

Next, we show that $\tilde{x} \in Fix(S)$. Assume that $\tilde{x} \notin \Omega$. From Opial's condition Lemma 4 and $\lim_{n\to\infty}\|Sx_n - x_n\| = 0$, we obtain

$$\liminf_{n\to\infty} \|x_n - \tilde{x}\| < \liminf_{n\to\infty} \|x_n - S\tilde{x}\| \tag{27}$$

holds for every $y \in H$ with $y \ne \tilde{x}$. We observe that

$$\begin{aligned}
\liminf_{n\to\infty} \|x_{n_i} - \tilde{x}\| &< \liminf_{n\to\infty} \|x_{n_i} - S\tilde{x}\| \\
&= \liminf_{n\to\infty} \|x_{n_i} - Sx_{n_i} + Sx_{n_i} - S\tilde{x}\| \\
&\le \liminf_{n\to\infty} (\|x_{n_i} - Sx_{n_i}\| + \|Sx_{n_i} - S\tilde{x}\|) \\
&\le \liminf_{n\to\infty} \|x_{n_i} - \tilde{x}\|.
\end{aligned}$$

This is contradiction. Thus $\tilde{x} = S\tilde{x}$, it means $\tilde{x} \in Fix(S)$. Next, we will show that $\tilde{x} \in VI(A, C)$. Since $\|u_n - x_n\| \to 0$, $\|w_n - x_n\| \to 0$ and $\|u_n - w_n\| \to 0$ as $n \to \infty$, we deduce that $x_{n_i} \rightharpoonup \tilde{x}$ and $w_{n_i} \rightharpoonup \tilde{x}$ Let

$$Tv = \begin{cases} Av + N_C v, & v \in C, \\ \emptyset, & v \notin C, \end{cases} \tag{28}$$

where $N_C v$ is the normal cone to C at $v \in C$. In this case, the mapping T is maximal monotone. Let $(v, x) \in T$. Then, we have $x \in Tv = Av + N_C v$ and hence $x - Av \in$

$N_C v$. So, we have $\langle v - u, x - Av\rangle \geq 0$, for all $u \in C$. On the other hand, from $w_n = P_C(z_n - \lambda_n A z_n)$ and $v \in C$, we have

$$\langle z_n - \lambda_n A z_n - w_n, w_n - v\rangle \geq 0$$
$$\langle v - w_n, \frac{w_n - z_n}{\lambda_n} + A z_n\rangle \geq 0.$$

From $\langle v - u, x - Av\rangle \geq 0$, for all $u \in C$ and $w_{n_i} \in C$, we have

$$\begin{aligned}\langle v - w_{n_i}, x\rangle &\geq \langle v - w_{n_i}, Av\rangle \geq \langle v - w_{n_i}, Av\rangle - \Big\langle v - w_{n_i}, \frac{w_{n_i} - z_{n_i}}{\lambda_{n_i}} + A z_{n_i}\Big\rangle \\ &= \langle v - w_{n_i}, Av - Aw_{n_i}\rangle + \langle v - w_{n_i}, Aw_{n_i} - Az_{n_i}\rangle \\ &\quad - \Big\langle v - w_{n_i}, \frac{w_{n_i} - z_{n_i}}{\lambda_{n_i}}\Big\rangle \\ &\geq \langle v - w_{n_i}, Aw_{n_i} - Az_{n_i}\rangle - \Big\langle v - w_{n_i}, \frac{w_{n_i} - z_{n_i}}{\lambda_{n_i}}\Big\rangle.\end{aligned}$$

Hence, we obtain $\langle v - \tilde{x}, x\rangle \geq 0$ as $i \to \infty$. Since T is maximal monotone, we have $\tilde{x} \in T^{-1}$ hence $\tilde{x} \in VI(A, C)$. Similarly way, we can prove $\tilde{x} \in VI(B, C)$. It implies that $\tilde{x} \in EP(F) \cap Fix(S) \cap VI(A, C) \cap VI(B, C)$. Therefore, $\tilde{x} \in \Omega$.

Step 7. Finally, we show that $\lim_{n\to\infty} x_n = \tilde{x}$, where $\tilde{x} = P_\Omega x_0$. Since Ω is nonempty closed convex subset of H, there exists a unique $x^* \in \Omega$ such that $x^* = P_\Omega x_0 \in C_n \cap D_n$ and $x_n = P_{D_n} x_0$, we have

$$\|x_n - x_0\| \leq \|x^* - x_0\|. \tag{29}$$

Let $\{x_{n_i}\}$ be subsequence of $\{x_n\}$ such that $x_{n_i} \rightharpoonup \tilde{x}$. It follows from $x^* = P_\Omega x_0$ and the lower semicontinuity of norm that

$$\|x^* - x_0\| \leq \|\tilde{x} - x_0\| \leq \liminf_{i\to\infty} \|x_{n_i} - x_0\| \leq \limsup_{i\to\infty} \|x_{n_i} - x_0\| \leq \|x^* - x_0\|. \tag{30}$$

Thus, we obtain that $\lim_{i\to\infty} \|x_{n_i} - x_0\| = \|\tilde{x} - x_0\| = \|x^* - x_0\|$. By Lemma 5 Kadec-Klee property of H, we obtain that

$$\lim_{i\to\infty} x_{n_i} = \tilde{x} = x^*. \tag{31}$$

Since $\{x_{n_i}\}$ is an arbitrary subsequence of $\{x_n\}$, we can conclude that $\{x_n\}$ converges strongly to $P_\Omega x$. □

Corollary 10 *Let C be a nonempty, closed and convex subset of a real Hilbert space H, let f be a bifunction $f : C \times C$ into R satisfying conditions (A1)–(A5), let A be an α-inverse-strongly monotone mapping of C into H and B be a β-inverse-strongly monotone mapping of C into H. Let S be a nonexpansive mapping from C to C and such that $\Omega := EP(F) \cap Fix(S) \cap VI(A, C) \cap VI(B, C) \neq \emptyset$. Let*

$\{x_n\}, \{y_n\}, \{z_n\}, \{u_n\}$ and $\{w_n\}$ be sequences generated by $x_0 \in C$. For every $n \in N$, where $\{\alpha_n\} \subset [0, 1)$ $\{\lambda_n\} \subset (0, 1]$, $\{\delta_n\} \subset (0, 2\alpha)$, and $\{\varphi_n\} \subset (0, 2\beta)$, following condition:

(1) $\{\lambda_n\} \subset [\lambda_{\min}, \lambda_{\max}]$, *where* $0 < \lambda_{\min} \leq \lambda_{\max} < \min\left\{\frac{1}{2c_1}, \frac{1}{2c_2}\right\}$,
(2) $\{\delta_n\} \subset [a, b]$ *for some* a, b *with* $0 < a < b < 2\alpha$,
(3) $\{\varphi_n\} \subset [c, d]$ *for some* d, e *with* $0 < c < d < 2\beta$,
(4) $\{\beta_n\} \subset [e, f]$ *for some* $0 \leq \xi < e \leq f < 1$, *and* $\lim_{n\to\infty} \alpha_n = 0$,

Then the sequence $\{x_n\}$ generated by Algorithm 9 converges strongly to the projection of x_0 on to the set Ω.

Proof Setting $\xi = 0$ in Theorem 1, we have the result.

If $A \equiv 0$ and $B \equiv 0$ in Algorithm 3 then $P_C \equiv I$ such that $u_n = w_n = P_C z_n = z_n$ by Theorem 1 we obtain the following Corollary 11.

Corollary 11 *Let C be a nonempty, closed and convex subset of a real Hilbert space H, let f be a bifunction $f : C \times C$ into R satisfying conditions $(A1) - (A5)$. Let S be a ξ-strict pseudocontraction mapping from C to C and such that $\Omega := EP(F) \cap Fix(S) \neq \emptyset$. Let $\{x_n\}, \{y_n\}, \{z_n\}$ be sequences generated by $x_0 \in C$ and*

$$\begin{cases} y_n = \arg\min\limits_{y\in C}\{\lambda_n F(x_n, y) + \dfrac{1}{2}\|y - x_n\|^2\}, \\ z_n = \arg\min\limits_{y\in C}\{\lambda_n F(y_n, y) + \dfrac{1}{2}\|y - x_n\|^2\}, \\ t_n = \alpha_n x_n + (1 - \alpha_n)[\beta_n S u_n + (1 - \beta_n)u_n], \\ C_n = \{z \in C : \|t_n - z\| \leq \|x_n - z\|\}, \\ D_n = \{z \in C : \langle x_n - z, x_0 - x\rangle \geq 0\}, \\ x_{n+1} = P_{C_n \cap D_n} x_0, \end{cases} \tag{32}$$

For every $n \in N$, where $\{\alpha_n\} \subset [0, 1)$, $\{\lambda_n\} \subset (0, 1]$, $\{\beta_n\} \subset [0, 1)$. Following Algorithm 12

Algorithm 12 Choose the sequences $\{\alpha_n\} \subset [0, 1)$, $\{\beta_n\} \subset (0, 1)$, $\{\lambda_n\} \subset (0, 1]$ and ξ be a constant in $[0, 1)$.

Step (1) Let $x_0 \in C$. Set $n = 0$
Step (2) Solve successively the strongly convex programs $\arg\min_{y\in C}\{\lambda_n F(x_n, y) + \frac{1}{2}\|y - x_n\|^2\}$ and $\arg\min_{y\in C}\{\lambda_n F(y_n, y) + \frac{1}{2}\|y - x_n\|^2\}$ to obtain the unique optimal solution y_n and z_n, respectively,
Step (3) Compute $t_n = \alpha_n x_n + (1 - \alpha_n)[\beta_n S u_n + (1 - \beta_n)u_n]$. If $y_n = x_n$ and $t_n = x_n$, then STOP: $x_n \in EP(F) \cap Fix(S) \cap VI(A, C) \cap VI(B, C)$. Otherwise, go to Step (4).
Step (4) Compute $x_{n+1} = P_{C_n \cap D_n} x_0$, where $C_n = \{z \in C : \|t_n - z\| \leq \|x_n - z\|\}$, $D_n = \{z \in C : \langle x_n - z, x_0 - x\rangle \geq 0\}$,
Step (5) Set $n := n + 1$, and go to Step (2).

Then the sequences $\{x_n\}$ *generated by Algorithm* 12 *converges strongly to the projection of* x_0 *on to the set* Ω.

Proof Setting $z_n = w_n = u_n$ in Theorem 1, we have the following result.

Acknowledgments The authors were supported by the Higher Education Research Promotion and National Research University Project of Thailand, Office of the Higher Education Commission (NRU-CSEC No. 55000613).

References

1. Ceng L-C, Al-Homidan S, Ansari QH, Yao J-C (2009) An iterative scheme for equilibrium problems and fixed point problems of strict pseudo-contraction mappings. J Comput Appl Math 223(2):967–974
2. Fan K (1972) A minimax inequality and applications. In: Shisha O (ed) Inequality III. Academic Press, New York, pp 103–113
3. Blum E, Oettli W (1994) From optimization and variational inequalities to equilibrium problems. Math Student 63:123–145
4. Flam SD, Antipin AS (1997) Equilibrium progamming using proximal-link algorithms. Math Program 78:29–41
5. Moudafi A, Thera M (1999) Proximal and dynamical approaches to equilibrium problems. In: Lecture note in economics and mathematical systems, vol 477. Springer, New York, pp 187–201.
6. Takahashi S, Takahashi W (2007) Viscosity approximation methods for equilibrium problems and fixed point problems in Hilbert spaces. J Math Anal Appl 331:506–515
7. Scherzer O (1995) Convergence criteria of iterative methods based on Landweber iteration for solving nonlinear problems. J Mathe Anal Appl 194(3):911–933
8. Mann WR (1953) Mean value methods in iteration. Proc Amer Math Soc 4:506–510
9. Reich S (1979) Weak convergence theorems for nonexpansive mappings. J Math Anal Appl 67:274–276
10. Genel A, Lindenstrass J (1975) An example concerning fixed points. Isael J Math 22:81–86
11. Nakajo K, Takahashi W (2003) Strong convergence theorems for nonexpansive mappings and nonexpansive semigroups. J Math Anal Appl 279:372–379
12. Takahashi W, Toyoda M (2003) Weak convergence theorems for nonexpansive mappings and monotone mappings. J Optim Theory Appl 118:417–428
13. Ceng LC, Ansari QH, Yao JC (2011) Relaxed extragradient iterative methods for variational inequalities. Appl Math Comput 218:1112–1123
14. Vuong PT, Strodiot JJ, Nguyen VH (2012) Extragradient methods and linesearch algorithms for solving Ky Fan inequalities and fixed point problems. J Optim Theory Appl 1–23.
15. Phiangsungnoen S, Kumam P (2013) A hybrid extragradient method for solving Ky Fan inequalities, variational inequalities and fixed point problems. In: Proceedings of the international multiConference of engineers and computer scientists, vol II, IMECS 2013, March 13–15, 2013. Hong Kong, pp 1042–1047.
16. Rockafellar RT (1970) On the maximality of sums of nonlinear monotone operators. Trans Amer Math Sco 149:75–88
17. Rockafellar RT (1976) Monotone operators and proximal point algorithm. SIAM J Control Optim 14:877–898
18. Opial Z (1967) Weak convergence of successive approximations for nonexpansive mappings. Bull Amer Math Soc 73:591–597
19. Goebel K, Kirk WA (1990) Topics in metric fixed point theory. Cambridge University Press, Cambridge

20. Takahashi W (2000) Nonlinear functional analysis. Yokohama Publishers, Yokohama
21. Mastroeni G (2003) On auxiliary principle for equilibrium problems. In: Daniele P, Gianessi F, Maugeri A (eds) Equilibrium problems and variational models. Kluwer Academic, Dordrecht, pp 289–298
22. Tran DQ, Muu LD, Nguyen VH (2008) Extragradient algorithms extended to equilibrium problems. Optimization 57:749–776
23. Anh PN (2011) A hybrid extragradient method extended to fixed point problems and equilibrium problem. Optimization. doi:url10.1080/02331934.2011.607497
24. Jaiboon C, Kumam P (2010) Strong convergence theorems for solving equilibrium problems and fixed point problems of ξ-strict pseudo-contraction mappings by two hybrid projection methods. J Comput Appl Math 230:722–732

Random Fuzzy Multiobjective Linear Programming with Variance Covariance Matrices

Hitoshi Yano and Kota Matsui

Abstract In this paper, we propose an interactive decision making method for random fuzzy multiobjective linear programming problems with variance covariance matrices (RFMOLP-VC). In the proposed method, it is assumed that the decision maker has fuzzy goals for not only permissible objective levels of a probability maximization model for RFMOLP-VC but also the corresponding distribution function values. Using the fuzzy decision, such two kinds of membership functions are integrated. In the integrated membership space, a satisfactory solution is obtained from among a Pareto optimal solution set through the interaction with the decision maker.

Keywords Interactive method · Multiobjective linear programming · Pareto optimality · Random fuzzy variables · Satisfactory solution · Variance covariance matrices

1 Introduction

In the real world decision making situations, we often have to make a decision under uncertainty. To deal with decision problems involving uncertainty, stochastic programming approaches [1–3, 10] and fuzzy programming approaches [16, 20, 28] have been developed. Recently, in order to deal with mathematical programming problems involving the randomness and the fuzziness, random fuzzy programming

H. Yano (✉)
Graduate School of Humanities and Social Sciences, Nagoya City University, Nagoya 467-8501, Japan
e-mail: yano@hum.nagoya-cu.ac.jp

K. Matsui
Graduate School of Information Science,
Nagoya University, Nagoya 4464-8601, Japan
e-mail: matsui@math.cm.is.nagoya-u.ac.jp

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 275, DOI: 10.1007/978-94-007-7684-5_27,

has been developed [11], in which the coefficients of the objective functions and/or the constraints are represented with random fuzzy variables [17, 18]. As a natural extension, a random fuzzy multiobjective programming problem (RFMOLP) was formulated and the interactive decision making methods were proposed to obtain a satisfactory solution of the decision maker from among a Pareto optimal solution set [12–15]. Moreover, in order to show the efficiency of random fuzzy programming techniques, real-world decision making problems under random fuzzy environments were formulated as random fuzzy programming problems, and the corresponding algorithms to obtain the optimal solutions were proposed [6, 8, 19, 22].

Under these circumstances, we focus on the interactive decision making methods [11–13] for RFMOLP to obtain a satisfactory solution, in which a probability maximization model or a fractile optimization model is adopted in order to deal with RFMOLP. In their proposed methods, it seems to be very difficult for the decision maker to specify in advance permissible objective levels or permissible probability levels appropriately. From such a point of view, under the assumption that the decision maker has fuzzy goals for permissible objective levels and the corresponding distribution functions for a probability maximization model, we have proposed an interactive decision making method for RFMOLP to obtain a satisfactory solution of the decision maker [24]. The proposed method can be regarded as a random fuzzy version of the interactive decision making methods for fuzzy random multiobjective programming problems [23, 26, 27]. However, in the proposed method [24], covariance between random fuzzy variables in the objective functions cannot be treated. In this paper, we formulate a random fuzzy multiobjective programming problem with variance–covariance matrices (RFMOLP-VC), and propose an interactive decision making method for RFMOLP-VC to obtain a satisfactory solution of the decision maker. In Sect. 2, RFMOLP-VC is formulated by using a concept of a possibility measure and a probability maximization model, and the D-Pareto optimal solution concept for RFMOLP-VC is introduced. For the reference membership values specified by the decision maker, the corresponding D-Pareto optimal solution is obtained by solving the minmax problem. It is shown that the optimal solution of the minimax problem can be easily obtained by the convex programming technique. In Sect. 3, the interactive algorithm to obtain the satisfactory solution from among a D-Pareto optimal solution set is proposed, which is based on the convex programming technique. In Sect. 4, In order to illustrate the proposed method, a crop planning problem at farm level [5, 7, 9, 21] is formulated as a numerical example, and the interactive processes under the hypothetical decision maker are demonstrated. Finally, in Sect. 5, we conclude this paper.

2 Problem Formulation

In this section, we focus on RFMOLP-VC in which random fuzzy variable coefficients are involved in objective functions.

[RFMOLP-VC]

$$\min_{x \in X} \overline{\tilde{C}} x = (\overline{\tilde{C}}_1 x, \ldots, \overline{\tilde{C}}_k x)$$

where $x = (x_1, x_2, \ldots, x_n)^T$ is an n dimensional decision variable column vector, $\overline{\tilde{C}}_i = (\overline{\tilde{C}}_{i1}, \ldots, \overline{\tilde{C}}_{in}), i = 1, \cdots, k$, are coefficient vectors of objective function $\overline{\tilde{C}}_i x$, whose elements are random fuzzy variables [17], and the symbols "-" and "$\sim$" mean randomness and fuzziness respectively.

In this paper, according to Katagiri et al. [11–13], we assume that a random fuzzy variable $\overline{\tilde{C}}_{ij}$ is normally distributed with the fuzzy number $\tilde{M}_{ij}$ as mean, and the positive-definite variance–covariance matrices $V_i, i = 1, \ldots, k$ between random fuzzy variables $\overline{\tilde{C}}_{ij_1}$ and $\overline{\tilde{C}}_{ij_2}$, $j_1, j_2 = 1, \cdots, n$ are given as:

$$V_i = \begin{pmatrix} \sigma_{i11}\sigma_{i12} \ldots \sigma_{i1n} \\ \sigma_{i21}\sigma_{i22} \ldots \sigma_{i2n} \\ \ldots\ldots\ldots\ldots \\ \sigma_{in1}\sigma_{in2} \ldots \sigma_{inn} \end{pmatrix}, \quad i = 1, \ldots, k. \tag{1}$$

As a result, we assume that a probability density function $f_{ij}(y)$ for a random fuzzy variable $\overline{\tilde{C}}_{ij}$ is formally represented with the following form.

$$f_{ij}(y) = \frac{1}{\sqrt{2\pi}\sigma_{ijj}} e^{-\frac{(y-\tilde{M}_{ij})^2}{2\sigma_{ijj}}}, \quad 1 \le i \le k, 1 \le j \le n \tag{2}$$

where $\tilde{M}_{ij}$ is an L-R fuzzy number characterized by the following membership function.

$$\mu_{\tilde{M}_{ij}}(t) = \begin{cases} L\left(\frac{m_{ij}-t}{\alpha_{ij}}\right), & m_{ij} \ge t \\ R\left(\frac{t-m_{ij}}{\beta_{ij}}\right), & m_{ij} \le t \end{cases} \tag{3}$$

L and R are called reference functions, m_{ij} is the mean value, and α_{ij}, β_{ij} are spread parameters [4].

Then, a random fuzzy variable $\overline{\tilde{C}}_{ij}$ can be characterized by the following membership function [11–13].

$$\mu_{\overline{\tilde{C}}_{ij}}(\overline{\gamma}_{ij}) = \sup_{s_{ij}} \{\mu_{\tilde{M}_{ij}}(s_{ij}) | \overline{\gamma}_{ij} \sim N(s_{ij}, \sigma_{ijj})\} \tag{4}$$

where $N(s_{ij}, \sigma_{ijj})$ means a normal distribution with mean s_{ij} and standard deviation $\sqrt{\sigma_{ijj}}$. Moreover, using Zadeh's extension principle [20, 28], the objective function $\overline{\tilde{C}}_i x$ become a random fuzzy variable characterized by the following membership function [11–13].

$$\mu_{\overline{\tilde{C}}_i x}(\bar{u}_i) = \sup_{(s_{i1},\ldots,s_{in})\in \mathrm{R}^n} \left\{ \min_{1\le j\le n} \mu_{\tilde{M}_{ij}}(s_{ij})\bar{u}_i \sim N\left(\sum_{j=1}^{n} s_{ij}x_j, x^T V_i x\right)\right\} \tag{5}$$

Unfortunately, we can not treat RFMOLP-VC directly because it is ill-defined. Katagiri et al. [11, 13] formulated RFMOLP-VC using permissible objective levels of a probability maximization model and the possibility measure. For permissible objective levels $f_i, i = 1, \ldots, k$ specified by the decision maker, a probability maximization model for RFMOLP-VC can be formulated as follows.

[**MOP1**(f)]

$$\max_{x\in X}(\Pr(\omega|\overline{\tilde{C}}_1(\omega)x \le f_1), \ldots, \Pr(\omega|\overline{\tilde{C}}_k(\omega)x \le f_k))$$

It should be noted here that the each objective function:

$$\tilde{P}_i(x, f_i) \stackrel{\text{def}}{=} \Pr(\omega|\overline{\tilde{C}}_i(\omega)x \le f_i), \tag{6}$$

becomes a fuzzy set and the corresponding membership function is defined as follows [11–13].

$$\mu_{\tilde{P}_i(x,f_i)}(p_i) \stackrel{\text{def}}{=} \sup_{\bar{u}_i} \left\{ \mu_{\overline{\tilde{C}}_i x}(\bar{u}_i) p_i = \Pr(\omega|\bar{u}_i(\omega) \le f_i), \quad \bar{u}_i \sim N\left(\sum_{j=1}^{n} s_{ij}x_j, x^T V_i x\right)\right\} \tag{7}$$

Katagiri et al. [11–13] showed that, from (5), the membership function (7) can be transformed as follows.

$$\mu_{\tilde{P}_i(x,f_i)}(p_i) = \sup_{(s_{i1},\ldots,s_{in})\in \mathrm{R}^n} \left\{ \min_{1\le j\le n} \mu_{\tilde{M}_{ij}}(s_{ij}) p_i = \Pr(\omega|\bar{u}_i(\omega) \le f_i), \quad \bar{u}_i \sim N\left(\sum_{j=1}^{n} s_{ij}x_j, x^T V_i x\right)\right\} \tag{8}$$

Using (6), MOP1(f) can be transformed as follows.

[**MOP2** (f)]

$$\max_{x\in X}(\tilde{P}_1(x, f_1), \ldots, \tilde{P}_k(x, f_k))$$

MOP2(f) is ill-defined yet, because objective functions of MOP2(f) are fuzzy sets depending on permissible objective level $f_i, i = 1, \ldots, k$. In order to deal with MOP2(f), let us assume that the decision maker has a fuzzy goal $\tilde{G}_i$ for each

objective function $\tilde{P}_i(x, f_i)$, which is expressed in words such as "$\tilde{P}_i(x, f_i)$ should be substantially less than p_i" For the corresponding membership function $\mu_{\tilde{G}_i}(p_i)$, we make the following assumption.

Assumption 1 $\mu_{\tilde{G}_i}(p_i), i = 1, \ldots, k$ are strictly increasing and continuous with respect to $p_i \in [p_{i\min}, p_{i\max}]$, and $\mu_{p_i}(p_{i\min}) = 0$, $\mu_{p_i}(p_{i\max}) = 1$, where $0.5 < p_{i\min}$ is a maximum value of an unacceptable levels and $p_{i\max} < 1$ is a minimum value of a sufficiently satisfactory levels.

Using the following possibility measure [4],

$$\Pi_{\tilde{P}_i(x,f_i)}(\tilde{G}_i) \stackrel{\text{def}}{=} \sup_{p_i} \min\{\mu_{\tilde{P}_i(x,f_i)}(p_i), \mu_{\tilde{G}_i}(p_i)\}, \tag{9}$$

Katagiri et al. [11–13] transformed MOP2(f) into the following well-defined multiobjective programming problem.

[MOP3(f)]

$$\max_{x \in X}(\Pi_{\tilde{P}_1(x,f_1)}(\tilde{G}_1), \ldots, \Pi_{\tilde{P}_k(x,f_k)}(\tilde{G}_k))$$

Unfortunately, in MOP3(f), the decision maker must specify permissible objective levels in advance. However, it seems very difficult to specify such values because $\Pi_{\tilde{P}_i(x,f_i)}(\tilde{G}_i)$ depends on a permissible objective level f_i. From such a point of view, in this paper, instead of MOP3(f), we consider the following extended problem where $f_i, i = 1, \ldots, k$ are not constants but decision variables.

[MOP4]

$$\max_{x \in X, f_i \in \mathrm{R}^1, i=1,\ldots,k}(\Pi_{\tilde{P}_1(x,f_1)}(\tilde{G}_1), \ldots, \Pi_{\tilde{P}_k(x,f_k)}(\tilde{G}_k), -f_1, \ldots, -f_k)$$

Considering the imprecise nature of the decision maker's judgment, we assume that the decision maker has a fuzzy goal for each permissible objective level. Such a fuzzy goal can be quantified by eliciting the corresponding membership function. Let us denote a membership function of a permissible objective level f_i as $\mu_{\tilde{F}_i}(f_i)$. For the membership function $\mu_{\tilde{F}_i}(f_i)$, we make the following assumption.

Assumption 2 $\mu_{\tilde{F}_i}(f_i), i = 1, \cdots, k$ are strictly decreasing and continuous with respect to $f_i \in [f_{i\min}, f_{i\max}]$, and $\mu_{f_i}(f_{i\min}) = 1$, $\mu_{f_i}(f_{i\max}) = 0$, where $f_{i\min}$ is a maximum value of a sufficiently satisfactory levels. and $f_{i\max}$ is a minimum value of an unacceptable levels.

Then, MOP4 can be transformed as the following multiobjective programming problem.

[MOP5]

$$\max_{x \in X, f_i \in \mathrm{R}^1, i=1,\ldots,k}(\Pi_{\tilde{P}_1(x,f_1)}(\tilde{G}_1), \ldots, \Pi_{\tilde{P}_k(x,f_k)}(\tilde{G}_k), \mu_{\tilde{F}_1}(f_1), \ldots, \mu_{\tilde{F}_k}(f_k))$$

It should be noted here that, $\Pi_{\tilde{P}_i(x,f_i)}(\tilde{G}_i)$ is strictly increasing with respect to f_i. If the decision maker adopts the fuzzy decision [20, 28] to integrate $\Pi_{\tilde{P}_i(x,f_i)}(\tilde{G}_i)$ and $\mu_{\tilde{F}_i}(f_i)$, MOP5 can be transformed into the following form.

[MOP6]

$$\max_{x\in X, f_i\in \mathrm{R}^1, i=1,\ldots,k} (\mu_{D_1}(x, f_1), \ldots, \mu_{D_k}(x, f_k))$$

where $\mu_{D_i}(x, f_i) = \min\{\Pi_{\tilde{P}_i(x,f_i)}(\tilde{G}_i), \mu_{\tilde{F}_i}(f_i)\}$. In order to deal with MOP6, we introduce a D-Pareto optimal solution concept.

Definition 1 $x^* \in X, f_i^* \in \mathrm{R}^1, i = 1, \ldots, k$ is said to be a D-Pareto optimal solution to MOP6, if and only if there does not exist another $x \in X, f_i \in \mathrm{R}^1, i = 1, \ldots, k$ such that $\mu_{D_i}(x, f_i) \geq \mu_{D_i}(x^*, f_i^*) \quad i = 1, \ldots, k$ with strict inequality holding for at least one i.

For generating a candidate of a satisfactory solution which is also D-Pareto optimal, the decision maker is asked to specify the reference membership values [20]. Once the reference membership values $\hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_k)$ are specified, the corresponding D-Pareto optimal solution is obtained by solving the following minmax problem.

[**MINMAX1**($\hat{\mu}$)]

$$\min_{x\in X, f_i\in \mathrm{R}^1, i=1,\ldots,k, \lambda\in\lambda} \lambda \tag{10}$$

subject to

$$\hat{\mu}_i - \Pi_{\tilde{P}_i(x,f_i)}(\tilde{G}_i) \leq \lambda, i = 1, \ldots, k \tag{11}$$

$$\hat{\mu}_i - \mu_{\tilde{F}_i}(f_i) \leq \lambda, i = 1, \ldots, k \tag{12}$$

From [11–13], each constraint of (11) can be equivalently transformed into the following form.

$$\begin{aligned} &\hat{\mu}_i - \Pi_{\tilde{P}_i(x,f_i)}(\tilde{G}_i) \leq \lambda \\ &\Leftrightarrow \sum_{j=1}^{n} \{m_{ij} - L^{-1}(\hat{\mu}_i - \lambda)\alpha_{ij}\}x_j \\ &+ \Phi^{-1}(\mu_{\tilde{G}_i}^{-1}(\hat{\mu}_i - \lambda))\sqrt{x^T V_i x} \leq f_i \end{aligned} \tag{13}$$

where $\Phi(\cdot)$ is a distribution function of the standard Gaussian random variable, $\Phi^{-1}(\cdot)$ is a corresponding inverse function, and $L^{-1}(\cdot)$, $\mu_{\tilde{G}_l}^{-1}(\cdot)$ are pseudo-inverse functions of $L(\cdot)$, $\mu_{\tilde{G}_i}(\cdot)$ respectively. Moreover, since the inequalities (12) can be transformed into $f_i \leq \mu_{\tilde{F}_i}^{-1}(\hat{\mu}_i - \lambda)$, MINMAX1($\hat{\mu}$) can be reduced to the following problem.

[**MINMAX2($\hat{\mu}$)]**

$$\min_{x\in X,\lambda\in\Lambda} \lambda \tag{14}$$

subject to

$$\sum_{j=1}^{n}\{m_{ij}-L^{-1}(\hat{\mu}_i-\lambda)\alpha_{ij}\}x_j+\Phi^{-1}(\mu_{\tilde{G}_i}^{-1}(\hat{\mu}_i-\lambda))\sqrt{x^T V_i x} \le \mu_{\tilde{F}_i}^{-1}(\hat{\mu}_i-\lambda), \quad i=1,\ldots,k \tag{15}$$

where

$$\Lambda \stackrel{\text{def}}{=} [\lambda_{\min},\lambda_{\max}] = [\max_{i=1,\ldots,k}\hat{\mu}_i - 1, \min_{i=1,\ldots,k}\hat{\mu}_i]. \tag{16}$$

The relationships between the optimal solution (x^*, λ^*) of MINMAX2($\hat{\mu}$) and D-Pareto optimal solutions can be characterized by the following theorem.

Theorem 1

(1) If $x^* \in X, \lambda^* \in \Lambda$ is a unique optimal solution of MINMAX2($\hat{\mu}$), then $x^* \in X, \mu_{\tilde{F}_i}^{-1}(\hat{\mu}_i - \lambda^*) \in \mathrm{R}^1, i = 1,\ldots,k$ is a D-Pareto optimal solution.

(2) If $x^* \in X, f_i^* \in \mathrm{R}^1, i = 1,\ldots,k$ is a D-Pareto optimal solution, then $x^* \in X, \quad \lambda^* = \hat{\mu}_i - \Pi_{\tilde{P}_i(x^*,f_i^*)}(\tilde{G}_i) = \hat{\mu}_i - \mu_{\tilde{F}_i}(f_i^*), i = 1,\ldots,k$ is an optimal solution of MINMAX2($\hat{\mu}$) for some reference membership values $\hat{\mu} = (\hat{\mu}_1,\ldots,\hat{\mu}_k)$.

Proof

(1) Assume that $x^* \in X, f_i^* \stackrel{\text{def}}{=} \mu_{\tilde{F}_i}^{-1}(\hat{\mu}_i - \lambda^*), i = 1,\ldots,k$ is not a D-Pareto optimal solution. Then, from (11), there exist $x \in X, f_i \in \mathrm{R}^1, i = 1,\ldots,k$ such that

$$\begin{aligned}\mu_{D_i}(x, f_i) &= \min\{\Pi_{\tilde{P}_i(x,f_i)}(\tilde{G}_i), \mu_{\tilde{F}_i}(f_i)\} \ge \mu_{D_i}(x^*, f_i^*)\\ &= \min\{\Pi_{\tilde{P}_i(x^*,f_i^*)}(\tilde{G}_i), \mu_{\tilde{F}_i}(f_i^*)\} = \hat{\mu}_i - \lambda^*,\\ i &= 1,\ldots,k,\end{aligned}$$

with strict inequality holding for at least one i. Then it holds that

$$\Pi_{\tilde{P}_i(x,f_i)}(\tilde{G}_i) \ge \hat{\mu}_i - \lambda^*, i = 1,\ldots,k, \tag{17}$$

$$\mu_{\tilde{F}_i}(f_i) \ge \hat{\mu}_i - \lambda^*, i = 1,\ldots,k. \tag{18}$$

From (13), (17) can be transformed as follows.

$$\sum_{j=1}^{n}\{m_{ij} - L^{-1}(\hat{\mu}_i - \lambda^*)\alpha_{ij}\}x_j + \Phi^{-1}(\mu_{\tilde{G}_i}^{-1}(\hat{\mu}_i - \lambda^*))\sqrt{x^T V_i x} \le f_i \tag{19}$$

From (18), it holds that $f_i \leq \mu_{\tilde{F}_i}^{-1}(\hat{\mu}_i - \lambda^*)$. As a result, there exists $x \in X$ such that

$$\sum_{j=1}^{n}\{m_{ij} - L^{-1}(\hat{\mu}_i - \lambda^*)\alpha_{ij}\}x_j + \Phi^{-1}(\mu_{\tilde{G}_i}^{-1}(\hat{\mu}_i - \lambda^*))\sqrt{x^T V_i x} \leq \mu_{\tilde{F}_i}^{-1}(\hat{\mu}_i - \lambda^*), \quad i = 1, \ldots, k, \tag{20}$$

which contradicts the fact that $x^* \in X, \lambda^* \in \Lambda$ is a unique optimal solution to MINMAX2($\hat{\mu}$).

(2) Assume that $x^* \in X, \lambda^* \in \Lambda$ is not an optimal solution to MINMAX2($\hat{\mu}$) for any reference membership values $\hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_k)$, which satisfy the equalities :

$$\hat{\mu}_i - \lambda^* = \Pi_{\tilde{P}_i(x^*, f_i^*)}\left(\tilde{G}_i\right) = \mu_{\tilde{F}_i}\left(f_i^*\right), \quad i = 1, \ldots, k. \tag{21}$$

Then, there exists some $x \in X, \lambda < \lambda^*$ such that

$$\sum_{j=1}^{n}\{m_{ij} - L^{-1}(\hat{\mu}_i - \lambda)\alpha_{ij}\}x_j + \Phi^{-1}(\mu_{\tilde{G}_i}^{-1}(\hat{\mu}_i - \lambda))\sqrt{x^T V_i x} \leq \mu_{\tilde{F}_i}^{-1}(\hat{\mu}_i - \lambda), \quad i = 1, \ldots, k. \tag{22}$$

This means that

$$\begin{aligned} \Pi_{\tilde{P}_i(x, f_i)}(\tilde{G}_i) \geq &\quad \hat{\mu}_i - \lambda > \hat{\mu}_i - \lambda^*, i = 1, \ldots, k \\ \mu_{\tilde{F}_i}(f_i) = &\quad \hat{\mu}_i - \lambda > \hat{\mu}_i - \lambda^*, i = 1, \ldots, k, \end{aligned}$$

where $f_i \stackrel{\text{def}}{=} \mu_{\tilde{F}_i}^{-1}(\hat{\mu}_i - \lambda)$. From (21), there exists $x \in X$, $f_i \in \mathrm{R}^1, i = 1, \ldots, k$ such that $\mu_{D_i}(x, f_i) > \mu_{D_i}(x^*, f_i^*), i = 1, \ldots, k$. This contradicts the fact that $x^* \in X$, $f_i^* \in \mathrm{R}^1, i = 1, \cdots, k$ is a D-Pareto optimal solution.

Since the constraints (15) are nonlinear, it is not easy to solve MINMAX2($\hat{\mu}$) directly. Before considering the algorithm to solve MINMAX2($\hat{\mu}$), we first define the following functions corresponding to (15).

$$g_i(x, \lambda) \stackrel{\text{def}}{=} \mu_{\tilde{F}_i}^{-1}(\hat{\mu}_i - \lambda) - \sum_{j=1}^{n}\{m_{ij} - L^{-1}(\hat{\mu}_i - \lambda)\alpha_{ij}\}x_j - \Phi^{-1}(\mu_{\tilde{G}_i}^{-1}(\hat{\mu}_i - \lambda))\sqrt{x^T V_i x}, \quad i = 1, \ldots, k \tag{23}$$

Because of $\mu_{\tilde{G}_i}^{-1}(\hat{\mu}_i - \lambda) > 0.5$ for any $\lambda \in \Lambda$, it holds that $\Phi^{-1}(\mu_{p_i}^{-1}(\hat{\mu}_i - \lambda)) > 0$. This means that $g_i(x, \lambda), i = 1, \cdots, k$ are concave with respect to $x \in X$ for any fixed value $\lambda \in \Lambda$. Let us define the following feasible set $X(\lambda)$ of MINMAX2($\hat{\mu}$) for some fixed value $\lambda \in \Lambda$.

$$X(\lambda) \stackrel{\text{def}}{=} \{x \in X | g_i(x, \lambda) \geq 0, i = 1, \ldots, k\} \tag{24}$$

Then, it is clear that $X(\lambda)$ is a convex set. $X(\lambda)$ satisfies the following property.

Property 1

If $\lambda_{\min} \leq \lambda_1 \leq \lambda_2 \leq \lambda_{\max}$, then it holds that $X(\lambda_1) \subset X(\lambda_2)$.

In the following, it is assumed that $X(\lambda_{\min}) = \phi$, $X(\lambda_{\max}) \neq \phi$. From Property 1, we can obtain the optimal solution (x^*, λ^*) of MINMAX2($\hat{\mu}$) using the following simple algorithm which is based on the bisection method and the convex programming technique.

[Algorithm 1]

Step 1: Set $\lambda_0 = \lambda_{\min}, \lambda_1 = \lambda_{\max}, \lambda \leftarrow (\lambda_0 + \lambda_1)/2$.

Step 2: Solve the following convex programming problem for the fixed value λ, and denote the optimal solution as $x(\lambda)$.

$$\max_{x \in X} g_j(x, \lambda) \tag{25}$$

subject to

$$g_i(x, \lambda) \geq 0, i = 1, \ldots, k, i \neq j \tag{26}$$

Step 3: If $g_\ell(x(\lambda), \lambda) \geq 0, \ell = 1, \ldots, k$ then set $\lambda_1 \leftarrow \lambda, \lambda \leftarrow (\lambda_0 + \lambda_1)/2$. Otherwise, set $\lambda_0 \leftarrow \lambda, \lambda \leftarrow (\lambda_0 + \lambda_1)/2$. If $|\lambda_1 - \lambda_0| < \epsilon$ then go to Step 4, where ϵ is a sufficiently small positive constant. Otherwise, go to Step 2.

Step 4: Set $\lambda^* \leftarrow \lambda$ and $x^* \leftarrow x(\lambda)$. The optimal solution (x^*, λ^*) of MINMAX2($\hat{\mu}$) is obtained.

3 An Interactive Algorithm

In this section, we propose an interactive algorithm to obtain a satisfactory solution from among a D-Pareto optimal solution set. From Theorem 1, it is not guaranteed that the optimal solution (x^*, λ^*) of MINMAX2($\hat{\mu}$) is D-Pareto optimal, if it is not unique. In order to guarantee the D-Pareto optimality , we first assume that k constraints (15) are active at the optimal solution (x^*, λ^*), *i.e.*,

$$\sum_{j=1}^{n}\{m_{ij} - L^{-1}(\hat{\mu}_i - \lambda^*)\alpha_{ij}\}x_j^* + \Phi^{-1}(\mu_{\tilde{G}_i}^{-1}(\hat{\mu}_i - \lambda^*))\sqrt{x^{*T}V_i x^*} = \mu_{\tilde{F}_i}^{-1}(\hat{\mu}_i - \lambda^*), \quad i = 1, \cdots, k. \tag{27}$$

If the ℓ-th constraint of (15) is inactive, *i.e.*,

$$\sum_{j=1}^{n}\{m_{\ell j} - L^{-1}(\hat{\mu}_\ell - \lambda^*)\alpha_{\ell j}\}x_j^* + \Phi^{-1}(\mu_{\tilde{G}_\ell}^{-1}(\hat{\mu}_\ell - \lambda^*))\sqrt{x^{*T}V_\ell x^*} < \mu_{\tilde{F}_\ell}^{-1}(\hat{\mu}_\ell - \lambda^*), \tag{28}$$

we can convert the inactive constraint (28) into the active one by applying the bisection method for the reference membership value $\hat{\mu}_\ell \in [\lambda^*, \lambda^* + 1]$.

For the optimal solution (x^*, λ^*) of MINMAX2$(\hat{\mu})$, where the active conditions (27) are satisfied, we solve the D-Pareto optimality test problem defined as follows.

[**D -Pareto optimality test problem**]

$$\max_{x \in X, \epsilon_i \geq 0, i=1,\cdots,k} w \stackrel{\text{def}}{=} \sum_{i=1}^{k} \epsilon_i \tag{29}$$

subject to

$$\sum_{j=1}^{n}\{m_{ij} - L^{-1}(\hat{\mu}_i - \lambda^*)\alpha_{ij}\}x_j + \Phi^{-1}(\mu_{\tilde{G}_i}^{-1}(\hat{\mu}_i - \lambda^*))\sqrt{x^T V_i x} + \epsilon_i \leq \sum_{j=1}^{n}\{m_{ij} - L^{-1}(\hat{\mu}_i - \lambda^*)\alpha_{ij}\}x_j^* + \Phi^{-1}(\mu_{\tilde{G}_i}^{-1}(\hat{\mu}_i - \lambda^*))\sqrt{x^{*T}V_i x^*}, \; i = 1, \ldots, k \tag{30}$$

For the optimal solution of the above test problem, the following theorem holds.

Theorem 2 For the optimal solution $\breve{x}, \breve{\epsilon}_i, i = 1, \ldots, k$ of the test problem (29)–(30), if $w = 0$ (equivalently, $\breve{\epsilon}_i = 0, i = 1, \ldots, k$), $x^* \in X$, $f_i^* \stackrel{\text{def}}{=} \mu_{\tilde{F}_i}^{-1}(\hat{\mu}_i - \lambda^*) \in \mathrm{R}^1, i = 1, \ldots, k$ is a D-Pareto optimal solution.

Proof

From the active condition (27) at the optimal solution (x^*, λ^*) of MINMAX2$(\hat{\mu})$, it holds that

$$\hat{\mu}_i - \lambda^* = \Pi_{\tilde{P}_i(x^*, f_i^*)}\left(\tilde{G}_i\right), i = 1, \ldots, k,$$
$$\hat{\mu}_i - \lambda^* = \mu_{\tilde{F}_i}\left(f_i^*\right), i = 1, \ldots, k.$$

Assume that $x^* \in X, \mu_{\tilde{F}_i}^{-1}(\hat{\mu}_i - \lambda^*), i = 1, \cdots, k$ is not a D-Pareto optimal solution. Then, there exist $x \in X$, $f_i \in \mathrm{R}^1, i = 1, \ldots, k$ such that

$$\mu_{D_i}(x, f_i) = \min\{\Pi_{\tilde{P}_i(x, f_i)}(\tilde{G}_i), \mu_{\tilde{F}_i}(f_i)\} \geq \mu_{D_i}(x^*, f_i^*)$$
$$= \hat{\mu}_i - \lambda^*, i = 1, \ldots, k,$$

with strict inequality holding for at least one i. This means that

$$\Pi_{\tilde{P}_i(x, f_i)}(\tilde{G}_i) \geq \hat{\mu}_i - \lambda^*, i = 1, \ldots, k, \tag{31}$$

$$\mu_{\tilde{F}_i}(f_i) \geq \hat{\mu}_i - \lambda^*, i = 1, \ldots, k. \tag{32}$$

From (13), (31) and (32), the following inequalities hold,

$$\sum_{j=1}^{n}\{m_{ij} - L^{-1}(\hat{\mu}_i - \lambda^*)\alpha_{ij}\}x_j$$
$$+ \Phi^{-1}(\mu_{\tilde{G}_i}^{-1}(\hat{\mu}_i - \lambda^*))\sqrt{x^T V_i x} \leq \mu_{\tilde{F}_i}^{-1}(\hat{\mu}_i - \lambda^*), i = 1, \ldots, k \tag{33}$$

with strict inequality holding for at least one i. This contradicts the fact that $\breve{\epsilon}_i = 0, i = 1, \ldots, k$.

Now, following the above discussions, we can present the interactive algorithm in order to derive a satisfactory solution from among a D-Pareto optimal solution set.

[Algorithm 2]

Step 1: The decision maker sets each of the membership functions $\mu_{\tilde{F}_i}(f_i), i = 1, \ldots, k$ of the fuzzy goal $\tilde{F}_i$ for permissible objective level f_i according to Assumption 2.

Step 2: Corresponding to the fuzzy goal $\tilde{G}_i$ for the probability that the objective function $\overline{\tilde{C}}_i x$ is less than f_i, the decision maker sets each of the membership functions $\mu_{\tilde{G}_i}(p_i), i = 1, \cdots, k$ according to Assumption 1.

Step 3: Set the initial reference membership values as $\hat{\mu}_i = 1, i = 1, \ldots, k$.

Step 4: Solve MINMAX2($\hat{\mu}$) by applying Algorithm 1, and obtain the optimal solution (x^*, λ^*). For the optimal solution (x^*, λ^*), The corresponding D-Pareto optimality test problem (29)–(30) is formulated and solved.

Step 5: If the decision maker is satisfied with the current values of the D-Pareto optimal solution $\mu_{D_i}(x^*, f_i^*), i = 1, \cdots, k$ where $f_i^* = \mu_{\tilde{F}_i}^{-1}(\hat{\mu}_i - \lambda^*)$, then stop. Otherwise, the decision maker updates his/her reference membership values $\hat{\mu}_i, i = 1, \ldots, k$, and return to Step 4.

4 A Numerical Example

In order to demonstrate our proposed decision making method, we consider the following crop planning problem [5, 7, 9, 21], in which a farmer or an agricultural manager wants to maximize his/her total profit (unit: 1,000 yen) and minimize his/her working time (unit: 1 h) by using his/her farmland effectively. In order to decide the planting ratio for five kinds of crops x_j, $j = 1, \ldots, 5$ (unit: 1,000 m^2) in his/her farmland, we formulate the following multiobjective fuzzy random programming problem.

[RFMOLP-VC]

$$\min \overline{\tilde{C}}_1 x = \sum_{j=1}^{5} \overline{\tilde{C}}_{1j} x_j \text{(profit maximization)}$$

$$\min \overline{\tilde{C}}_2 x = \sum_{j=1}^{5} \overline{\tilde{C}}_{2j} x_j \text{(labor minimization)}$$

subject to

$$x \in X = \left\{ x \in \mathrm{R}^5 | x_1 + x_2 + x_3 + x_4 + x_5 \leq 10, x_j \geq 0, 1 \leq j \leq 5 \right\}$$

where $x = (x_1, x_2, x_3, x_4, x_5)^T$ is a five dimensional decision variable column vector, each element x_j means the cultivation area for crop j. $\overline{\tilde{C}}_{1j}$, $j = 1, \cdots, 5$ are profit coefficients at the unit area for crop j, and $\overline{\tilde{C}}_{2j}$, $j = 1, \cdots, 5$ are working time coefficients for growing crop j at the unit area, each of which is defined as a random fuzzy variable. Random fuzzy variables $\overline{\tilde{C}}_{ij}$, $i = 1, 2$, $j = 1, \ldots, 5$ are normally distributed with the fuzzy number $\tilde{M}_{ij}$ as mean and σ_{ijj} as variance, and $\tilde{M}_{ij}$ is set as Yano and Matsui [25]. The variance–covariance matrices of $\overline{\tilde{C}}_1$ and $\overline{\tilde{C}}_2$ are as follows.

$$V_1 = \begin{pmatrix} \sigma_{111} & \sigma_{112} & \sigma_{113} & \sigma_{114} & \sigma_{115} \\ \sigma_{121} & \sigma_{122} & \sigma_{123} & \sigma_{124} & \sigma_{125} \\ \sigma_{131} & \sigma_{132} & \sigma_{133} & \sigma_{134} & \sigma_{135} \\ \sigma_{141} & \sigma_{142} & \sigma_{143} & \sigma_{144} & \sigma_{145} \\ \sigma_{151} & \sigma_{152} & \sigma_{153} & \sigma_{154} & \sigma_{155} \end{pmatrix} = \begin{pmatrix} 2.93 & -3.48 & 2.05 & 0.62 & 1.39 \\ -3.48 & 8.15 & -1.46 & 2.28 & -1.05 \\ 2.05 & -1.46 & 4.35 & 2.12 & 1.83 \\ 0.62 & 2.28 & 2.12 & 4.89 & -0.07 \\ 1.39 & -1.05 & 1.83 & -0.07 & 1.97 \end{pmatrix}$$

$$V_2 = \begin{pmatrix} \sigma_{211} & \sigma_{212} & \sigma_{213} & \sigma_{214} & \sigma_{215} \\ \sigma_{221} & \sigma_{222} & \sigma_{223} & \sigma_{224} & \sigma_{225} \\ \sigma_{231} & \sigma_{232} & \sigma_{233} & \sigma_{234} & \sigma_{235} \\ \sigma_{241} & \sigma_{242} & \sigma_{243} & \sigma_{244} & \sigma_{245} \\ \sigma_{251} & \sigma_{252} & \sigma_{253} & \sigma_{254} & \sigma_{255} \end{pmatrix} = \begin{pmatrix} 2.34 & -1.14 & -1.75 & -1.64 & 2.35 \\ -1.14 & 4.31 & 2.51 & 2.42 & 0.61 \\ -1.75 & 2.51 & 6.39 & 3.86 & -0.13 \\ -1.64 & 2.42 & 3.86 & 2.74 & -0.62 \\ 2.35 & 0.61 & -0.13 & -0.62 & 6.55 \end{pmatrix}$$

In RFMOLP-VC, let us assume that the hypothetical decision maker adopts the same membership functions as $\mu_{\tilde{F}_i}(\cdot)$, $\mu_{\tilde{G}_i}(\cdot)$, $i = 1, 2$ of [25] (Step 1, 2). Set the initial reference membership values as $(\hat{\mu}_1, \hat{\mu}_2) = (1, 1)$ (Step 3), and solve MINMAX2$(\hat{\mu})$ (Step 4). Then the corresponding D-Pareto optimal solution(x^*, λ^*) is obtained as follows, where $f_i^* = \mu_{\tilde{F}_i}^{-1}(\hat{\mu}_i - \lambda^*)$, $i = 1, 2$.

$$(\mu_{D_1}(x^*, f_1^*), \mu_{D_2}(x^*, f_2^*)) = (0.7148, 0.7148),$$

$$(f_1^*, f_2^*) = (-1464.62, 1285.55),$$

$$(p_1^*, p_2^*) = (0.8072, 0.8741).$$

The hypothetical decision maker is not satisfied with the current value of the D-Pareto optimal solution (x^*, f_i^*), and, in order to improve $\mu_{D_1}(\cdot)$ at the expense of $\mu_{D_2}(\cdot)$, he/she updates his/her reference membership values as $(\hat{\mu}_1, \hat{\mu}_2) = (1, 0.8)$ (Step 5). Then, the corresponding D-Pareto optimal solution is obtained as follows (Step 4):

$$\begin{aligned}(\mu_{D_1}(x^*, f_1^*), \mu_{D_2}(x^*, f_2^*)) &= (0.7967, 0.5967),\\ (f_1^*, f_2^*) &= (-1517.87, 1320.98),\\ (p_1^*, p_2^*) &= (0.8195, 0.8596).\end{aligned}$$

In order to compare our proposed approach with the previous ones, let us consider the following multiobjective programming problem based on a probability maximization model for RFMOLP-VC.

[**MOP2'(*f*)**]

$$\max_{x \in X} (\Pi_{\tilde{P}_1(x, f_1)}(\tilde{G}_1), \Pi_{\tilde{P}_2(x, f_2)}(\tilde{G}_2))$$

where f_1 and f_2 are permissible objective levels specified by the decision maker in his/her subjective manner. Once the reference membership values $\hat{\mu} = (\hat{\mu}_1, \hat{\mu}_2)$ are specified by the decision maker, the corresponding Pareto optimal solution is obtained by the following minmax problem.

[**MINMAX3($\hat{\boldsymbol{\mu}}$,$\boldsymbol{f}$)**]

$$\min_{x \in X, \lambda \in \Lambda} \lambda$$

subject to

$$\hat{\mu}_i - \Pi_{\tilde{P}_i(x, f_i)}(\tilde{G}_i) \leq \lambda, i = 1, 2$$

We can easily solve MINMAX3$(\hat{\mu}, f)$ by applying Algorithm 1, because the constraint set (34) are convex for any fixed $\lambda \in \Lambda$. Let us assume that the decision maker sets his/her reference membership values as $(\hat{\mu}_1, \hat{\mu}_2) = (1, 1)$ and permissible objective levels as $(f_1, f_2) = (-1550, 1240)$. Then, the corresponding Pareto optimal solution can be obtained as shown as:

$$
\begin{aligned}
(f_1, f_2) &= (-1550, 1240)\,, \\
\left(p_1^*, p_2^*\right) &= (0.7731, 0.8487)\,, \\
\left(\mu_{\tilde{F}_1}(f_1), \mu_{\tilde{F}_2}(f_2)\right) &= (0.7692, 0.8666)\,, \\
\left(\mu_{\tilde{G}_1}(p_1^*), \mu_{\tilde{G}_2}(p_2^*)\right) &= (0.4876, 0.4876)\,.
\end{aligned}
$$

It is clear that, in the proposed method, a proper balance between permissible probability levels and the corresponding objective functions in a probability maximization model is attained in membership space. On the other hand, In a probability maximization model based method, although permissible objective levels are improved in comparison with the proposed method, the corresponding probability function values was changed for the worse.

5 Conclusion

In this paper, we have proposed an interactive decision making method for RFMOLP-VC based on a probability maximization model to obtain a satisfactory solution from among a Pareto optimal solution set. In the proposed method, the decision maker is required to specify the membership functions for the fuzzy goals of not only the permissible objective levels in a probability maximization model but also the corresponding distribution function. Such two kinds of membership functions are integrated and, in the integrated membership space, a D-Pareto optimal solution concept is introduced. The satisfactory solution can be obtained by updating the reference membership values and solving the corresponding minmax problem by applying the convex programming technique. At any D-Pareto optimal solution, it is guaranteed that a proper balance between permissible objective levels and the corresponding distribution function values in a probability maximization model is attained.

References

1. Birge JR, Louveaux F (1997) Introduction to stochastic programming. Springer
2. Charnes A, Cooper WW (1959) Chance constrained programming. Manage Sci 6:73–79
3. Danzig GB (1955) Linear programming under uncertainty. Manage Sci 1:197–206
4. Dubois D, Prade H (1980) Fuzzy sets and systems: theory and applications. Academic Press
5. Glen JJ (1987) Mathematical models in farm planning : a survey. Oper Res 35:641–666
6. Hasuike T, Katagiri K, Ishii H (2009) Portfolio selection problems with random fuzzy variable returns. Fuzzy Sets Syst 160:2579–2596
7. Hayashi K (2000) Multicriteria analysis for agricultural resource management : a critical survey and future perspectives. Eur J Oper Res 122:486–500
8. Huang X (2007) Optimal Project Selection with Random Fuzzy Parameters. Int J Prod Econ 106:513–522

9. Itoh T, Ishii H, Nanseki T (2003) A model of crop planning under uncertainty in agricultural management. Int J Prod Econ 81–82:555–558
10. Kall P, Mayer J (2005) Stochastic linear programming models, theory, and computation. Springer, Berlin
11. Katagiri H, Hasuike T, Ishii H (2008) A random fuzzy programming models based on possibilistic programming. In: Proceedings of 2008 IEEE international conference on systems, man and, cybernetics (SMC2008), pp 1788–1793
12. Katagiri H, Sakawa M, Matsui T (2011) Interactive multiobjective random fuzzy programming problems through the possibility-based fractile model. In: Proceedings of 2011 IEEE international conference on systems, man and, cybernetics (SMC2011), pp 3144–3148
13. Katagiri H, Sakawa M, Matsui T (2011) An interactive satisficing method for multiobjective random fuzzy programming problems through the possibility-based probability model. In: Proceedings of 2011 IEEE international conference on fuzzy systems, pp 1778–1782
14. Katagiri H, Uno T, Kato K, Tsuda H, Tsubaki H (2012) Interactive multiobjective random fuzzy programming: necessity-based value at risk model. In: Proceedings of 2012 IEEE international conference on systems, man and, cybernetics (SMC2012), pp 727–732
15. Katagiri H, Uno T, Kato K, Tsuda H, Tsubaki H (2013) Random fuzzy multi-objective linear programming: optimization of possibilistic value at risk (PVaR). Expert Syst Appl 40:563–574
16. Lai YJ, Hwang CL (1992) Fuzzy mathematical programming. Springer, Berlin
17. Liu B (2002) Random fuzzy dependent-chance programming and its hybrid intelligent algorithm. Inf Sci 141:259–271
18. Liu B (2004) Uncertainty theory. Springer, Berlin
19. Sakalli US, Baykoc OF (2010) An application of investment decision with random fuzzy outcomes. Expert Syst Appl 37:3405–3414
20. Sakawa M (1993) Fuzzy sets and interactive multiobjective optimization. Plenum Press, New York
21. Toyonaga T, Itoh T, Ishii H (2005) A crop planning with fuzzy random profit coefficients. Fuzzy Optim Decis Making 4:51–69
22. Xu J, Yao L, Zhao X (2011) A multi-objective chance-constrained network optimal model with random fuzzy coefficients and its application to logistics distribution center location problem. Fuzzy Optim Decis Making 10:255–285
23. Yano H (2012) Interactive decision making for fuzzy random multiobjective linear programming problems with variance-covariance matrices through probability maximization. In: Proceedings of the 6th international conference on soft computing and intelligent systems and 13th international symposium on advanced intelligent systems, SCIS-ISIS 2012, 20–24 Nov 2012, Kobe, pp 965–970
24. Yano H, Matsui M (2013) Random fuzzy multiobjective linear programming through probability maximization. In: Proceedings of the international multi conference of engineers and computer scientists 2013, IMECS 2013, lecture notes in engineering and computer Science, 13–15 Mar 2013, Hong Kong, pp 1135–1140
25. Yano H, Matsui M (2013) Random fuzzy multiobjective linear programming through probability maximization and its application to farm planning. IAENG Int J Appl Math 43(2):87–93
26. Yano H, Sakawa M (2012) Interactive multiobjective fuzzy random linear programming through fractile criteria. Advances in fuzzy systems, vol 2012. Hindawi Publishing Corp., Article ID:521080
27. Yano H, Sakawa M (2013) Interactive fuzzy programming for multiobjective fuzzy random linear programming problems through possibility-based probability maximization. Int J Oper Res. doi:10.1007/s12351-013-0135-4
28. Zimmermann H-J (1987) Fuzzy sets, decision-making and expert systems. Kluwer Academic Publishers, Boston

The Different Ways of Using Utility Function with Multi-choice Goal Programming

Ching-Ter Chang and Zheng-Yun Zhuang

Abstract By studying the way a previous study is using utility function with multi-choice goal programming (MCGP), some drawbacks are identified. These drawbacks can mistakenly result in an incomplete representativeness of the original MCGP with utility functions model and thus lead to the inability of the model to appropriately assess and express the real preference structure of a decision maker (DM). This study recommends some points and proposes another ways of using utility functions with MCGP. These new ways are validated by using them to express different DM preference structures pertaining to a goal, which underlies his/her real oral statements during decision making.

Keywords Decision maker statements · Decision making · Goal programming · Multi-choice goal programming · Preference structure · Utility function

1 Introduction

Decision-making via goal programming (GP) is ubiquitous nowadays. As a good tool for solving multi-criteria decision-making (MCDM) problems with multiple objectives, GP has gained its wide popularity [1]. It takes into account multi-criteria and multi-objective concerns of multiple goals, and is based on mathematical programming, which is quite different from other algorithm-based decision approaches

Z.-Y. Zhuang (✉)
School of Computer and Computing Science, City College, Zhejiang University, Hangzhou 310015, China
e-mail: zhuangzy@zucc.edu.on

C.-T. Chang
The Graduate Institute of Buisness Management, Chang Gung University, 259 Webhua 1st Rd.,Taoyuan 333, Taiwan
e-mail: chingter@mail.cgu.edu.tw

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*, Lecture Notes in Electrical Engineering 275, DOI: 10.1007/978-94-007-7684-5_28,

[2, 3]. It is still popular now and continues to be irrigated by researchers and practitioners [4]. As a special extension of linear programming, GP was first introduced by Charnes and Cooper [5]. Since then, important extensions and numerous applications have been proposed [6].

The literature is abundant with applications of GP. Over the decades, GP has been used to support real-world decision-making processes in many fields such as communication, energy, manufacturing, medical healthcare, vendor selection, pricing, and so on [7–13].

The literature is also abundant with *GP variants* (i.e., the GP extension models or specific-purposed formulations for GP). The extension models include, for example, weighted GP (WGP) [14], interactive GP [15], integer GP instead of continuous GP, interval GP (IGP), fuzzy GP (FGP) [16], multi-choice GP (MCGP) [17], multi-segment GP (MSGP) [18], percentage GP (%GP) [19], etc. There are studies about the reformulations of either the objective measurement or the goal constraints, too. They are the value-added components of GP that can enhance the solving range and widen the use of GP in different application scenarios. This category includes, but not limited to (for space reasons we just list some recent works), formulating arbitrary penalty function for interval programming [20], formulating the S-shaped penalty function for IGP [21], formulating procurement risk using a possibility formulation for fuzzy multi-objective programming [22], weighted max-min model for FGP [23], and so on.

However, the above stratifications of GP variants are, in fact, categorized *methodologically*. When they are categorized through their original ideas, perhaps the article by Tamiz et al. published in 1998 [24], which is an overview of GP modeling techniques that has categorized the *considerations*, can help.

In the field of GP, as pointed out in [24], considerations in making the GP variants include Pareto optimality consideration, normalization technique consideration, preferential weights selection consideration and GP's utility interpretation consideration. As can be identified, issues about the last two considerations are strongly associated with DM's preference structure, so that works in GP to properly express the DM's preferences mainly fall in the last two considerations [25].

Studies about the preference structure issues have also become a subset of GP researches. In 1978, Zimmermann proposed the concept of fuzzy programming (FP) [26], in which the right-hand-side (RHS) of a constraint can be fuzzified with utility functions, and then Narasimhan (1980) applied the "fuzzy subsets" concepts to GP in a fuzzy environment [16]. Martel and Aouni [27] utilized the Promethee method to build the preference structure of DM for GP and avoided incommensurability problems between criteria that different measurement units. Later, Yang et al. [28] proposed the fuzzy programming model with non-linear membership functions, using piecewise line segments to approximate the non-linear functions . Mohamed [29] discussed the relationships (how one can lead to another) between GP and FP. Romero [14] summarized the general structure of the lexicographic and weighted achievement functions for GP, to deal with the different philosophies of DM preference. Chang [30] proposed an approximation approach for the S-shaped membership functions, while Chang [31] later integrated the concept of utility function into MCGP.

To represent the preference structure of DMs, Chang's last model (hereafter, briefed as the "MCGP+U" model), involves using multiple possible choices of RHS values for each goal constraint to represent the multiple goal aspiration levels of a DM. The study have used utility function to "glue-up" the RHS values and proposed a model that integrates the utility function concept into MCGP, to solve the decision problem.

Within the MCGP+U model, for a "the-less-the-better" criterion, no matter the utility function is a purely linear one or a piecewise-linear one (or comes in any other shape else, e.g., Fig. 1), it always begins with 1 when the achieved goal value is at the *minimum bound* and ends with 0 when the achieved goal value is at the *maximum bound*. The situation is vice versa for a "the-more-the-better" criterion. The "maximum and minimum bounds" are respectively defined by the maximal possible aspiration level of a goal and the minimal possible level, among the multiple choices of aspiration levels listed by the DM.

However, using the utility function in this way can be inappropriate. For example, when people are applying MCGP+U to a real world decision-making case, the slope of the utility function of any specific "the-less-the- better" goal constraint is not necessary to end (i.e., to reach 0) at the maximum bound (i.e., the maximal listed aspiration level). Rather, the location at which the slope of the utility function falls to 0 should depend on the DM's statement which reveals his/her real preference structure in mind.

Thus, there is a need to change, or revise at least, the way utility function is used with MCGP, so as to reflect the real preference structure of a DM and to correct the possible misleading caused by the previous MCGP+U study.

To have a better understanding about the above core claim mentioned by this study, the original way of utility function using is examined in Sect. 2, while the main shortcomings of using utility functions in this manner is identified. By observation of some statements of the DMs during decision making, Sect. 3 proposes some new ways and examines if these new ways can perfectly match and express the DM preference structure. Section 4 concludes this study.

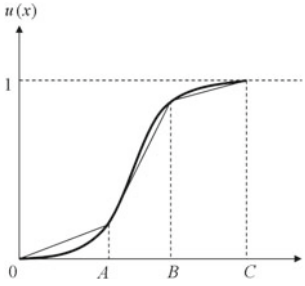


Fig. 1 A non-linear utility function approached by piecewise line segments

2 The Way Utility Function was Used with MCGP

For simplicity of illustration, the multi-objective numerical decision problem case in Chang's study [31] is adopted here, as follows:

(MODM Problem P1)
(Goal 1) $8x_1 + 9x_2 + 6.5x_3 \geq 50$
(Goal 2) $x_1 + 0.2x_2 + 0.5x_3 \leq 5$, with utility function shown in Fig. 2.
(Goal 3) $2x_1 + 3x_2 + 2.4x_3 \geq 10$, with utility function shown in Fig. 3.
(Constraints) $5x_1 + 7x_2 + 5.5x_3 \leq 40$.

where x_1, x_2, x_3 are the market share of each product, in units and should be no less than 2 units; for details please check the original article.

There are some facts one can read from these figures, regardless of what the approach is taken to formulate these utility functions in [31].

Firstly, the left- and right- triangles depicted in Figs. 2 and 3 represent the two usual types of utility functions (i.e., the-less-the-better and the usual the-more-the-better) that are widely accepted by researchers or practitioners when they are incorporating utility function for solving problems.

Next, if the real shape of a utility function is not purely linear such as the shapes shown in Figs. 2 or 3, piecewise linear approach can be used to approximate and represent it. In fact, once any utility function is approximated by, or in itself, piecewise

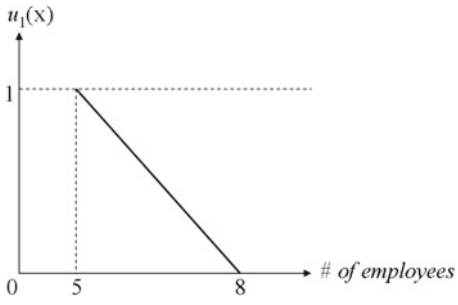


Fig. 2 The linear and *left-triangular* utility function

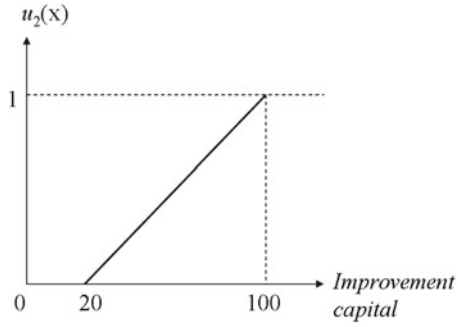


Fig. 3 The linear and *right-triangular* utility function

line segments, the work left is to find a proper method to model these pieces. For example, Yang et al. [28] proposed a smart approach that uses only 1 binary variable to formulate an S-shaped utility function that is composed of 3 linear pieces. Chang [30] proposed another approach that can deal with an S-shaped utility function which are not purely increasing (concave) or purely decreasing (convex).

Thirdly, as mentioned previously, it may be questionable that such utility functions can fully represent all the facts about a DM's requirement pertaining to his/her real preference structure, for the reason that it simply bounds the feasible achievement levels of a goal with a fixed interval. For example, for Goal 2, the possible levels for this goal to achieve are hardly constrained by [5, 8], so to Goal 3, wherein the achieve-able goal value are bounded by [20, 100].

However, in real decision cases, the situation for a goal to be achieved outside the sloping range (e.g. the # of employees < 5 cases) is very possible and reasonable, if doing so leads to more beneficial decisions. That is, allowing the goal achievement to fall in some place outside the defined sloping range is necessary. For example, it can be easily seen that in Fig. 1, [0, 5] and [8, unlimited) can both be encountered as the intervals for Goal 2 to achieve, rather than the sloping range [5, 8] merely. These two additional intervals should continuously produce utility levels of 1 and 0, respectively.

Fourthly, as mentioned previously, it is not necessary for the slope of a "the-less-the-better" utility function to fall to 0 rightly at the place where the maximum bound is. The situation for a "the-more-the-better" utility function is analogous. That is, with the original MCGP+U model, each goal constraint has multiple choices of the RHS values to serve as the aspiration levels of the goal. When a DM has multiple possible aspiration levels pertaining to a goal, the range of a utility function to rise up or to slope down (i.e., the "sloping range") is defined by the maximal and minimal values of these seen levels.

Nevertheless, this is not always true in practice. In Chang's study [31], for a goal, the "maximum bound" of the utility function is tantamount to the maximal seen aspiration level of that goal. However, the maximal seen aspiration level is usually not equivalent to the point the DM becomes fully frustrated, which means location where the sloping range ends should be reconsidered. As can be imagined, the point where the DM becomes fully unsatisfied usually comes earlier than the maximal seen aspiration level.

3 New Ways of Using Utility Function with MCGP

By observation of the above mentioned facts, this study suggests another ways to use utility function for MCGP. To differentiate these new ways from the original way, a thorough dissection is required.

3.1 The Example DM Preference Statement

As can be seen, via the original way to use utility function for MCGP+U, when a DM has multiple possible aspiration levels pertaining to a goal, the sloping range of the utility function is defined by the maximum and minimum bounds (i.e., maximal and minimal values of the seen aspiration levels claimed by the DM). Besides, this range is exactly the target range for this goal to achieve.

As mentioned previously, using utility functions in this manner might be inappropriate for the sake of the inability to represent a DM's preference properly and fully. Such using way might have missed a critical fact that the sloping range should not be defined and bounded by the maximum and minimal values among the multiple, possible aspiration levels of a goal. Rather, it involves the DM's real perception in regard to the durable ranges of that goal. Moreover, it might have also missed another critical fact that the target range for a goal to achieve cannot simply conform to the pre-defined sloping range of the utility function.

Take G2 in Fig. 2 for example, given that the DM has listed the possible aspiration levels (multiple choices of the RHS) of this goal as 5, 6, 7 and 8. Then, will the DM feel unhappy with lesser than 5 employees? And is 8 always the end (i.e., the utility function value is becoming 0) of the slope range of the utility function? What if the DM expressed a statement like *S*1?

Statement *S*1: "*A total number of* 5, 6, 7, 8 *employees are all possible and a total number less than or equal to* 5 *is strongly acceptable, but for some reasons a total number greater than* 7 *is very unacceptable.*"

The statement "a total number of 5, 6, 7, 8 employees are all possible" in *S*1 means that 5, 6, 7, 8 are all possible choices of the RHS aspiration levels of the goal constraint. But the statement "for some reasons a total number greater than 7 is very unacceptable" means that the end of the sloping range of the utility function to reach 0 is not 8, but instead, it can be 7, 7.5, or so. Note that here there can be many utility function interpretations for the claim "greater than 7 is very unacceptable". Anyhow, with such a consideration, the original way of use of utility function by Chang [31] may distort a DM's real preference structure and might be infeasible to produce in practice.

Moreover, the sloping range of Goal 2 is closed. That is, only the interval [5,8] is allowed for Goal 2 to achieve, which is in fact, the sloping range of the utility function exactly. However, inside S1, there is a statement which states that "a total number less than or equal to 5 is strongly acceptable". This means that a total number of employees under 5 (e.g., 1, 2, 3, or 4, while 0 is quite impossible) also makes the DM perfectly happy. Unfortunately, if a DM had taken the version of utility function in Fig. 2, the model could not solve out any answer that leads to 1, 2, 3 and 4 employees just because the utility function is, in itself, bounded by the "at least 5" condition. This can again distort a DM's real preference and can be another flaw that deters any practitioner from applying the MCGP+U model.

3.2 The New Way to Use Utility Functions with MCGP

Therefore, this study disseminates the idea of making some adjustments to the use of utility function for MCGP. For example, the utility function in Fig. 2 can be adjusted as the one shown in Fig. 4, or the one in Fig. 5.

In Fig. 4, the thin solid line, which conforms to Fig. 2, is the original version of the utility function $u(x)$ that described the RHS of Goal 2. As discussed, the feasible range of this goal is bounded by [5,8], which is, exactly, also the sloping range of its utility function.

On the contrary, the bold solid line, $u_1'(x)$, is the proposed version of utility function. In compare with $u_1(x)$ that contains solely one sloping line segment, $u_1'(x)$ has three line segments including one flat segment, one sloping segment and one flat segment that follows, each of which is respectively delimited by $x = [0, 5)$, [5,7] and (7,8]. This relaxes the limitation of the feasible range of Goal 2 that had to be [5,8]. The feasible range of Goal 2 becomes [0,8] instead, to include the possibly forgotten feasible numbers of {0, 1, 2, 3, 4} employees, and to include the possible levels of {5, 6, 7, 8} employees, which is the original feasible range of $u_1(x)$.

In addition, as can be seen in Fig. 4, within the original feasible range of $u_1(x)$, which is [5,8], there are two piecewise line segments delimited by $x = [5, 7]$ and $x = (7, 8]$ respectively.

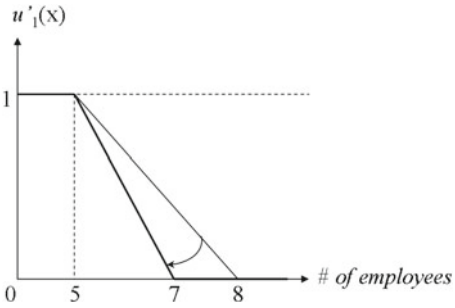


Fig. 4 The revised utility function for goal G2

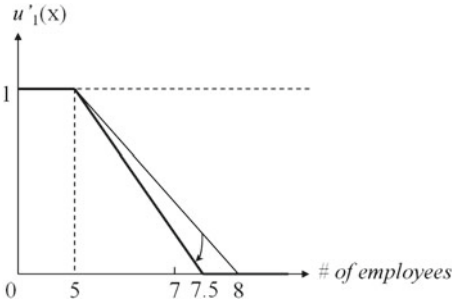


Fig. 5 Another revised utility function for goal G2

The former part, delimited by $x = [5, 7]$, is the interval that $u_1'(x)$ really slopes down. To conform to the statement S1, the sloping range changes from [5, 8] [of $u_1(x)$] to [5, 7] [of $u_1'(x)$]. So the slope should shrink proportionally, no matter what the shape of the slope is (e.g., the linearly-shaped one like this case, or an S-shaped or concave-shaped one for some other cases). The arrow in Fig. 4 displays this shrinking process. After shrinking, the slope of the utility function becomes abrupt because the slope range is lessened but the maximal height of the function remains the same (i.e., 1).

The latter tail line segment delimited by $x = (7, 8]$ is a flat segment that produces a satisfaction level of 0. This can meet the DM's claim "but a total number greater than 7 is very unacceptable" in $S1$. Beware that this segment, although seems trivial, is necessary to stretch till the maximal possible aspiration level defined by the DM, so as to faithfully and completely represent the full feasible range in a DM's mind.

Neglecting this tail part can result in serious consequence. For example, if one excluded this part (i.e., the interval (7, 8]) from modeling, the model would have never got an answer that achieves a goal value of 8 for Goal 2, whatever the other constraints are. Consider the case in which a DM is applying MCGP with utility function and there is a must to list 8 as a possible aspiration level of Goal 2 (which implies the achievement of this criterion can be sacrificed when leveraging with other criteria). If merely the first flat range [0, 5) and the slope range [5, 7) were taken into account, the model could miss the possible aspiration level of 8, and more critically, would distort the optimal solution because those feasible solutions that achieves Goal 2 with a level of 8 are totally ignored.

Figure 4 is based on interpreting "an employee number greater than 7 is very unacceptable" as "the DM becomes fully unhappy right at 7 and the slope ends at 7". Because this goal has integer target values, if we interpret the same statement as "the DM becomes fully unhappy at 7.5 and the slope should end at 7.5", the utility function should be the one shown in Fig. 5.

In summary, these new forms of $u_1'(x)$ can fully describe a DM statement such as $S1$, as it can be used to denote the real satisfactory level of a DM more precisely.

3.3 Extension to the New Utility Function

Let us look back to Fig. 4 and note a little revision here. If we intentionally let the tail line segment of $u_1'(x)$ to penetrate through $x = 8$, the tail segment can span from 7 to a certain bound or the unlimited [i.e., $x = [7, B_1]$ or $x = [7, \infty)$, where B_1 is the DM-perceived upper bound of the x-axis of $u_1'(x)$].

This can be also an important improvement for the original way of use of utility function with MCGP. Such improvement is from the fact that if Goal 2 is not an important goal (e.g., its deviation is not with priority or is assigned a extra low weight in the objective function), a DM can allow it to be "further sacrificed" by extending the tail segment to be ended with 9, 10, 11 and so on, solving the model again, and seeing how the optimal solution is changing.

For example, the DM's statement can be S1′ as follows, which states more than *S*1:

Statement S1′: "*For our company, any total number of employees over* 8 *is able to be exercised and a total number less than or equal to* 5 *is strongly preferable. But for some internal reasons, a total number greater than* 7 *is very unacceptable*".

Shall this be the case, the proposed using way in this study, such as $u_1'(x)$, can perfectly serve the purpose, while the original way of use, e.g. $u_1(x)$, may not.

3.4 The Statement for a "The-More-the-Better" Criterion

Section 3.2 and 3.3 demonstrated the new ways of using utility function with MCGP and reshaped the utility function of Goal 2 for its "the-less-the-better"-typed criterion. For a more concrete illustration and a full coverage of the idea revealed by this study, an additional case about a goal with a "the-more-the-better"-typed criterion is given.

Consider the capital improvement goal (Goal 3) in P1. The original claim that the DM has made is as follows, which reflects no more than a greedy DM's prospect:

Statement *S*2: "*A capital improvement between* 20 *and* 100 *M NT$ is preferred. But for some reasons, an improvement under* 35 *M is quite unacceptable. In fact, any capital improvement level more than* 100 *M is, of course, very welcomed and I will be fully satisfied with it.*"

In this case, the sloping range of the utility function in Fig. 3 should also shrink, with a front bottom field that produces the altitude (satisfaction level) of 0, followed by an endless Qinghai-Tibet plateau that produces the altitude of 1 consistently. This leads to another new utility function whose shape is quite similar to the topography around Lhasa, as Fig. 6 has shown.

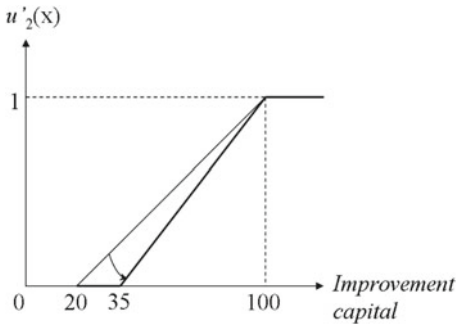


Fig. 6 The revised utility function for MCGP: the-more-the-better case

4 Conclusion and Future Work

This study proposes novel ways of using utility functions when such functions are to be incorporated with existing MCGP models.

Some crucial shortcomings when the original MCGP+U model is to be applied in practice are examined: the missed ranges of the aspiration levels that the original MCGP+U model might fail to formulate. They include the additional "plain field" range which represents the possible choices of aspiration levels that are acceptable by a DM but can lead to 0 satisfaction of the DM. They also include the "plateau" range which represents the other possible choices of goal levels that makes the DM perfectly happy and the possible extended parts of these ranges. Ignoring these facts during modeling could impair the descriptiveness of the model to represent a DM's real preference structure.

In consideration of the aforementioned shortcomings, this study proposes the idea of changing the way utility functions are used with a MCGP model. These new ways express a DM's preference structure more precisely. This point is evidential from the fact that they can faithfully formulate the DM statements based on his/her real preference structure pertaining to the goals.

This study also paves ways to future works.

This study studies the topic of "how the using style of utility functions can be changed". Although the result of this study is quite interesting, this study does not discuss the topics of "how these new ways is to be mathematically formulated" as well as of "how the formulations can be incorporated with MCGP". These become future topics worthy of note.

At last, the core idea to change the way how utility function Utility function is used implies that perhaps there is room for other GP variants, which also incorporate the utility function concept (i.e., other GP extension models with utility functions), to be improved.

Acknowledgments This work was supported in part by the National Science Council of Taiwan under Grant NARPD3B0052.

References

1. Aouni B, Kettani O (2001) Goal programming mode: a glorious history and a promising future. Eur J Oper Res 133:225–231
2. Kyngäs N, Nurmi K, Kyngäs J (2013) The workforce optimization process using the PEAST algorithm. Lecture notes in Engineering and Computer Science: proceedings of The international multiConference of Engineers and Computer Scientists 2013, Hong Kong, 13–15 March 2013, pp 1048–1056.
3. Gorbenko A, Popov V (2012) The Hamiltonian alternating path problem. IAENG Int J Appl Math 42(4):204–213
4. Aouni B, La Torre D (2010) A generalized stochastic goal programming model. Appl Math Comput 215(12):4347–4357

5. Charnes A, Cooper WW (1977) Goal programming and multiple objective optimization. Eur J Oper Res 1:39–51
6. Larbani M, Aouni B (2010) A new approach for generating efficient solutions within the goal programming model. J Oper Res Soc 62:175–182
7. Blake JT, Carter MW (2002) A goal programming approach to strategic resource allocation in acute care hospitals. Eur J Oper Res 140(3):541–561
8. Demirtas EA, Ustun O (2009) Analytic network process and multi-period goal programming integration in purchasing decisions. Comput Ind Eng 56(2):677–690
9. Lin HW, Nagalingam SV, Lin GCI (2009) An interactive meta-goal-programming-based decision analysis methodology to support collaborative manufacturing. Robot Comput Integr Manuf 25(1):135–154
10. Samouilidis JE, Pappas IA (1980) A goal programming approach to energy forecasting. Eur J Oper Res 5(5):321–331
11. Korhonen P, Soismaa M (1988) A multiple criteria model for pricing alcoholic beverages. Eur J Oper Res 37(2):165–175
12. Gómez-Limón JA, Riesgo L (2004) Irrigation water pricing: differential impacts on irrigated farms. Agric Econ 31(1):47–66
13. Senthilkumar K, Lubbers MTMH, De Ridder N, Bindraban PS, Thiyagarajan TM, Giller KE (2011) Policies to support economic and environmental goals at farm and regional scales: outcomes for rice farmers in Southern India depend on their resource endowment. Agric Syst 104(1):82–93
14. Romero C (2004) A general structure of achievement function for a goal programming model. Eur J Oper Res 153(3):675–686
15. Dyer JS (1972) Interactive goal programming. Manage Sci 19(1):62–70
16. Narasimhan R (1980) Goal programming in a fuzzy environment. Decis Sci 11(2):325–336
17. Chang CT (2007) Multi-choice goal programming. Omega 35:389–396
18. Liao CN (2009) Formulating the multi-segment goal programming. Comput Ind Eng 56:138–141
19. Chang CT, Chen HM, Zhuang ZY (2012) Revised multi-segment goal programming: Percentage goal programming. Comput Ind Eng 63:1235–1242
20. Lu HC, Chen TL (2013) Efficient model for interval goal programming with arbitrary penalty function. Optim Lett 7(2):325–341
21. Chang CT, Lin TC (2009) Interval goal programming for S-shaped penalty function. Eur J Oper Res 199:9–20
22. Wu DD, Zhang Y, Wu D, Olson DL (2010) Fuzzy multi-objective programming for supplier selection and risk modeling: a possibility approach. Eur J Oper Res 200:774–787
23. Lin CC (2004) A weighted max-min model for fuzzy goal programming. Fuzzy Sets Syst 142:407–420
24. Tamiz M, Jones DF, Romero C (1998) Goal programming for decision making: an overview of the current state-of-the-art. Eur J Oper Res 111(3):569–581
25. Chang CT, Zhuang ZY (2013) A new way using utility functions for multi-choice goal programming models. Lecture Notes in Engineering and Computer Science: Proceedings of the international multi conference of Engineers and Computer Scientists 2013, Hong Kong, 13–15 March 2013, pp 1141–1145
26. Zimmermann HJ, Fuzzy programming and linear programming with several objective functions. Fuzzy Sets Syst 1:45–55
27. Martel J-M, Aouni B (1990) Incorporating the decision-maker's preference in the goal-programming model. J Oper Res Soc 41(12):1121–1132
28. Yang T, Ignizio JP, Kim HJ (1991) Fuzzy programming with nonlinear membership function: piecewise linear approximation. Fuzzy Sets Syst 41:39–53
29. Mohamed RH (1997) The relationship between goal programming and fuzzy programming. Fuzzy Sets Syst 89:215–222
30. Chang CT (2010) An approximation approach for representing S-shaped membership functions. IEEE Trans Fuzzy Syst 18(2):412–424
31. Chang CT (2011) Multi-choice goal programming with utility function. Eur J Oper Res 215:439–445

A New Viscosity Cesàro Mean Approximation Method for a General System of Finite Variational Inequalities and Fixed Point Problems in Banach Spaces

Poom Kumam, Somyot Plubtieng and Phayap Katchang

Abstract We introduce the new strong convergence theorem by using the Cesàro mean approximation method and viscosity method with weak contraction for finding a common solution of fixed point problems for nonexpansive mappings and general system of finite variational inequalities for finite different inverse-strongly accretive operators in Banach spaces. Our results are extended and improved of some authors' recent results of the literature works in involving this field.

Keywords Accretive operator · Banach space · Cesàro mean · Fixed point · Nonexpansive · Variational inequality · Viscosity · Weak contraction

1 Introduction

Let E be a real Banach space with norm $\|\cdot\|$ and C be a nonempty closed convex subset of E. Let E^* be the dual space of E and $\langle\cdot,\cdot\rangle$ denote the pairing between E and E^*. We recall the following concepts (See also [1] for). Let $T : C \to C$ a nonlinear mapping. We use $F(T)$ to denote the set of fixed points of T, that is, $F(T) = \{x \in C : Tx = x\}$. A mapping T is called $nonexpansive$ if $\|Tx - Ty\| \leq$

P. Katchang (✉)
Division of Mathematics, Faculty of Science and Agricultural Technology, Rajamangala University of Technology Lanna Tak (RMTUL Tak), Tak 63000, Thailand
e-mail: p.katchang@hotmail.com

P. Kumam
Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi (KMUTT), Bang Mod, Thung Khru , Bangkok 10140, Thailand
e-mail: poom.kum@kmutt.ac.th

S. Plubtieng
Department of Mathematics, Faculty of Science, Naresuan University, Phitsanulok 65000, Thailand
e-mail: somyotp@nu.ac.th

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*, Lecture Notes in Electrical Engineering 275, DOI: 10.1007/978-94-007-7684-5_29,

$\|x-y\|,\ \forall x, y \in C$ and $T_n x = \frac{1}{n+1}\sum_{i=0}^{n} T^i x$ is called Cesàro means. A mapping f is called *weakly contractive* on a closed convex set C in the Banach space E if there exists $\varphi : [0,\infty) \to [0,\infty)$ is a continuous and strictly increasing function such that φ is positive on $(0,\infty)$, $\varphi(0) = 0$, $\lim_{t\to\infty} \varphi(t) = \infty$ and $x, y \in C$

$$\|f(x) - f(y)\| \le \|x - y\| - \varphi(\|x - y\|). \tag{1}$$

If $\varphi(t) = (1-k)t$, then f is called to be *contractive* with the contractive coefficient k. If $\varphi(t) = 0$, then f is said to be *nonexpansive*.

For $q > 1$, the *generalized duality mapping* $J_q : E \to 2^{E^*}$ is defined by

$$J_q(x) = \{f \in E^* : \langle x, f\rangle = \|x\|^q, \|f\| = \|x\|^{q-1}\}$$

for all $x \in E$. In particular, if $q = 2$, the mapping J_2 is called the *normalized duality mapping* and usually write $J_2 = J$. Further, we have the following properties of the generalized duality mapping J_q: (1) $J_q(x) = \|x\|^{q-2} J_2(x)$ for all $x \in E$ with $x \neq 0$; (2) $J_q(tx) = t^{q-1} J_q(x)$ for all $x \in E$ and $t \in [0,\infty)$; and (3) $J_q(-x) = -J_q(x)$ for all $x \in E$. It is known that if X is smooth, then J is single-valued, which is denoted by j. Recall that the duality mapping j is said to be weakly sequentially continuous if for each $x_n \to x$ weakly, we have $j(x_n) \to j(x)$ weakly-*. We know that if X admits a weakly sequentially continuous duality mapping, then X is smooth (for the details, see [1–3]).

Let $U = \{x \in E : \|x\| = 1\}$. A Banach space E is said to *uniformly convex* if, for any $\epsilon \in (0, 2]$, there exists $\delta > 0$ such that, for any $x, y \in U$, $\|x - y\| \ge \epsilon$ implies $\|\frac{x+y}{2}\| \le 1 - \delta$. It is known that a uniformly convex Banach space is reflexive and strictly convex. A Banach space E is said to be *smooth* if the limit $\lim_{t\to 0} \frac{\|x+ty\|-\|x\|}{t}$ exists for all $x, y \in U$. It is also said to be *uniformly smooth* if the limit is attained uniformly for $x, y \in U$. The *modulus of smoothness* of E is defined by

$$\rho(\tau) = \sup\left\{\frac{1}{2}(\|x + y\| + \|x - y\|) - 1 : x, y \in E, \|x\| = 1, \|y\| = \tau\right\},$$

where $\rho : [0,\infty) \to [0,\infty)$ is a function. It is known that E is uniformly smooth if and only if $\lim_{\tau\to 0} \frac{\rho(\tau)}{\tau} = 0$. Let q be a fixed real number with $1 < q \le 2$. A Banach space E is said to be *q-uniformly smooth* if there exists a constant $c > 0$ such that $\rho(\tau) \le c\tau^q$ for all $\tau > 0$: see, for instance, [1, 4].

We note that E is a uniformly smooth Banach space if and only if J_q is single-valued and uniformly continuous on any bounded subset of E. Typical examples of both uniformly convex and uniformly smooth Banach spaces are L^p, where $p > 1$. More precisely, L^p is $\min\{p, 2\}$-uniformly smooth for every $p > 1$. Note also that no Banach space is q-uniformly smooth for $q > 2$; see [1, 5] for more details.

Recall that an operator $A : C \to E$ is said to be *accretive* if there exists $j(x-y) \in J(x - y)$ such that

$$\langle Ax - Ay, j(x - y)\rangle \geq 0$$

for all $x, y \in C$. A mapping $A : C \to E$ is said to be β-*strongly accretive* if there exists a constant $\beta > 0$ such that

$$\langle Ax - Ay, j(x - y)\rangle \geq \beta\|x - y\|^2, \quad \forall x, y \in C.$$

An operator $A : C \to E$ is said to be β-*inverse strongly accretive* if, for any $\beta > 0$,

$$\langle Ax - Ay, j(x - y)\rangle \geq \beta\|Ax - Ay\|^2$$

for all $x, y \in C$. Evidently, the definition of the inverse strongly accretive operator is based on that of the inverse strongly monotone operator.

To convey an idea of the *variational inequality*, let C be a closed and convex set in a real Hilbert space H. For a given operator A, we consider the problem of finding $x^* \in C$ such that

$$\langle Ax^*, x - x^*\rangle \geq 0$$

for all $x \in C$, which is known as the variational inequality, introduced and studied by Stampacchia [6] in 1964 in the field of potential theory. In 2006, Aoyama et al. [4] first considered the following generalized variational inequality problem in a smooth Banach space. Let A be an accretive operator of C into E. Find a point $x \in C$ such that

$$\langle Ax, j(y - x)\rangle \geq 0 \tag{2}$$

for all $y \in C$. The set of solution of (2) is denoted by $VI(C, A)$. This problem is connected with the fixed point problem for nonlinear mappings, the problem of finding a zero point of an accretive operator and so on. For the problem of finding a zero point of an accretive operator by the proximal point algorithm, see Kamimura and Takahashi [7, 8]. In order to find a solution of the variational inequality (2), Aoyama et al. [4] proved the strong convergence theorem in the framework of Banach spaces which is generalized Iiduka et al. [9] from Hilbert spaces.

Motivated by Aoyama et al. [4] and also Ceng et al. [10], Qin et al. [11] and Yao et al. [3] first considered the following a *new general system of variational inequalities* in Banach spaces:

Let $A : C \to E$ be an β-inverse strongly accretive mapping. Find $(x^*, y^*) \in C \times C$ such that

$$\begin{cases} \langle \lambda Ay^* + x^* - y^*, j(x - x^*)\rangle \geq 0 \ \ \forall x \in C, \\ \langle \mu Ax^* + y^* - x^*, j(x - y^*)\rangle \geq 0 \ \ \forall x \in C. \end{cases} \tag{3}$$

Let C be nonempty closed convex subset of a real Banach space E. For given two operators $A, B : C \to E$, consider the problem of finding $(x^*, y^*) \in C \times C$ such that

$$\begin{cases} \langle \lambda A y^* + x^* - y^*, j(x - x^*) \rangle \geq 0 \ \ \forall x \in C, \\ \langle \mu B x^* + y^* - x^*, j(x - y^*) \rangle \geq 0 \ \ \forall x \in C, \end{cases} \tag{4}$$

where λ and μ are two positive real numbers. This system is called the *general system of variational inequalities* in a real Banach spaces. If we add up the requirement that $A = B$, then the problem (4) is reduced to the system (3).

By the following the general system of variational inequalities, we extend into the *general system of finite variational inequalities* is to find $(x_1^*, x_2^*, \ldots, x_M^*) \in C \times C \times \cdots \times C$ and is defined by

$$\begin{cases} \langle \lambda_M A_M x_M^* + x_1^* - x_M^*, j(x - x_1^*) \rangle \geq 0, \ \ \forall x \in C, \\ \langle \lambda_{M-1} A_{M-1} x_{M-1}^* + x_M^* - x_{M-1}^*, j(x - x_M^*) \rangle \geq 0, \ \ \forall x \in C, \\ \vdots \\ \langle \lambda_2 A_2 x_2^* + x_3^* - x_2^*, j(x - x_3^*) \rangle \geq 0, \ \ \forall x \in C, \\ \langle \lambda_1 A_1 x_1^* + x_2^* - x_1^*, j(x - x_2^*) \rangle \geq 0, \ \ \forall x \in C, \end{cases} \tag{5}$$

where $\{A_l\}_{l=1}^M : C \to E$ is a family of mappings, $\lambda_l \geq 0, l \in \{1, 2, \ldots, M\}$. The set of solution of (5) is denoted by $GSVI(C, A_l)$. In particular, if $M = 2$, $A_1 = B$, $A_2 = A$, $\lambda_1 = \mu$, $\lambda_2 = \lambda$, $x_1^* = x^*$ and $x_2^* = y^*$, then the problem (5) is reduced to the problem (4).

In 1997, Shimizu and Takahashi [12] originally studied the convergence of an iteration process $\{x_n\}$ for a family of nonexpansive mappings in the framework of a real Hilbert space. They restate the sequence $\{x_n\}$ as follows:

$$x_{n+1} = \alpha_n x + (1 - \alpha_n) \frac{1}{n+1} \sum_{j=0}^{n} T^j x_n, \ \ for \ \ n = 0, 1, 2, \ldots \tag{6}$$

where x_0 and x are all elements of C and α_n is an appropriate in $[0, 1]$. They proved that $\{x_n\}$ converges strongly to an element of fixed point of T which is the nearest to x.

In this paper, motivated and inspired by the idea of Katchang et al. [13], Yao et al. [3], Ceng et al. [10] and Witthayarat et al. [14]. We introduce a new iterative scheme with weak contraction for finding solutions of a new general system of finite variational inequalities (5) for finite different inverse-strongly accretive operators and solutions of fixed point problems for nonexpansive mapping in a Banach space. Consequently, we obtain new strong convergence theorems for fixed point problems which solves the general system of variational inequalities (4). Moreover, using the above theorem, we can apply to finding solutions of zeros of accretive operators and the class of k-strictly pseudocontractive mappings. This paper extend

and improve the corresponding results of Yao et al. [3], Ceng et al. [10], Shimizu and Takahashi [12] and many authors.

2 Preliminaries

We always assume that E is a real Banach space and C be a nonempty closed convex subset of E. Let D be a subset of C and $Q : C \to D$. Then Q is said to *sunny* if

$$Q(Qx + t(x - Qx)) = Qx,$$

whenever $Qx + t(x - Qx) \in C$ for $x \in C$ and $t \geq 0$. A subset D of C is said to be a *sunny nonexpansive retract* of C if there exists a sunny nonexpansive retraction Q of C onto D. A mapping $Q : C \to C$ is called a *retraction* if $Q^2 = Q$. If a mapping $Q : C \to C$ is a retraction, then $Qz = z$ for all z is in the range of Q. For example, see [4, 15] for more details. The following result describes a characterization of sunny nonexpansive retractions on a smooth Banach space.

A Banach space E is said to satisfy *Opial's condition* if for any sequence $\{x_n\}$ in E, $x_n \rightharpoonup x (n \to \infty)$ implies

$$\limsup_{n\to\infty} \|x_n - x\| < \limsup_{n\to\infty} \|x_n - y\|, \forall y \in E \text{ with } x \neq y.$$

By [16, Theorem 1], it is well known that if E admits a weakly sequentially continuous duality mapping, then E satisfies Opial's condition, and E is smooth.

Proposition 1 ([17]) *Let E be a smooth Banach space and let C be a nonempty subset of E. Let $Q : E \to C$ be a retraction and let J be the normalized duality mapping on E. Then the following are equivalent:*

(1) Q is sunny and nonexpansive;
(2) $\|Qx - Qy\|^2 \leq \langle x - y, J(Qx - Qy)\rangle, \forall x, y \in E$;
(3) $\langle x - Qx, J(y - Qx)\rangle \leq 0, \forall x \in E, y \in C$.

Proposition 2 ([18]) *Let C be a nonempty closed convex subset of a uniformly convex and uniformly smooth Banach space E and let T be a nonexpansive mapping of C into itself with $F(T) \neq \emptyset$. Then the set $F(T)$ is a sunny nonexpansive retract of C.*

We need the following lemmas for proving our main results.

Lemma 1 ([5]) *Let E be a real 2-uniformly smooth Banach space with the best smooth constant K. Then the following inequality holds:*

$$\|x + y\|^2 \leq \|x\|^2 + 2\langle y, Jx\rangle + 2\|Ky\|^2, \quad \forall x, y \in E.$$

Lemma 2 ([19]) *Let $\{x_n\}$ and $\{y_n\}$ be bounded sequences in a Banach space X and let $\{\beta_n\}$ be a sequence in $[0, 1]$ with $0 < \liminf_{n\to\infty} \beta_n \leq \limsup_{n\to\infty} \beta_n < 1$. Suppose $x_{n+1} = (1-\beta_n)y_n + \beta_n x_n$ for all integers $n \geq 0$ and $\limsup_{n\to\infty}(\|y_{n+1} - y_n\| - \|x_{n+1} - x_n\|) \leq 0$. Then, $\lim_{n\to\infty} \|y_n - x_n\| = 0$.*

Lemma 3 (*Lemma 2.2 in*[20]) *Let $\{a_n\}$ and $\{b_n\}$ be two nonnegative real number sequences and $\{\alpha_n\}$ a positive real number sequence satisfying the conditions: $\sum_{n=1}^{\infty} \alpha_n = \infty$ and $\lim_{n\to\infty} \frac{b_n}{\alpha_n} = 0$. Let the recursive inequality*

$$a_{n+1} \leq a_n - \alpha_n \varphi(a_n) + b_n, \ \ n \geq 0$$

where $\varphi(a)$ is a continuous and strict increasing function for all $a \geq 0$ with $\varphi(0) = 0$. Then $\lim_{n\to\infty} a_n = 0$.

Lemma 4 ([21]) *Let E be a uniformly convex Banach space and $B_r(0) := \{x \in E : \|x\| \leq r\}$ be a closed ball of E. Then there exists a continuous strictly increasing convex function $g : [0, \infty) \to [0, \infty)$ with $g(0) = 0$ such that*

$$\|\lambda x + \mu y + \gamma z\|^2 \leq \lambda\|x\|^2 + \mu\|y\|^2 + \gamma\|z\|^2 - \lambda\mu g(\|x - y\|)$$

for all $x, y, z \in B_r(0)$ and $\lambda, \mu, \gamma \in [0, 1]$ with $\lambda + \mu + \gamma = 1$.

Lemma 5 ([22]) *Let C be a nonempty bounded closed convex subset of a uniformly convex Banach space E and let T be nonexpansive mapping of C into itself. If $\{x_n\}$ is a sequence of C such that $x_n \to x$ weakly and $x_n - Tx_n \to 0$ strongly, then x is s fixed point of T.*

Lemma 6 ([23]) *Let C be a nonempty bounded closed convex subset of a uniformly convex Banach space E and $T : C \to C$ a nonexpansive mapping. For each $x \in C$ and the Cesàro means $T_n x = \frac{1}{n+1}\sum_{i=0}^{n} T^i x$, then $\limsup_{n\to\infty} \|T_n x - T(T_n x)\| = 0$.*

Lemma 7 ([3, Lemma 3.1]*; see also* [4, Lemma 8]) *Let C be a nonempty closed convex subset of a real 2-uniformly smooth Banach space E. Let the mapping $A : C \to E$ be β-inverse-strongly accretive. Then, we have*

$$\|(I - \lambda A)x - (I - \lambda A)y\|^2 \leq \|x - y\|^2 + 2\lambda(\lambda K^2 - \beta)\|Ax - Ay\|^2.$$

If $\beta \geq \lambda K^2$, then $I - \lambda A$ is nonexpansive.

Lemma 8 *Let C be a nonempty closed convex subset of a real 2-uniformly smooth Banach space E. Let Q_C be the sunny nonexpansive retraction from E onto C. Let the mapping $A_l : C \to H$ be a β_l-inverse-strongly accretive such that $\beta_l \geq \lambda_l K^2$ where $l \in \{1, 2, \ldots, M\}$. If $\mathcal{Q} : C \to C$ be a mapping defined by*

$$\mathcal{Q}(x) = Q_C(I-\lambda_M A_M)Q_C(I-\lambda_{M-1}A_{M-1})\ldots Q_C(I-\lambda_2 A_2)Q_C(I-\lambda_1 A_1)x, \forall x \in C,$$

then $\mathscr{Q}$ is nonexpansive.

Proof Taking $\mathscr{Q}_C^l = Q_C(I - \lambda_l A_l)Q_C(I - \lambda_{l-1}A_{l-1})\dots Q_C(I - \lambda_2 A_2)Q_C(I - \lambda_1 A_1)$, $l \in \{1, 2, 3, \dots, M\}$ and $\mathscr{Q}_C^0 = I$, where I is the identity mapping on H. Then we have $\mathscr{Q} = \mathscr{Q}_C^M$. For any $x, y \in C$, we have

$$\begin{aligned}
\|\mathscr{Q}(x) - \mathscr{Q}(y)\| &= \|\mathscr{Q}_C^M x - \mathscr{Q}_C^M y\| \\
&= \|Q_C(I - \lambda_M A_M)\mathscr{Q}_C^{M-1}x - Q_C(I - \lambda_M A_M)\mathscr{Q}_C^{M-1}y\| \\
&\le \|(I - \lambda_M A_M)\mathscr{Q}_C^{M-1}x - (I - \lambda_M A_M)\mathscr{Q}_C^{M-1}y\| \\
&\le \|\mathscr{Q}_C^{M-1}x - \mathscr{Q}_C^{M-1}y\| \\
&\vdots \\
&\le \|\mathscr{Q}_C^0 x - \mathscr{Q}_C^0 y\| \\
&= \|x - y\|.
\end{aligned}$$

Therefore $\mathscr{Q}$ is nonexpansive. □

Lemma 9 *Let C be a nonempty closed convex subset of a real smooth Banach space E. Let Q_C be the sunny nonexpansive retraction from E onto C. Let $A_l : C \to H$ be nonlinear mapping, where $l \in \{1, 2, \dots, M\}$. For $x_l^* \in C, l \in \{1, 2, \dots, M\}$, $(x_1^*, x_2^*, \dots, x_M^*)$ is a solution of problem (5) if and only if*

$$\begin{cases}
x_1^* = Q_C(I - \lambda_M A_M)x_M^* \\
x_2^* = Q_C(I - \lambda_1 A_1)x_1^* \\
x_3^* = Q_C(I - \lambda_2 A_2)x_2^* \\
\vdots \\
x_M^* = Q_C(I - \lambda_{M-1} A_{M-1})x_{M-1}^*,
\end{cases} \tag{7}$$

that is

$$x_1^* = Q_C(I - \lambda_M A_M)Q_C(I - \lambda_{M-1}A_{M-1})\dots Q_C(I - \lambda_2 A_2)Q_C(I - \lambda_1 A_1)x_1^*.$$

Proof From (5), we rewrite as

$$\begin{cases}
\langle x_1^* - (x_M^* - \lambda_M A_M x_M^*), j(x - x_1^*)\rangle \ge 0, \ \forall x \in C, \\
\langle x_M^* - (x_{M-1}^* - \lambda_{M-1} A_{M-1} x_{M-1}^*), j(x - x_M^*)\rangle \ge 0, \ \forall x \in C, \\
\vdots \\
\langle x_3^* - (x_2^* - \lambda_2 A_2 x_2^*), j(x - x_3^*)\rangle \ge 0, \ \forall x \in C, \\
\langle x_2^* - (x_1^* - \lambda_1 A_1 x_1^*), j(x - x_2^*)\rangle \ge 0, \ \forall x \in C.
\end{cases} \tag{8}$$

Using Proposition 1 (3), the system (8) equivalent to (7). □

Throughout this paper, the set of fixed points of the mapping $\mathscr{Q}$ is denoted by $F(\mathscr{Q})$.

3 Main Results

In this section, we prove a strong convergence theorem. The next result states the main result of this work.

Theorem 1 *Let E be a uniformly convex and 2-uniformly smooth Banach space which admits a weakly sequentially continuous duality mapping and C be a nonempty closed convex subset of E. Let $T^i : C \to C$ be a nonexpansive mappings for all $i = 1, 2, 3, \ldots, n$ and Q_C be a sunny nonexpansive retraction from E onto C. Let $A_l : C \to E$ be a β_l-inverse-strongly accretive such that $\beta_l \geq \lambda_l K^2$ where $l \in \{1, 2, \ldots, M\}$ and K be the best smooth constant. Let f be a weakly contractive of C into itself with function φ. Suppose $\mathscr{F} := F(\mathscr{Q}) \cap \left(\cap_{i=1}^{n} F(T^i)\right) \neq \emptyset$ where $\mathscr{Q}$ defined by Lemma 8. For arbitrary given $x_0 = x \in C$, the sequence $\{x_n\}$ generated by*

$$\begin{cases} y_n = Q_C(I - \lambda_M A_M)Q_C(I - \lambda_{M-1}A_{M-1}) \ldots Q_C(I - \lambda_2 A_2)Q_C(I - \lambda_1 A_1)x_n, \\ x_{n+1} = \alpha_n f(x_n) + \beta_n x_n + \gamma_n \frac{1}{n+1} \sum_{i=0}^{n} T^i y_n. \end{cases} \tag{9}$$

where the sequences $\{\alpha_n\}$, $\{\beta_n\}$ and $\{\gamma_n\}$ in $(0, 1)$ satisfy $\alpha_n + \beta_n + \gamma_n = 1$, $n \geq 1$ and $\lambda_l, l = 1, 2, \ldots, M$ are positive real numbers. The following conditions are satisfied:
(C1). $\lim_{n\to\infty} \alpha_n = 0$ *and* $\sum_{n=0}^{\infty} \alpha_n = \infty$;
(C2). $0 < \liminf_{n\to\infty} \beta_n \leq \limsup_{n\to\infty} \beta_n < 1$.
Then $\{x_n\}$ converges strongly to $\bar{x}_1 = Q_{\mathscr{F}} f(\bar{x}_1)$ and $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_M)$ is a solution of the problem (5), where $Q_{\mathscr{F}}$ is the sunny nonexpansive retraction of C onto $\mathscr{F}$.

Proof First, we prove that $\{x_n\}$ is bounded. Let $p \in \mathscr{F}$, taking

$$\mathscr{Q}_C^l = Q_C(I - \lambda_l A_l)Q_C(I - \lambda_{l-1}A_{l-1}) \ldots Q_C(I - \lambda_2 A_2)Q_C(I - \lambda_1 A_1),$$

$l \in \{1, 2, 3, \ldots, M\}$, $\mathscr{Q}_C^0 = I$, where I is the identity mapping on E. From the definition of Q_C is nonexpansive then $\mathscr{Q}_C^l, l \in \{1, 2, 3, \ldots, M\}$ also. We note that

$$\|y_n - p\| = \|\mathscr{Q}_C^l x_n - \mathscr{Q}_C^l p\| \leq \|x_n - p\|. \tag{10}$$

On the other hand, let $T_n = \frac{1}{n+1} \sum_{i=0}^{n} T^i$, we have

$$\|T_n x - T_n y\| = \left\| \frac{1}{n+1} \sum_{i=0}^{n} T^i x - \frac{1}{n+1} \sum_{i=0}^{n} T^i y \right\|$$

$$\begin{aligned}
&\leq \frac{1}{n+1}\sum_{i=0}^{n}\|T^i x - T^i y\| \\
&\leq \frac{1}{n+1}\sum_{i=0}^{n}\|x-y\| \\
&= \frac{n+1}{n+1}\|x-y\| \\
&= \|x-y\|, \qquad (11)
\end{aligned}$$

which implies that T_n is nonexpansive. Since $p \in \mathscr{F}$, we have $T_n p = \frac{1}{n+1}\sum_{i=0}^{n} T^i p = \frac{1}{n+1}\sum_{i=0}^{n} p = p$ for all $x, y \in C$. It follows from (9), (10) and (11), we also have

$$\begin{aligned}
\|x_{n+1}-p\| &= \|\alpha_n f(x_n) + \beta_n x_n + \gamma_n T_n y_n - p\| \\
&\leq \alpha_n\|f(x_n)-p\| + \beta_n\|x_n-p\| + \gamma_n\|T_n y_n - p\| \\
&\leq \alpha_n\big[\|x_n-p\| - \varphi(\|x_n-p\|)\big] + \alpha_n\|f(p)-p\| + \beta_n\|x_n-p\| \\
&\quad +\gamma_n\|y_n-p\| \\
&\leq \|x_n-p\| - \alpha_n\varphi(\|x_n-p\|) + \alpha_n\|f(p)-p\| \\
&\leq \max\{\|x_1-p\|, \varphi(\|x_1-p\|), \|f(p)-p\|\}. \qquad (12)
\end{aligned}$$

This implies that $\{x_n\}$ is bounded, so are $\{f(x_n)\}$, $\{y_n\}$, and $\{T_n y_n\}$.

Next, we show that $\lim_{n\to\infty}\|x_{n+1}-x_n\| = 0$. Notice that

$$\begin{aligned}
\|y_{n+1}-y_n\| &= \|\mathscr{Q}_C^M x_{n+1} - \mathscr{Q}_C^M x_n\| \\
&= \|Q_C(I-\lambda_M A_M)\mathscr{Q}_C^{M-1}x_{n+1} - Q_C(I-\lambda_M A_M)\mathscr{Q}_C^{M-1}x_n\| \\
&\leq \|(I-\lambda_M A_M)\mathscr{Q}_C^{M-1}x_{n+1} - (I-\lambda_M A_M)\mathscr{Q}_C^{M-1}x_n\| \\
&\leq \|\mathscr{Q}_C^{M-1}x_{n+1} - \mathscr{Q}_C^{M-1}x_n\| \\
&\vdots \\
&\leq \|\mathscr{Q}_C^0 x_{n+1} - \mathscr{Q}_C^0 x_n\| \\
&= \|x_{n+1}-x_n\|
\end{aligned}$$

and

$$\begin{aligned}
\|T_{n+1}y_{n+1} - T_n y_n\| &\leq \|T_{n+1}y_{n+1} - T_{n+1}y_n\| + \|T_{n+1}y_n - T_n y_n\| \\
&\leq \|y_{n+1}-y_n\| + \Big\|\frac{1}{n+2}\sum_{i=0}^{n+1}T^i y_n - \frac{1}{n+1}\sum_{i=0}^{n}T^i y_n\Big\| \\
&= \|y_{n+1}-y_n\|
\end{aligned}$$

$$
\begin{aligned}
&+\Big\|\frac{1}{n+2}\sum_{i=0}^{n}T^{i}y_n+\frac{1}{n+2}T^{n+1}y_n-\frac{1}{n+1}\sum_{i=0}^{n}T^{i}y_n\Big\|\\
&=\|y_{n+1}-y_n\|+\Big\|-\frac{1}{(n+1)(n+2)}\sum_{i=0}^{n}T^{i}y_n+\frac{1}{n+2}T^{n+1}y_n\Big\|\\
&\leq\|y_{n+1}-y_n\|+\frac{1}{(n+1)(n+2)}\sum_{i=0}^{n}\|T^{i}y_n\|+\frac{1}{n+2}\|T^{n+1}y_n\|\\
&\leq\|y_{n+1}-y_n\|+\frac{1}{(n+1)(n+2)}\sum_{i=0}^{n}(\|T^{i}y_n-T^{i}p\|+\|p\|)\\
&\quad+\frac{1}{n+2}(\|T^{n+1}y_n-T^{n+1}p\|+\|p\|)\\
&\leq\|y_{n+1}-y_n\|+\frac{1}{(n+1)(n+2)}\sum_{i=0}^{n}(\|y_n-p\|+\|p\|)\\
&\quad+\frac{1}{n+2}(\|y_n-p\|+\|p\|)\\
&\leq\|y_{n+1}-y_n\|+\frac{n+1}{(n+1)(n+2)}(\|y_n-p\|+\|p\|)\\
&\quad+\frac{1}{n+2}\|y_n-p\|+\frac{1}{n+2}\|p\|\\
&=\|y_{n+1}-y_n\|+\frac{2}{n+2}\|y_n-p\|+\frac{2}{n+2}\|p\|\\
&\leq\|x_{n+1}-x_n\|+\frac{2}{n+2}\|y_n-p\|+\frac{2}{n+2}\|p\|.
\end{aligned}
$$

Setting $x_{n+1}=(1-\beta_n)z_n+\beta_n x_n$ for all $n\geq 0$, we see that $z_n=\frac{x_{n+1}-\beta_n x_n}{1-\beta_n}$, then we have

$$
\begin{aligned}
\|z_{n+1}-z_n\|&=\|\frac{x_{n+2}-\beta_{n+1}x_{n+1}}{1-\beta_{n+1}}-\frac{x_{n+1}-\beta_n x_n}{1-\beta_n}\|\\
&=\|\frac{\alpha_{n+1}f(x_{n+1})+\gamma_{n+1}T_{n+1}y_{n+1}}{1-\beta_{n+1}}-\frac{\alpha_n f(x_n)+\gamma_n T_n y_n}{1-\beta_n}\|\\
&=\|\frac{\alpha_{n+1}f(x_{n+1})+\gamma_{n+1}T_{n+1}y_{n+1}}{1-\beta_{n+1}}-\frac{\alpha_{n+1}f(x_n)}{1-\beta_{n+1}}+\frac{\alpha_{n+1}f(x_n)}{1-\beta_{n+1}}\\
&\quad-\frac{\gamma_{n+1}T_n y_n}{1-\beta_{n+1}}+\frac{\gamma_{n+1}T_n y_n}{1-\beta_{n+1}}-\frac{\alpha_n f(x_n)+\gamma_n T_n y_n}{1-\beta_n}\|\\
&=\|\frac{\alpha_{n+1}}{1-\beta_{n+1}}(f(x_{n+1})-f(x_n))+\frac{\gamma_{n+1}}{1-\beta_{n+1}}(T_{n+1}y_{n+1}-T_n y_n)\\
&\quad+(\frac{\alpha_{n+1}}{1-\beta_{n+1}}-\frac{\alpha_n}{1-\beta_n})f(x_n)+(\frac{\gamma_{n+1}}{1-\beta_{n+1}}-\frac{\gamma_n}{1-\beta_n})T_n y_n\|\\
&\leq\frac{\alpha_{n+1}}{1-\beta_{n+1}}\|f(x_{n+1})-f(x_n)\|+\frac{\gamma_{n+1}}{1-\beta_{n+1}}\|T_{n+1}y_{n+1}-T_n y_n\|\\
&\quad+|\frac{\alpha_{n+1}}{1-\beta_{n+1}}-\frac{\alpha_n}{1-\beta_n}|\|f(x_n)\|
\end{aligned}
$$

$$
\begin{aligned}
&+|\frac{1-\beta_{n+1}-\alpha_{n+1}}{1-\beta_{n+1}}-\frac{1-\beta_n-\alpha_n}{1-\beta_n}|\|T_n y_n\| \\
&\leq \frac{\alpha_{n+1}}{1-\beta_{n+1}}\Big[\|x_{n+1}-x_n\|-\varphi(\|x_{n+1}-x_n\|)\Big] \\
&\quad+\frac{\gamma_{n+1}}{1-\beta_{n+1}}\|T_{n+1}y_{n+1}-T_n y_n\|+|\frac{\alpha_{n+1}}{1-\beta_{n+1}}-\frac{\alpha_n}{1-\beta_n}|\|f(x_n)\| \\
&\quad+|\frac{\alpha_{n+1}}{1-\beta_{n+1}}-\frac{\alpha_n}{1-\beta_n}|\|T_n y_n\| \\
&\leq \frac{\alpha_{n+1}}{1-\beta_{n+1}}\|x_{n+1}-x_n\|+\frac{\gamma_{n+1}}{1-\beta_{n+1}}\|T_{n+1}y_{n+1}-T_n y_n\| \\
&\quad+|\frac{\alpha_{n+1}}{1-\beta_{n+1}}-\frac{\alpha_n}{1-\beta_n}|(\|f(x_n)\|+\|T_n y_n\|) \\
&\leq \frac{\alpha_{n+1}}{1-\beta_{n+1}}\|x_{n+1}-x_n\|+\|T_{n+1}y_{n+1}-T_n y_n\| \\
&\quad+|\frac{\alpha_{n+1}}{1-\beta_{n+1}}-\frac{\alpha_n}{1-\beta_n}|(\|f(x_n)\|+\|T_n y_n\|) \\
&\leq \frac{\alpha_{n+1}}{1-\beta_{n+1}}\|x_{n+1}-x_n\|+\|x_{n+1}-x_n\|+\frac{2}{n+2}\|y_n-p\|+\frac{2}{n+2}\|p\| \\
&\quad+|\frac{\alpha_{n+1}}{1-\beta_{n+1}}-\frac{\alpha_n}{1-\beta_n}|(\|f(x_n)\|+\|T_n y_n\|).
\end{aligned}
$$

Therefore

$$
\begin{aligned}
\|z_{n+1}-z_n\|-\|x_{n+1}-x_n\| \leq{}& \frac{\alpha_{n+1}}{1-\beta_{n+1}}\|x_{n+1}-x_n\|+\frac{2}{n+2}\|y_n-p\|+\frac{2}{n+2}\|p\| \\
&+|\frac{\alpha_{n+1}}{1-\beta_{n+1}}-\frac{\alpha_n}{1-\beta_n}|(\|f(x_n)\|+\|T_n y_n\|).
\end{aligned}
$$

It follows from the condition (C1) and (C2) and $n \to \infty$, which implies that

$$\limsup_{n\to\infty}(\|z_{n+1}-z_n\|-\|x_{n+1}-x_n\|) \leq 0.$$

Applying Lemma 2, we obtain $\lim_{n\to\infty}\|z_n - x_n\| = 0$ and also

$$\|x_{n+1}-x_n\| = (1-\beta_n)\|z_n - x_n\| \to 0$$

as $n \to \infty$. Therefore, we have

$$\lim_{n\to\infty}\|x_{n+1}-x_n\| = 0. \tag{13}$$

Next, we show that $\lim_{n\to\infty}\|T_n y_n - y_n\| = 0$. Since $p \in \mathscr{F}$, from Lemma 4, we obtain

$$\|x_{n+1}-p\|^2 = \|\alpha_n f(x_n)+\beta_n x_n+\gamma_n T_n y_n - p\|^2$$

$$\begin{aligned}
&\leq \alpha_n \|f(x_n) - p\|^2 + (1 - \alpha_n - \gamma_n)\|x_n - p\|^2 + \gamma_n \|y_n - p\|^2 \\
&= \alpha_n \|f(x_n) - p\|^2 + (1 - \alpha_n)\|x_n - p\|^2 - \gamma_n(\|x_n - p\|^2 - \|y_n - p\|^2) \\
&= \alpha_n \|f(x_n) - p\|^2 + (1 - \alpha_n)\|x_n - p\|^2 \\
&\quad -\gamma_n(\|x_n - p\| - \|y_n - p\|)(\|x_n - p\| + \|y_n - p\|) \\
&\leq \alpha_n \|f(x_n) - p\|^2 + \|x_n - p\|^2 - \gamma_n \|x_n - y_n\|^2.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\gamma_n \|x_n - y_n\|^2 &\leq \alpha_n \|f(x_n) - p\|^2 + \|x_n - p\|^2 - \|x_{n+1} - p\|^2 \\
&\leq \alpha_n \|f(x_n) - p\|^2 + (\|x_n - p\| + \|x_{n+1} - p\|)\|x_n - x_{n+1}\|.
\end{aligned}$$

From the condition (C1) and (13), this implies that $\|x_n - y_n\| \to 0$ as $n \to \infty$. Now, we note that

$$\begin{aligned}
\|x_n - T_n y_n\| &\leq \|x_n - x_{n+1}\| + \|x_{n+1} - T_n y_n\| \\
&= \|x_n - x_{n+1}\| + \|\alpha_n f(x_n) + \beta_n x_n + \gamma_n T_n y_n - T_n y_n\| \\
&= \|x_n - x_{n+1}\| + \|\alpha_n (f(x_n) - T_n y_n) + \beta_n (x_n - T_n y_n)\| \\
&\leq \|x_n - x_{n+1}\| + \alpha_n \|f(x_n) - T_n y_n\| + \beta_n \|x_n - T_n y_n\|.
\end{aligned}$$

Therefore, we get

$$\|x_n - T_n y_n\| \leq \frac{1}{1 - \beta_n}\|x_n - x_{n+1}\| + \frac{\alpha_n}{1 - \beta_n}\|f(x_n) - T_n y_n\|.$$

From the condition (C1), (C2) and (13), this implies that $\|x_n - T_n y_n\| \to 0$ as $n \to \infty$. Since

$$\|T_n y_n - y_n\| \leq \|T_n y_n - x_n\| + \|x_n - y_n\|,$$

and hence it follows that $\lim_{n\to\infty} \|T_n y_n - y_n\| = 0$.

Next, we prove that $z \in \mathscr{F} := F(\mathscr{Q}) \cap \left(\cap_{i=1}^{n} F(T^i)\right)$.

(a) First, we show that $z \in F(T_n) = \cap_{i=1}^{n} F(T^i)$. To show this, we choose a subsequence $\{y_{n_i}\}$ of $\{y_n\}$. Since $\{y_{n_i}\}$ is bounded, we have that a subsequence $\{y_{n_{i_j}}\}$ of $\{y_{n_i}\}$ converges weakly to z. We may assume without loss of generality that $y_{n_i} \rightharpoonup z$. Since $\|T_n y_n - y_n\| \to 0$, we obtain $T_n y_{n_i} \rightharpoonup z$. Then we can obtain $z \in \mathscr{F}$. Assume that $z \notin F(T_n)$. Since $y_{n_i} \rightharpoonup z$ and $T_n z \neq z$, from Opial's condition, we get

$$\begin{aligned}&\liminf_{i\to\infty}\|y_{n_i}-z\| < \liminf_{i\to\infty}\|y_{n_i}-T_nz\|\\ &\le \liminf_{i\to\infty}(\|y_{n_i}-T_ny_{n_i}\|+\|T_ny_{n_i}-T_nz\|)\\ &\le \liminf_{i\to\infty}\|y_{n_i}-z\|.\end{aligned}\tag{14}$$

This is a contradiction. Thus, we have $z \in F(T_n)$.

(b) Next, we show that $z \in F(\mathscr{Q})$. From Lemma 8, we know that $\mathscr{Q} = \mathscr{Q}_C^M$ is nonexpansive, it follows that

$$\|y_n - \mathscr{Q}y_n\| = \|\mathscr{Q}_C^M x_n - \mathscr{Q}_C^M y_n\| \le \|x_n - y_n\|.$$

Thus $\lim_{n\to\infty}\|y_n - \mathscr{Q}y_n\| = 0$. Since $\mathscr{Q}$ is nonexpansive, we get

$$\begin{aligned}\|x_n - \mathscr{Q}x_n\| &\le \|x_n - y_n\| + \|y_n - \mathscr{Q}y_n\| + \|\mathscr{Q}y_n - \mathscr{Q}x_n\|\\ &\le 2\|x_n - y_n\| + \|y_n - \mathscr{Q}y_n\|,\end{aligned}$$

and so

$$\lim_{n\to\infty}\|x_n - \mathscr{Q}x_n\| = 0.\tag{15}$$

By Lemma 5 and (15), we have $z \in F(\mathscr{Q})$. Therefore $z \in \mathscr{F}$.

Next, we show that $\limsup_{n\to\infty}\langle (f-I)\bar{x}_1, J(x_n - \bar{x}_1)\rangle \le 0$, where $\bar{x}_1 = Q_{\mathscr{F}}f(\bar{x}_1)$. Since $\{x_n\}$ is bounded, we can choose a sequence $\{x_{n_i}\}$ of $\{x_n\}$ which $x_{n_i} \rightharpoonup z$ such that

$$\limsup_{n\to\infty}\langle (f-I)\bar{x}_1, J(x_n - \bar{x}_1)\rangle = \lim_{i\to\infty}\langle (f-I)\bar{x}_1, J(x_{n_i} - \bar{x}_1)\rangle.\tag{16}$$

Now, from (16), Proposition 1 (3) and the weakly sequential continuity of the duality mapping J, we have

$$\begin{aligned}\limsup_{n\to\infty}\langle (f-I)\bar{x}_1, J(x_n - \bar{x}_1)\rangle &= \lim_{i\to\infty}\langle (f-I)\bar{x}_1, J(x_{n_i} - \bar{x}_1)\rangle\\ &= \langle (f-I)\bar{x}_1, J(z - \bar{x}_1)\rangle \le 0.\end{aligned}\tag{17}$$

From (13), it follows that

$$\limsup_{n\to\infty}\langle (f-I)\bar{x}_1, J(x_{n+1} - \bar{x}_1)\rangle \le 0.\tag{18}$$

Finally, we show that $\{x_n\}$ converges strongly to $\bar{x}_1 = Q_{\mathscr{F}}f(\bar{x}_1)$. We compute that

$$\begin{aligned}\|x_{n+1} - \bar{x}_1\|^2 &= \langle x_{n+1} - \bar{x}_1, J(x_{n+1} - \bar{x}_1)\rangle\\ &= \langle \alpha_n f(x_n) + \beta_n x_n + \gamma_n T_n y_n - \bar{x}_1, J(x_{n+1} - \bar{x}_1)\rangle\end{aligned}$$

$$
\begin{aligned}
&= \langle \alpha_n(f(x_n) - \bar{x}_1) + \beta_n(x_n - \bar{x}_1) + \gamma_n(T_n y_n - \bar{x}_1), J(x_{n+1} - \bar{x}_1)\rangle \\
&= \alpha_n\langle f(x_n) - f(\bar{x}_1), J(x_{n+1} - \bar{x}_1)\rangle + \alpha_n\langle f(\bar{x}_1) - \bar{x}_1, J(x_{n+1} - \bar{x}_1)\rangle \\
&\quad + \beta_n\langle x_n - \bar{x}_1, J(x_{n+1} - \bar{x}_1)\rangle + \gamma_n\langle T_n y_n - \bar{x}_1, J(x_{n+1} - \bar{x}_1)\rangle \\
&\leq \alpha_n\Big[\|x_n - \bar{x}_1\| - \varphi(\|x_n - \bar{x}_1\|)\Big]\|x_{n+1} - \bar{x}_1\| \\
&\quad + \alpha_n\langle f(\bar{x}_1) - \bar{x}_1, J(x_{n+1} - \bar{x}_1)\rangle + \beta_n\|x_n - \bar{x}_1\|\|x_{n+1} - \bar{x}_1\| \\
&\quad + \gamma_n\|y_n - \bar{x}_1\|\|x_{n+1} - \bar{x}_1\| \\
&\leq \alpha_n\|x_n - \bar{x}_1\|\|x_{n+1} - \bar{x}_1\| - \alpha_n\varphi(\|x_n - \bar{x}_1\|)\|x_{n+1} - \bar{x}_1\| \\
&\quad + \alpha_n\langle f(\bar{x}_1) - \bar{x}_1, J(x_{n+1} - \bar{x}_1)\rangle + \beta_n\|x_n - \bar{x}_1\|\|x_{n+1} - \bar{x}_1\| \\
&\quad + \gamma_n\|x_n - \bar{x}_1\|\|x_{n+1} - \bar{x}_1\| \\
&= \|x_n - \bar{x}_1\|\|x_{n+1} - \bar{x}_1\| - \alpha_n\varphi(\|x_n - \bar{x}_1\|)\|x_{n+1} - \bar{x}_1\| \\
&\quad + \alpha_n\langle f(\bar{x}_1) - \bar{x}_1, J(x_{n+1} - \bar{x}_1)\rangle \\
&= \frac{1}{2}\Big(\|x_n - \bar{x}_1\|^2 + \|x_{n+1} - \bar{x}_1\|^2\Big) - \alpha_n\varphi(\|x_n - \bar{x}_1\|)\|x_{n+1} - \bar{x}_1\| \\
&\quad + \alpha_n\langle f(\bar{x}_1) - \bar{x}_1, J(x_{n+1} - \bar{x}_1)\rangle.
\end{aligned}
$$

By (12) and since $\{x_{n+1} - \bar{x}_1\}$ bounded i.e., there exist $M > 0$ such that $\|x_{n+1} - \bar{x}_1\| \leq M$, which implies that

$$
\begin{aligned}
\|x_{n+1} - \bar{x}_1\|^2 \leq \|x_n - \bar{x}_1\|^2 &- 2\alpha_n M\varphi(\|x_n - \bar{x}_1\|) \\
&+ 2\alpha_n\langle f(\bar{x}_1) - \bar{x}_1, J(x_{n+1} - \bar{x}_1)\rangle. \qquad (19)
\end{aligned}
$$

Now, from (C1) and applying Lemma 3 to (19), we get $\|x_n - \bar{x}_1\| \to 0$ as $n \to \infty$. This completes the proof. □

Corollary 1 *Let E be a uniformly convex and 2-uniformly smooth Banach space which admits a weakly sequentially continuous duality mapping and C be a nonempty closed convex subset of E. Let $T^i : C \to C$ be a nonexpansive mappings for all $i = 1, 2, 3, \ldots, n$ and Q_C be a sunny nonexpansive retraction from E onto C. Let $A : C \to E$ be a β-inverse-strongly accretive such that $\beta \geq \lambda K^2$ and K be the best smooth constant. Let f be a contraction of C into itself with function φ. Suppose $\mathscr{F} := VI(C, A) \cap \left(\cap_{i=1}^{n} F(T^i)\right) \neq \emptyset$. For arbitrary given $x_0 = x \in C$, the sequence $\{x_n\}$ generated by*

$$
\begin{cases}
y_n = Q_C(I - \lambda A)x_n, \\
x_{n+1} = \alpha_n f(x_n) + \beta_n x_n + \gamma_n \frac{1}{n+1}\sum_{i=0}^{n} T^i y_n
\end{cases} \qquad (20)
$$

and satisfy the conditions (C1) and (C2) in the Theorem 1. *Then $\{x_n\}$ converges strongly to $\bar{x} = Q_{\mathscr{F}} f(\bar{x})$.*

Proof Put $M = 1$ and f is a contraction in the Theorem 1, we can conclude the desired conclusion easily. This completes the proof. □

Corollary 2 *Let E be a uniformly convex and 2-uniformly smooth Banach space which admits a weakly sequentially continuous duality mapping and C be a nonempty closed convex subset of E. Let $T : C \to C$ be a nonexpansive mapping and Q_C be a sunny nonexpansive retraction from E onto C. Let $A_l : C \to E$ be a β_l-inverse-strongly accretive such that $\beta_l \geq \lambda_l K^2$ where $l \in \{1, 2, \ldots, M\}$ and K be the best smooth constant. Let f be a contraction of C into itself. Suppose $\mathscr{F} := F(\mathscr{Q}) \cap F(T) \neq \emptyset$ where $\mathscr{Q}$ defined by Lemma 8. For arbitrary given $x_0 = x \in C$, the sequence $\{x_n\}$ generated by*

$$\begin{cases} y_n = Q_C(I - \lambda_M A_M) Q_C(I - \lambda_{M-1} A_{M-1}) \ldots Q_C(I - \lambda_2 A_2) Q_C(I - \lambda_1 A_1) x_n, \\ x_{n+1} = \alpha_n f(x_n) + \beta_n x_n + \gamma_n T y_n \end{cases} \tag{21}$$

and satisfy the conditions (C1) and (C2) in the Theorem 1. Then $\{x_n\}$ converges strongly to $\bar{x}_1 = Q_{\mathscr{F}} f(\bar{x}_1)$.

Proof Take $T^i = T$ for all i and f is a contraction in the Theorem 1, we can conclude the desired conclusion easily. This completes the proof. □

Acknowledgments This research was supported by the Commission on Higher Education, the Thailand Research Fund and the Rajamangala University of Technology Lanna Tak (Grant no. MRG5580233).

References

1. Takahashi W (2000) Nonlinear functional analysis. Fixed point theory and its applications. Yokohama Publishers, Yokohama
2. Takahashi W (2007) Viscosity approximation methods for resolvents of accretive operators in Banach spaces. J Fixed Point Theory Appl 1:135–147
3. Yao Y, Noor MA, Noor KI, Liou Y-C, Yaqoob H (2010) Modified extragradient methods for a system of variational inequalities in Banach spaces. Acta Applicandae Mathematicae 110:1211–1224
4. Aoyama K, Iiduka H, Takahashi W (2006) Weak convergence of an iterative sequence for accretive operators in Banach spaces. Fixed Point Theory Appl. Article ID 35390:1–13
5. Xu HK (1991) Inequalities in Banach spaces with applications. Nonlinear Anal 16:1127–1138
6. Stampacchi G (1964) Formes bilineaires coercivites sur les ensembles convexes. C R Acad Sci Paris 258:4413–4416
7. Kamimura S, Takahashi W (2000) Approximating solutions of maximal monotone operators in Hilbert space. J Approximation Theory 106:226–240
8. Kamimura S, Takahashi W (2000) Weak and strong convergence of solutions to accretive operator inclusions and applications. Set-Valued Anal 8:361–374
9. Iiduka H, Takahashi W, Toyoda M (2004) Approximation of solutions of variational inequalities for monotone mappings. Panam Mathl J 14:49–61
10. Ceng L-C, Wang C-Y, Yao J-C (2008) Strong convergence theorems by a relaxed extragradient method for a general system of variational inequalities. Math Methods Oper Res 67:375–390
11. Qin X, Cho SY, Kang SM (2009) Convergence of an iterative algorithm for systems of variational inequalities and nonexpansive mappings with applications. J Comput Appl Math 233:231–240
12. Shimizu T, Takahashi W (1997) Strong convergence to common fixed points of families of nonexpansive mappings. J Math Anal Appl 211:71–83

13. Katchang P, Plubtieng S, Kumam P (2013) A viscosity method for solving a general system of finite variational inequalities for finite accretive operators. Lecture notes in engineering and computer science, In: Proceedings of the international multiConference of engineers and computer scientists 2013, 13–15 Mar 2013, Hong Kong, pp 1076–1081
14. Witthayarat U, Cho YJ, Kumam P (2012) Convergence of an iterative algorithm for common solutions for zeros of maximal accretive operator with applications. J Appl Math, Article ID 185104 17 pages
15. Takahashi W (2000) Convex analysis and approximation fixed points. Yokohama Publishers, Yokohama
16. Gossez JP, Dozo EL (1972) Some geometric properties related to the fixed point theory for nonexpansive mappings. Pacific J Math 40:565–573
17. Reich S (1973) Asymptotic behavior of contractions in Banach spaces. J Math Anal Appl 44:57–70
18. Kitahara S, Takahashi W (1993) Image recovery by convex combinations of sunny nonexpansive retractions. Method Nonlinear Anal 2:333–342
19. Suzuki T (2005) Strong convergence of Krasnoselskii and Mann's type sequences for one-parameter nonexpansive semigroups without Bochner integrals. J Math Anal Appl 305:227–239
20. Li S, Su Y, Zhang L, Zhao H, Li L (2011) Viscosity approximation methods with weak contraction for L-Lipschitzian pseudocontractive self-mapping. Nonlinear Anal 74:1031–1039
21. Cho YJ, Zhou HY, Guo G (2004) Weak and strong convergence theorems for three-step iterations with errors for asymptotically nonexpansive mappings. Compt Math Appl 47:707–717
22. Browder FE (1976) Nonlinear operators and nonlinear equations of evolution in Banach spaces. In: Proceedings of the symposium on pure mathematics, vol 18, pp 78–81
23. Bruck RE (1981) On the convex approximation property and the asymptotic behavior of nonlinear contractions in Banach spaces. Israel J Math 38:304–314

Optimal Models for Adding Relation to an Organization Structure with Different Numbers of Subordinates at Each Level

Kiyoshi Sawada, Hidefumi Kawakatsu and Takashi Mitsuishi

Abstract This study aims at revealing optimal additional relations to a pyramid organization such that the communication of information between every member in the organization becomes the most efficient. This paper proposes models of adding a relation between two members in the same level of the organization structure in which each member of m-th level below the top has different number of subordinates, $m + 2$ in Model-1 and $2m + 2$ in Model-2. The total shortening distance which is the sum of shortening lengths of shortest paths between every pair of all nodes is formulated and is illustrated with numerical examples to obtain an optimal level of two members between which a relation is added for each of Model-1 and Model-2.

Keywords Adding relation · Optimal model · Organization structure · Pyramid organization · Rooted tree · Total shortening distance

1 Introduction

Our studies aim at revealing optimal additional relations to a pyramid organization such that the communication of information between every member in the organization becomes the most efficient. We have obtained an optimal set of additional edges

K. Sawada (✉)
Department of Policy Studies, University of Marketing and Distribution Sciences,
3-1 Gakuen-nishi-machi, Nishi-ku, Kobe 651-2188, Japan
e-mail: Kiyoshi_Sawada@red.umds.ac.jp

H. Kawakatsu
Department of Economics and Information Science, Onomichi City University,
1600-2 Hisayamadacho, Onomichi 722-8506, Japan
e-mail: kawakatsu@onomichi-u.ac.jp

T. Mitsuishi
Department of Commerce, University of Marketing and Distribution Sciences,
3-1 Gakuen-nishi-machi, Nishi-ku, Kobe 651-2188, Japan
e-mail: takashi_mitsuishi@red.umds.ac.jp

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 275, DOI: 10.1007/978-94-007-7684-5_30,

to a complete K-ary tree of height $H (H = 1, 2, \ldots)$ minimizing the sum of lengths of shortest paths between every pair of all nodes in the complete K-ary tree for the following three models in [5]: (1) a model of adding an edge between two nodes with the same depth, (2) a model of adding edges between every pair of nodes with the same depth, and (3) a model of adding edges between every pair of siblings with the same depth. A complete K-ary tree is a rooted tree in which all leaves have the same depth and all internal nodes have $K (K = 2, 3, \ldots)$ children [1]. We have also obtained an optimal set of levels for a model of adding edges between every pair of nodes with the same depth in multi-levels in a complete binary tree [3].

The complete K-ary tree expresses a pyramid organization in which every member except the top should have a single superior. Nodes and edges in the complete K-ary tree correspond to members and relations between members in the organization respectively. Then the pyramid organization structure is characterized by the number of subordinates of each member [2, 6], that is, K which is the number of children of each node and the number of levels in the organization, that is, H which is the height of the complete K-ary tree. Moreover, the path between each node in the complete K-ary tree is equivalent to the route of communication of information between each member in the organization, and adding edges to the complete K-ary tree is equivalent to forming additional relations other than that between each superior and his subordinates.

The above models give us optimal additional relations to the organization structure of a complete K-ary tree, but these models cannot be applied to adding relations to an organization structure which is not a complete K-ary tree. This paper expands the above model (1) into a model of adding an edge between two nodes with the same depth in a rooted tree with different numbers of children at each depth, that is, a model of adding a ralation between two members of the same level in a pyramid organization structure with different numbers of subordinates at each level. This paper assumes that each node with a depth m has $m + 2$ children in Model-1 [4] and $2m + 2$ children in Model-2.

If $l_{i,j} (= l_{j,i})$ denotes the distance, which is the number of edges in the shortest path from a node v_i to a node v_j in the rooted tree, then $\sum_{i<j} l_{i,j}$ is the total distance. Furthermore, if $l'_{i,j}$ denotes the distance from v_i to v_j after adding an edge, $l_{i,j} - l'_{i,j}$ is called the shortening distance between v_i and v_j, and $\sum_{i<j} (l_{i,j} - l'_{i,j})$ is called the total shortening distance. Minimizing the total distance is equivalent to maximizing the total shortening distance.

Section 2 formulates the total shortening distance of each of Model-1 and Model-2. Section 3 shows an optimal depth of the deepest common ancestor of the two nodes on which the adding edge is incident and Sect. 4 illustrates an optimal depth of adding an edge which maximizes the total shortening distance with numerical examples for each of two models.

2 Formulation of Total Shortening Distance

This section formulates the total shortening distance when a new edge between two nodes with the same depth $N(N = 1, 2, \ldots, H)$ is added to a rooted tree of height $H(H = 1, 2, \ldots)$ in which each node with a depth $m(m = 0, 1, \ldots, H - 1)$ has $m + 2$ children in Model-1 and $2m + 2$ children in Model-2.

We can add a new edge between two nodes with the same depth N in the rooted tree of the above two models in N ways that lead to non-isomorphic graphs. Let $S_{1,H}(N, L)$ and $S_{2,H}(N, L)$ denote the total shortening distance by adding the new edge, where $L(L = 0, 1, 2, \ldots, N - 1)$ is the depth of the deepest common ancestor of the two nodes on which the new edge is incident for Model-1 and Model-2 respectively.

We formulate $S_{1,H}(N, L)$ and $S_{2,H}(N, L)$ in the following.

Let v_0^X and v_0^Y denote the two nodes on which the adding edge is incident. Let v_k^X and v_k^Y denote ancestors of v_0^X and v_0^Y, respectively, with depth $N - k$ for $k = 1, 2, \ldots, N - L - 1$. The sets of descendants of v_0^X and v_0^Y are denoted by V_0^X and V_0^Y respectively. (Note that every node is a descendant of itself [1].) Let V_k^X denote the set obtained by removing the descendants of v_{k-1}^X from the set of descendants of v_k^X and let V_k^Y denote the set obtained by removing the descendants of v_{k-1}^Y from the set of descendants of v_k^Y, where $k = 1, 2, \ldots, N - L - 1$.

Since addition of the new edge doesn't shorten distances between pairs of nodes other than between pairs of nodes in $V_k^X(k = 0, 1, 2, \ldots, N - L - 1)$ and nodes in $V_k^Y(k = 0, 1, 2, \ldots, N - L - 1)$, the total shortening distance can be formulated by adding up the following three sums of shortening distances:

1. The sum of shortening distances between every pair of nodes in V_0^X and nodes in V_0^Y.
2. The sum of shortening distances between every pair of nodes in V_0^X and nodes in $V_k^Y(k = 1, 2, \ldots, N - L - 1)$ and between every pair of nodes in V_0^Y and nodes in $V_k^X(k = 1, 2, \ldots, N - L - 1)$.
3. The sum of shortening distances between every pair of nodes in $V_k^X(k = 1, 2, \ldots, N - L - 1)$ and nodes in $V_k^Y(k = 1, 2, \ldots, N - L - 1)$.

2.1 Formulation of Model-1

This subsection formulates the total shortening distance of Model-1 in which each node with a depth m has $m + 2$ children.

The sum of shortening distances between every pair of nodes in V_0^X and nodes in V_0^Y is given by

$$A_{1,H}(N, L) = \left(\sum_{i=N}^{H-1} \prod_{j=N}^{i} (j + 2) + 1 \right)^2 (2N - 2L - 1)$$

$$= \left(\sum_{i=N}^{H-1} \frac{(i+2)!}{(N+1)!} + 1 \right)^2 (2N - 2L - 1) \tag{2.1}$$

where we define $\sum_{i=h}^{h-1} \cdot = 0$. The sum of shortening distances between every pair of nodes in V_0^X and nodes in $V_k^Y (k = 1, 2, \ldots, N - L - 1)$ and between every pair of nodes in V_0^Y and nodes in $V_k^X (k = 1, 2, \ldots, N - L - 1)$ is given by

$$\begin{aligned} B_{1,H}(N, L) &= 2 \left(\sum_{i=N}^{H-1} \prod_{j=N}^{i} (j+2) + 1 \right) \\ &\quad \times \sum_{i=L+1}^{N-1} \left\{ (i+1) \left(\sum_{j=i+1}^{H-1} \prod_{k=i+1}^{j} (k+2) + 1 \right) + 1 \right\} (2i - 2L - 1) \\ &= 2 \left(\sum_{i=N}^{H-1} \frac{(i+2)!}{(N+1)!} + 1 \right) \\ &\quad \times \sum_{i=L+1}^{N-1} \left\{ (i+1) \left(\sum_{j=i+1}^{H-1} \frac{(j+2)!}{(i+2)!} + 1 \right) + 1 \right\} (2i - 2L - 1) \end{aligned} \tag{2.2}$$

and the sum of shortening distances between every pair of nodes in $V_k^X (k = 1, 2, \ldots, N - L - 1)$ and nodes in $V_k^Y (k = 1, 2, \ldots, N - L - 1)$ is given by

$$\begin{aligned} &C_{1,H}(N, L) \\ &= \sum_{i=L+2}^{N-1} \left\{ (i+1) \left(\sum_{j=i+1}^{H-1} \prod_{k=i+1}^{j} (k+2) + 1 \right) + 1 \right\} \\ &\quad \times \sum_{j=N+L-i+1}^{N-1} \left\{ (j+1) \left(\sum_{k=j+1}^{H-1} \prod_{l=j+1}^{k} (l+2) + 1 \right) + 1 \right\} (2i + 2j - 2N - 2L - 1) \\ &= \sum_{i=L+2}^{N-1} \left\{ (i+1) \left(\sum_{j=i+1}^{H-1} \frac{(j+2)!}{(i+2)!} + 1 \right) + 1 \right\} \\ &\quad \times \sum_{j=N+L-i+1}^{N-1} \left\{ (j+1) \left(\sum_{k=j+1}^{H-1} \frac{(k+2)!}{(j+2)!} + 1 \right) + 1 \right\} (2i + 2j - 2N - 2L - 1) \end{aligned} \tag{2.3}$$

where we define $\sum_{i=h}^{h-2} \cdot = 0$.

From the above equations, the total shortening distance $S_{1,H}(N, L)$ is given by

$$
\begin{aligned}
&S_{1,H}(N, L)\\
&= A_{1,H}(N, L) + B_{1,H}(N, L) + C_{1,H}(N, L)\\
&= \left(\sum_{i=N}^{H-1} \frac{(i+2)!}{(N+1)!} + 1\right)^2 (2N - 2L - 1) \; + \; 2\left(\sum_{i=N}^{H-1} \frac{(i+2)!}{(N+1)!} + 1\right)\\
&\quad \times \sum_{i=L+1}^{N-1} \left\{(i+1)\left(\sum_{j=i+1}^{H-1} \frac{(j+2)!}{(i+2)!} + 1\right) + 1\right\} (2i - 2L - 1)\\
&\quad + \sum_{i=L+2}^{N-1} \left\{(i+1)\left(\sum_{j=i+1}^{H-1} \frac{(j+2)!}{(i+2)!} + 1\right) + 1\right\}\\
&\quad \times \sum_{j=N+L-i+1}^{N-1} \left\{(j+1)\left(\sum_{k=j+1}^{H-1} \frac{(k+2)!}{(j+2)!} + 1\right) + 1\right\} (2i + 2j - 2N - 2L - 1).
\end{aligned}
\tag{2.4}
$$

2.2 Formulation of Model-2

This subsection formulates the total shortening distance of Model-2 in which each node with a depth m has $2m + 2$ children.

The sum of shortening distances between every pair of nodes in V_0^X and nodes in V_0^Y is given by

$$
\begin{aligned}
A_{2,H}(N, L) &= \left(\sum_{i=N}^{H-1} \prod_{j=N}^{i} (2j + 2) + 1\right)^2 (2N - 2L - 1)\\
&= \left(\sum_{i=N}^{H-1} 2^{i-N+1} \frac{(i+1)!}{N!} + 1\right)^2 (2N - 2L - 1)
\end{aligned}
\tag{2.5}
$$

where we define $\sum_{i=h}^{h-1} \cdot = 0$. The sum of shortening distances between every pair of nodes in V_0^X and nodes in $V_k^Y (k = 1, 2, \ldots, N - L - 1)$ and between every pair of nodes in V_0^Y and nodes in $V_k^X (k = 1, 2, \ldots, N - L - 1)$ is given by

$$
\begin{aligned}
B_{2,H}(N, L) &= 2\left(\sum_{i=N}^{H-1} \prod_{j=N}^{i} (2j + 2) + 1\right)\\
&\quad \times \sum_{i=L+1}^{N-1} \left\{(2i + 1)\left(\sum_{j=i+1}^{H-1} \prod_{k=i+1}^{j} (2k + 2) + 1\right) + 1\right\} (2i - 2L - 1)
\end{aligned}
$$

$$= 2\left(\sum_{i=N}^{H-1} 2^{i-N+1}\frac{(i+1)!}{N!} + 1\right) \times \sum_{i=L+1}^{N-1}\left\{(2i+1)\left(\sum_{j=i+1}^{H-1} 2^{j-i}\frac{(j+1)!}{(i+1)!} + 1\right) + 1\right\}(2i-2L-1) \tag{2.6}$$

and the sum of shortening distances between every pair of nodes in $V_k^X (k = 1, 2, \ldots, N-L-1)$ and nodes in $V_k^Y (k = 1, 2, \ldots, N-L-1)$ is given by

$$\begin{aligned} C_{2,H}(N,L) &= \sum_{i=L+2}^{N-1}\left\{(2i+1)\left(\sum_{j=i+1}^{H-1}\prod_{k=i+1}^{j}(2k+2)+1\right)+1\right\} \\ &\quad \times \sum_{j=N+L-i+1}^{N-1}\left\{(2j+1)\left(\sum_{k=j+1}^{H-1}\prod_{l=j+1}^{k}(2l+2)+1\right)+1\right\} \\ &\quad \times (2i+2j-2N-2L-1) \\ &= \sum_{i=L+2}^{N-1}\left\{(2i+1)\left(\sum_{j=i+1}^{H-1} 2^{j-i}\frac{(j+1)!}{(i+1)!}+1\right)+1\right\} \\ &\quad \times \sum_{j=N+L-i+1}^{N-1}\left\{(2j+1)\left(\sum_{k=j+1}^{H-1} 2^{k-j}\frac{(k+1)!}{(j+1)!}+1\right)+1\right\} \\ &\quad \times (2i+2j-2N-2L-1) \end{aligned} \tag{2.7}$$

where we define $\sum_{i=h}^{h-2} \cdot = 0$.

From the above equations, the total shortening distance $S_{2,H}(N, L)$ is given by

$$\begin{aligned} S_{2,H}(N,L) &= A_{2,H}(N,L) + B_{2,H}(N,L) + C_{2,H}(N,L) \\ &= \left(\sum_{i=N}^{H-1} 2^{i-N+1}\frac{(i+1)!}{N!}+1\right)^2 (2N-2L-1) \\ &\quad + 2\left(\sum_{i=N}^{H-1} 2^{i-N+1}\frac{(i+1)!}{N!}+1\right) \\ &\quad \times \sum_{i=L+1}^{N-1}\left\{(2i+1)\left(\sum_{j=i+1}^{H-1} 2^{j-i}\frac{(j+1)!}{(i+1)!}+1\right)+1\right\}(2i-2L-1) \end{aligned}$$

$$+\sum_{i=L+2}^{N-1}\left\{(2i+1)\left(\sum_{j=i+1}^{H-1}2^{j-i}\frac{(j+1)!}{(i+1)!}+1\right)+1\right\}$$
$$\times\sum_{j=N+L-i+1}^{N-1}\left\{(2j+1)\left(\sum_{k=j+1}^{H-1}2^{k-j}\frac{(k+1)!}{(j+1)!}+1\right)+1\right\}$$
$$\times(2i+2j-2N-2L-1). \tag{2.8}$$

3 An Optimal Depth L^* for Each N

This section shows an optimal depth L^* of the deepest common ancestor of the two nodes with a depth N on which the adding edge is incident which maximizes the total shortening distance where each node of a depth m has $m+2$ children in Model-1 and $2m+2$ children in Model-2.

3.1 An Optimal Depth L^* of Model-1

Theorem 1 *$L^*=0$ maximizes $S_{1,H}(N,L)$ for each N.*

Proof If $N=1$, then $L^*=0$ trivially. If $N\geq 2$, then $L^*=0$ for each N since

$$\begin{aligned}
&S_{1,H}(N,L+1)-S_{1,H}(N,L)\\
&=-2\left(\sum_{i=N}^{H-1}\frac{(i+2)!}{(N+1)!}+1\right)^2\\
&\quad-2\left(\sum_{i=N}^{H-1}\frac{(i+2)!}{(N+1)!}+1\right)\left\{(L+2)\left(\sum_{j=L+2}^{H-1}\frac{(j+2)!}{(L+3)!}+1\right)+1\right\}\\
&\quad-4\left(\sum_{i=N}^{H-1}\frac{(i+2)!}{(N+1)!}+1\right)\sum_{i=L+2}^{N-1}\left\{(i+1)\left(\sum_{j=i+1}^{H-1}\frac{(j+2)!}{(i+2)!}+1\right)+1\right\}\\
&\quad-\sum_{i=L+3}^{N-1}\left\{(i+1)\left(\sum_{j=i+1}^{H-1}\frac{(j+2)!}{(i+2)!}+1\right)+1\right\}\\
&\quad\times\left[\left\{(N+L-i+2)\left(\sum_{k=N+L-i+2}^{H-1}\frac{(k+2)!}{(N+L-i+3)!}+1\right)+1\right\}\right.
\end{aligned}$$

$$+ 2 \sum_{j=N+L-i+2}^{N-1} \left\{ (j+1) \left(\sum_{k=j+1}^{H-1} \frac{(k+2)!}{(j+2)!} + 1 \right) + 1 \right\} \Bigg] + t_{1,H}(N, L)$$
$$< 0 \tag{3.9}$$

where

$$t_{1,H}(N, L) = -\left\{ (L+3) \left(\sum_{j=L+3}^{H-1} \frac{(j+2)!}{(L+4)!} + 1 \right) + 1 \right\} \times \left\{ N \left(\sum_{k=N}^{H-1} \frac{(k+2)!}{(N+1)!} + 1 \right) + 1 \right\} \tag{3.10}$$

for $L = 0, 1, 2, \ldots, N-3$ and

$$t_{1,H}(N, L) = 0 \tag{3.11}$$

for $L = N-2$. □

Theorem 1 shows that the most efficient additional relation between two members in the same level N of the organization structure in which each member of m-th level below the top has $m+2$ subordinates is that between two members which doesn't have common superiors except the top.

3.2 An Optimal Depth L of Model-2*

Theorem 2 *$L^* = 0$ maximizes $S_{2,H}(N, L)$ for each N.*

Proof If $N = 1$, then $L^* = 0$ trivially. If $N \geq 2$, then $L^* = 0$ for each N since

$$\begin{aligned}
&S_{2,H}(N, L+1) - S_{2,H}(N, L) \\
&= -2 \left(\sum_{i=N}^{H-1} 2^{i-N+1} \frac{(i+1)!}{N!} + 1 \right)^2 \\
&\quad - 2 \left(\sum_{i=N}^{H-1} 2^{i-N+1} \frac{(i+1)!}{N!} + 1 \right) \left\{ (2L+3) \left(\sum_{j=L+2}^{H-1} 2^{j-L-1} \frac{(j+1)!}{(L+2)!} + 1 \right) + 1 \right\} \\
&\quad - 4 \left(\sum_{i=N}^{H-1} 2^{i-N+1} \frac{(i+1)!}{N!} + 1 \right) \sum_{i=L+2}^{N-1} \left\{ (2i+1) \left(\sum_{j=i+1}^{H-1} 2^{j-i} \frac{(j+1)!}{(i+1)!} + 1 \right) + 1 \right\}
\end{aligned}$$

$$
\begin{aligned}
&- \sum_{i=L+3}^{N-1} \left\{ (2i+1) \left(\sum_{j=i+1}^{H-1} 2^{j-i} \frac{(j+1)!}{(i+1)!} + 1 \right) + 1 \right\} \\
&\times \left[\left\{ (2N+2L-2i+3) \left(\sum_{k=N+L-i+2}^{H-1} 2^{i+k-N-L-1} \frac{(k+1)!}{(N+L-i+2)!} + 1 \right) + 1 \right\} \right. \\
&\left. + 2 \sum_{j=N+L-i+2}^{N-1} \left\{ (2j+1) \left(\sum_{k=j+1}^{H-1} 2^{k-j} \frac{(k+1)!}{(j+1)!} + 1 \right) + 1 \right\} \right] + t_{2,H}(N, L) \\
&< 0 \qquad (3.12)
\end{aligned}
$$

where

$$
\begin{aligned}
t_{2,H}(N, L) = &- \left\{ (2L+5) \left(\sum_{j=L+3}^{H-1} 2^{j-L-2} \frac{(j+1)!}{(L+3)!} + 1 \right) + 1 \right\} \\
&\times \left\{ (2N-1) \left(\sum_{k=N}^{H-1} 2^{k-N+1} \frac{(k+1)!}{N!} + 1 \right) + 1 \right\} \qquad (3.13)
\end{aligned}
$$

for $L = 0, 1, 2, \ldots, N-3$ and

$$
t_{2,H}(N, L) = 0 \qquad (3.14)
$$

for $L = N-2$. □

Theorem 2 shows that the most efficient additional relation between two members in the same level N of the organization structure in which each member of m-th level below the top has $2m+2$ subordinates is that between two members which doesn't have common superiors except the top.

4 Numerical Examples

This section illustrates the total shortening distance with numerical examples in the case of $L = 0$ to obtain an optimal depth N^* of two nodes on which the adding edge is incident for each of Model-1 and Model-2.

4.1 Numerical Examples of Model-1

Let $\hat{S}_{1,H}(N)$ denote the total shortening distance of Model-1 when $L = 0$, then $\hat{S}_{1,H}(N)$ becomes

$$
\begin{aligned}
\hat{S}_{1,H}(N) &= S_{1,H}(N,0) \\
&= \left(\sum_{i=N}^{H-1} \frac{(i+2)!}{(N+1)!} + 1\right)^2 (2N-1) + 2\left(\sum_{i=N}^{H-1} \frac{(i+2)!}{(N+1)!} + 1\right) \\
&\quad \times \sum_{i=1}^{N-1} \left\{ (i+1)\left(\sum_{j=i+1}^{H-1} \frac{(j+2)!}{(i+2)!} + 1\right) + 1 \right\} (2i-1) \\
&\quad + \sum_{i=2}^{N-1} \left\{ (i+1)\left(\sum_{j=i+1}^{H-1} \frac{(j+2)!}{(i+2)!} + 1\right) + 1 \right\} \\
&\quad \times \sum_{j=N-i+1}^{N-1} \left\{ (j+1)\left(\sum_{k=j+1}^{H-1} \frac{(k+2)!}{(j+2)!} + 1\right) + 1 \right\} (2i+2j-2N-1)
\end{aligned}
\tag{4.15}
$$

Table 1 shows numerical examples of the total shortening distance $\hat{S}_{1,H}(N)$ in the case of $H = 1, 2, \ldots, 8$ and $N = 1, 2, \ldots, H$.

Table 1 reveals that $N^* = 1$ maximizes $\hat{S}_{1,H}(N)$ irrespective of H when $H = 1, 2, \ldots, 8$. This means that the most efficient level of adding a relation to the organization structure in which each member of m-th level below the top has $m + 2$ subordinates is the first level below the top when the organization structure has few levels.

4.2 Numerical Examples of Model-2

Let $\hat{S}_{2,H}(N)$ denote the total shortening distance of Model-2 when $L = 0$, then $\hat{S}_{2,H}(N)$ becomes

Table 1 Total shortening distance $\hat{S}_{1,H}(N)$

N	$H=1$	$H=2$	$H=3$	$H=4$	$H=5$	$H=6$	$H=7$	$H=8$
1	1	16	256	5776	190096	8737936	534349456	41843157136
2	–	9	185	4425	147465	6793545	415584585	32544495945
3	–	–	67	1837	62857	2909197	178089517	13947480397
4	–	–	–	538	19870	931354	57121306	4474677274
5	–	–	–	–	4992	244210	15070434	1181504850
6	–	–	–	–	–	54087	3418791	268847367
7	–	–	–	–	–	–	679033	54120403
8	–	–	–	–	–	–	–	9759060

Table 2 Total shortening distance $\hat{S}_{2,H}(N)$

N	$H=1$	$H=2$	$H=3$	$H=4$	$H=5$	$H=6$	$H=7$	$H=8$
1	1	25	841	48841	4583881	634082761	120923803081	30345786707401
2	–	11	455	27335	2577095	356655815	68019421895	17069501579975
3	–	–	121	7993	762553	105663673	20153735353	5057627445433
4	–	–	–	1623	162023	22554023	4303530023	1080012099623
5	–	–	–	–	27001	3842905	734641561	184390607257
6	–	–	–	–	–	545895	105514647	26504722839
7	–	–	–	–	–	–	13092393	3306964137
8	–	–	–	–	–	–	–	364688967

$$
\begin{aligned}
\hat{S}_{2,H}(N) &= S_{2,H}(N,0) \\
&= \left(\sum_{i=N}^{H-1} 2^{i-N+1}\frac{(i+1)!}{N!}+1\right)^2 (2N-1) + 2\left(\sum_{i=N}^{H-1} 2^{i-N+1}\frac{(i+1)!}{N!}+1\right) \\
&\quad \times \sum_{i=1}^{N-1}\left\{(2i+1)\left(\sum_{j=i+1}^{H-1} 2^{j-i}\frac{(j+1)!}{(i+1)!}+1\right)+1\right\}(2i-1) \\
&\quad + \sum_{i=2}^{N-1}\left\{(2i+1)\left(\sum_{j=i+1}^{H-1} 2^{j-i}\frac{(j+1)!}{(i+1)!}+1\right)+1\right\} \\
&\quad \times \sum_{j=N-i+1}^{N-1}\left\{(2j+1)\left(\sum_{k=j+1}^{H-1} 2^{k-j}\frac{(k+1)!}{(j+1)!}+1\right)+1\right\}(2i+2j-2N-1).
\end{aligned}
\tag{4.16}
$$

Table 2 shows numerical examples of the total shortening distance $\hat{S}_{2,H}(N)$ in the case of $H = 1, 2, \ldots, 8$ and $N = 1, 2, \ldots, H$.

Table 2 reveals that $N^* = 1$ maximizes $\hat{S}_{2,H}(N)$ irrespective of H when $H = 1, 2, \ldots, 8$. This means that the most efficient level of adding a relation to the organization structure in which each member of m-th level below the top has $2m + 2$ subordinates is the first level below the top when the organization structure has few levels.

5 Conclusions

This study considered revealing an optimal additional relation to a pyramid organization with different numbers of subordinates at each level such that the communication of information between every member in the organization becomes the most efficient. For models of adding an edge between two nodes with the same depth of the rooted

tree in which each node with a depth m has $m + 2$ children in Model-1 and $2m + 2$ children in Model-2, we formulated the total shortening distance and showed that an optimal depth of the deepest common ancestor of the two nodes on which the adding edge is incident is $L^* = 0$ for each N in Theorem 1 for Model-1 and in Theorem 2 for Model-2. Furthermore, we illustrated an optimal depth $N^* = 1$ of adding an edge which maximizes the total shortening distance with numerical examples for each of two models.

Theorem 1 and 2 and numerical examples reveal that the most efficient manner of adding a relation between two members in the same level of the organization structure of Model-1 and Model-2 is to add the relation between two members in the first level below the top when the organization structure has few levels.

References

1. Cormen TH, Leiserson CE, Rivest RL, Stein C (2001) Introduction to algorithms, 2nd edn. MIT Press, Cambridge
2. Robbins SP (2003) Essentials of organizational behavior, 7th edn. Prentice Hall, Upper Saddle River
3. Sawada K, Amano K (2009) A model of adding relations in multi-levels to a formal organization structure with two subordinates. IAENG Trans Eng Technol 3:109–116
4. Sawada K, Kawakatsu H, Mitsuishi T (2013) A model for adding an efficient relation to an organization structure with different numbers of subordinates at each level. Lecture notes in engineering and computer science: proceedings of the international multiConference of engineers and computer scientists 2013, Hong Kong, pp 1057–1060, 13–15 Mar 2013.
5. Sawada K, Wilson R (2006) Models of adding relations to an organization structure of a complete K-ary tree. Eur J Oper Res 174:1491–1500
6. Takahara Y, Mesarovic M (2003) Organization structure: cybernetic systems foundation. Kluwer Academic/Plenum Publishers, NY

Author Index

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 275, DOI: 10.1007/978-94-007-7684-5,

Subject Index

G.-C. Yang et al. (eds.), *Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 275, DOI: 10.1007/978-94-007-7684-5,