Mathematics for Industry 25

Bob Anderssen
Philip Broadbridge
Yasuhide Fukumoto
Naoyuki Kamiyama
Yoshihiro Mizoguchi
Konrad Polthier
Osamu Saeki *Editors*

The Role and Importance of Mathematics in Innovation

Proceedings of the Forum "Math-for-Industry" 2015

Springer

Mathematics for Industry

Volume 25

Aims & Scope

The meaning of "Mathematics for Industry" (sometimes abbreviated as MI or MfI) is different from that of "Mathematics in Industry" (or of "Industrial Mathematics"). The latter is restrictive: it tends to be identified with the actual mathematics that specifically arises in the daily management and operation of manufacturing. The former, however, denotes a new research field in mathematics that may serve as a foundation for creating future technologies. This concept was born from the integration and reorganization of pure and applied mathematics in the present day into a fluid and versatile form capable of stimulating awareness of the importance of mathematics in industry, as well as responding to the needs of industrial technologies. The history of this integration and reorganization indicates that this basic idea will someday find increasing utility. Mathematics can be a key technology in modern society.

The series aims to promote this trend by (1) providing comprehensive content on applications of mathematics, especially to industry technologies via various types of scientific research, (2) introducing basic, useful, necessary and crucial knowledge for several applications through concrete subjects, and (3) introducing new research results and developments for applications of mathematics in the real world. These points may provide the basis for opening a new mathematics-oriented technological world and even new research fields of mathematics.

More information about this series at http://www.springer.com/series/13254

Bob Anderssen · Philip Broadbridge
Yasuhide Fukumoto · Naoyuki Kamiyama
Yoshihiro Mizoguchi · Konrad Polthier
Osamu Saeki
Editors

The Role and Importance of Mathematics in Innovation

Proceedings of the Forum "Math-for-Industry" 2015

Springer

Editors
Bob Anderssen
CSIRO Mathematics, Informatics and Statistics
Canberra, ACT
Australia

Philip Broadbridge
Department of Mathematics and Statistics
La Trobe University
Melbourne, VIC
Australia

Yasuhide Fukumoto
Institute of Mathematics for Industry
Kyushu University
Fukuoka
Japan

Naoyuki Kamiyama
Institute of Mathematics for Industry
Kyushu University
Fukuoka
Japan

Yoshihiro Mizoguchi
Institute of Mathematics for Industry
Kyushu University
Fukuoka
Japan

Konrad Polthier
AG Mathematical Geometry Processing
Freie Universität Berlin
Berlin
Germany

Osamu Saeki
Institute of Mathematics for Industry
Kyushu University
Fukuoka
Japan

ISSN 2198-350X
Mathematics for Industry
ISBN 978-981-10-0961-7
DOI 10.1007/978-981-10-0962-4

ISSN 2198-3518 (electronic)
ISBN 978-981-10-0962-4 (eBook)

Library of Congress Control Number: 2016945787

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer Science+Business Media Singapore Pte Ltd.

Preface

This book contains the proceedings of the Forum "Math-for-Industry" 2015 held at the Institute of Mathematics for Industry, Kyushu University, October 26–30, 2015, for which the unifying theme was "The Role and Importance of Mathematics in Innovation". Selected papers presented at the forum are collected here.

Innovation is in fact the cornerstone of creativity in all human endeavors. It involves "seeing" things from an entirely new, sometimes quite elementary, perspective. Innovation in mathematics is the bread and butter of mathematical creativity. Historical examples of mathematical innovation that have had profound and lasting impacts on the subsequent development of mathematics include the logarithm, complex numbers, non-Euclidean geometry, and calculus. Equally important is innovation in the performance of mathematics which can be disarmingly simple but have profound consequences. Examples include adding zero, multiplication by one, and seeing a new interpretation that simplifies matters. This book illustrates two different types of key roles that mathematics plays in supporting innovation in science, technology, and daily life:

(1) Needs-based. Once a need or an opportunity for innovation has been identified, the subsequent experimentation and/or lateral thinking utilizes mathematics to assist with sorting through the possibilities and putting matters on a more rigorous foundation. An example is the development of Wi-Fi.
(2) Idea-based. After an idea for an innovation has materialized, mathematical models of the possible implementations play a key role. An example is the design of the next model of an automobile that exploits recent developments in materials and technology. Being able to innovate comes from experiencing and understanding how innovation occurs in mathematics, science, and technology.

The contents of this volume report on productive and successful interaction between industry and mathematicians, as well as on the cross-fertilization and collaboration that occurred. The book contains excellent examples of the roles of

mathematics in innovation and, thereby, the importance and relevance of the concept Mathematics_FOR_Industry.

We would like to thank the participants of the forum, especially the members of the Scientific Board of the Forum. Without their cooperation and support, we would never have experienced the great excitement and success of the forum. Moreover, we would like to express our deep appreciation for the great help of the conference secretaries during the preparation and organization of the forum, and to Chiemi Furutani for the proceedings.

Fukuoka, Japan
April 2016

Yasuhide Fukumoto
On behalf of
The Organizing Committee of the Forum "Math-for-Industry" 2015
and
The Editorial Committee of the Proceedings

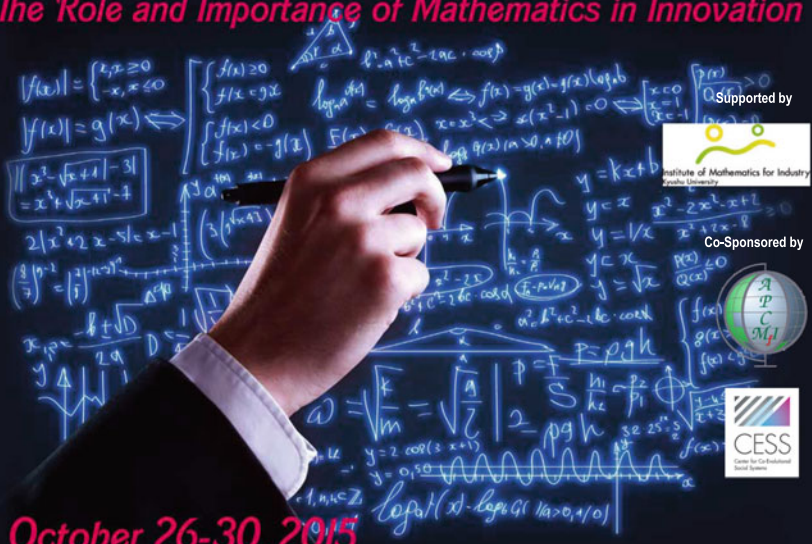
FMI2008	FMI2009	FMI2010	FMI2011	FMI2012	FMI2013	FMИ2014	FMИ2015
Tokyo	Fukuoka	Fukuoka	Honolulu	Fukuoka	Fukuoka	Fukuoka	Fukuoka
Sep. 19-17	Nov. 9-13	Oct. 21-23	Oct. 24-28	Oct. 22-26	Nov. 4-8	Oct.27-31	Oct.26-30
The 1st Forum:Consortium Math For Industry	Casimir Force, Casimir Operators and the Riemann Hypothesis	Information Security, Visualization, and Inverse Problems, on the basis of Optimization Techniques	TSUNAMI - Mathematical Modelling- Using Mathematics for Natural Disaster Prediction, Recovery and Provision for the Future -	Information Recovery and Discovery	-The Impact of Applications on Mathematics-	-Applications + Practical Conceptualization + Mathematics = fruitful Innovation-	Mathematics The Role and Importance of Mathematics in Innovation
							

九州大学 KYUSHU UNIVERSITY

Institute of Mathematics for Industry Kyushu University

Forum "Math-for-Industry" 2015

The Role and Importance of Mathematics in Innovation



Supported by

Institute of Mathematics for Industry Kyushu University

Co-Sponsored by

CESS Center for Co-Evolutional Social Systems

October 26-30, 2015

Venue: Auditorium, Institute of Mathematics for Industry, Kyushu University Ito Campus, Fukuoka

Invited Speakers

Kazushi Ahara	Meiji University
Pierluigi Cesana	IMI Australia Branch
Nguyen Huu Du	Vietnam Institute for Advanced Study in Mathematics
Marcio Gameiro	University of São Paulo
Peter Höfner	NICTA and UNSW
Marcel Jackson	La Trobe University
Manish Jain	ARMORWAY INC.
Akifumi Kira	IMI
Rafael López	Universidad de Granada
Gaven Martin	Massey University
Mary Myerscough	University of Sydney
Tak Nishida	Intec Innovative Technologies USA, Inc.
Kotaro Ohori	FUJITSU LABORATORIES LTD.
Takayuki Osogami	IBM Research - Tokyo
Taehyun Rhee	Victoria University of Wellington
Arnab Roy	Fujitsu Laboratories of America
Takenobu Seito	Bank of Japan
Vitaly Shumeiko	Chalmers University of Technology
Matthew Simpson	Queensland University of Technology
Takashi Suzuki	Osaka University
Dimetre Triadis	IMI Australia Branch
Graham Weir	Wellington
Dmitry Znamenskiy	Philips Research
Laura Karantgis	La Trobe University
Kentaro Okamoto	Graduate School of Mathematics, Kyushu University
Yuichiro Yasui	CESS, Kyushu University

Organizing Committee

Bob Anderssen	CSIRO
Philip Broadbridge	La Trobe University
Yasuhide Fukumoto	IMI,Chair
Naoyuki Kamiyama	IMI
Yoshihiro Mizoguchi	IMI
Konrad Polthier	Freie Universität Berlin and MATHEON
Osamu Saeki	IMI

Scientific Board

Hirokazu Anai	FUJITSU LABORATORIES LTD.
Yasuaki Hiraoka	Tohoku University
Robert McKibbin	Massey University
Ryuei Nishii	IMI
Kanzo Okada	IMI
Wil Schilders	TU Eindhoven
Tsuyoshi Takagi	IMI

Supported by

Institute of Mathematics for Industry, Kyushu University

FUJITSU LABORATORIES LTD.

FMfI2015 Organizing Office

IMI: Institute of Mathematics for Industry, Kyushu University
fmi2015@imi.kyushu-u.ac.jp http://fmi2015.imi.kyushu-u.ac.jp
Tel: 81 92 802 4401/4404 Fax: 81 92 802 4405

Forum "Math-for-Industry" 2014

Applications + Practical Conceptualization + Mathematics = fruitful Innovation

October 27(Mon) - 31(Fri), 2014
Venue: Nishijin Plaza, Kyushu University
2-16-23 Nishijin, Fukuoka City

Invited Speakers

Gary Froyland	University of New South Wales
Masahito Hasegawa	RIMS, Kyoto University
Hans-Christian Hege	Zuse-Institute Berlin (ZIB)
Thorsten Koch	TU Berlin / Zuse-Institute Berlin (ZIB)
Kerry Landman	The University of Melbourne
Vladimir Lorman	CNRS & Université Montpellier 2
Reinout Quispel	La Trobe University

Ernesto G. Birgin	University of São Paulo
Daniel Braak	University of Augsburg
Iñigo L. Egusquiza	University of the Basque Country, Bilbao
Farid Melgani	University of Trento
Robert Norman	RMIT University
Konrad Polthier	Freie Universität Berlin
Enrique Solano	University of the Basque Country, Bilbao
Akira Takada	Asahi Glass Co., Ltd.
Roger C. E. Tan	National University of Singapore

Supported by

Institute of Mathematics for Industry
Kyushu University

Co-Sponsored by

CESS
Center for Co-Evolutional Social Systems

Kenichi Arai	NTT Communication Science Laboratories
Zainal Abdul Aziz	Universiti Teknologi Malaysia
Troy Farrell	Queensland University of Technology
Luke Fullard	Massey University
Arnab Ganguly	University of Louisville
Sachiko Ishida	Meiji University
Shinsaku Kiyomoto	KDDI R&D Laboratories, Inc.
Akira Ohata	Toyota Motor Corporation
Takashi Sasaki	Yokogawa Electric Corporation
Toshinao Yoshiba	Bank of Japan

Alexandra Hogan	Australian National University
Keita Iida	Tohoku University
Yasuaki Kobayashi	Hokkaido University
Lucas Lamata	University of the Basque Country, Bilbao
Mikel Sanz	University of the Basque Country, Bilbao
Eriko Shinkawa	Graduate School of Mathematics, Kyushu University

Organizing Committee

Bob Anderssen	CSIRO, Australia
Philip Broadbridge	La Trobe University
Yasuhide Fukumoto	IMI
Kenji Kajiwara	IMI
Tsuyoshi Takagi	IMI
Evgeny Verbitskiy	Leiden University / University of Groningen
Masato Wakayama	IMI, Chair

Scientific Board

Hirokazu Anai	FUJITSU LABORATORIES LTD.
Shin-Ichiro Ei	Hokkaido University
Tim Hoffmann	Technische Universität München
Miyuki Koiso	IMI
Ryuei Nishii	IMI
Jill Pipher	Brown University
Hisayoshi Sato	Yokohama Research Laboratory, Hitachi, Ltd.
Tomoyuki Shirai	IMI
Katsuyuki Takashima	Mitsubishi Electric Corporation
Hayato Waki	IMI
Takeshi Yamada	NTT Communication Science Laboratories

Asia-Pacific Consortium of Mathematics for Industry

九州大学
KYUSHU UNIVERSITY

Supported by

Institute of Mathematics for Industry, Kyushu University

FMfI2014 Organizing Office

IMI: Institute of Mathematics for Industry, Kyushu University
fmi2014@imi.kyushu-u.ac.jp http://fmi2014.imi.kyushu-u.ac.jp
Tel: 81 92 802 4401/4404 Fax: 81 92 802 4405

Forum "Math-for-Industry" 2015

-The Role and Importance of Mathematics in Innovation-

October 26 to 30, 2015, IMI Auditorium, Kyushu University Ito Campus, Fukuoka, Japan

October 26, 2015

8:30 - Registration

10:15 - 10:25 **Opening Ceremony**

Session Chair: Osamu Saeki

Gaven Martin
Massey University
10:30 - 11:20 **New Approaches to the Modelling of Highly Elastic Media**

COFFEE BREAK

Vitaly Shumeiko
Chalmers University of Technology
11:45 - 12:30 **Parametric Effects in Quantum Electrical Circuits and Applications to Quantum Information**

LUNCH

Session Chair: Tomoyuki Shirai

Nguyen Huu Du
Vietnam Institute for Advanced Study in Mathematics
14:00 - 14:50 **Dynamic Behavior of Kolmogorov Systems Perturbed by Noise**

Takayuki Osogami
IBM Research - Tokyo
14:55 - 15:40 **Human Choice and Good Choice**

COFFEE BREAK

Session Chair: Bob Anderssen

16:00 - 18:00 **Poster Session**
PHOTO

18:00 - 19:30 **Welcome Party**
(Snack & Beverages)
& Poster Viewing

October 27, 2015

Registration

Session Chair: Ryuei Nishii

9:45 - 10:35 **Aleksandar Staykov**
ICNER, Kyushu University
Application of the Non-equilibrium Green's Function Method in the Design of Nanoscale Electronic Devices

10:40 - 11:25 **Tak Nishida**
Intec Innovative Technologies USA, Inc.
Internet-of-Things, Big data, and Entropy

COFFEE BREAK

Dmitry Znamenskiy
Philips Research
11:45 - 12:30 **On 3D Scanning Technologies for Respiratory Mask Design**

12:30 - 12:45 **Notice of & Invitation to APCMI!**

LUNCH

Session Chair: Kenji Kajiwara

Takashi Suzuki
Osaka University
14:00 - 14:50 **Mathematical Oncology and Applications**

Mary Myerscough
University of Sydney
14:55 - 15:40 **Why Do Hives Die? Using Mathematics to Solve the Problem of Honey Bee Colony Collapse**

COFFEE BREAK

Young Reseacher Session
16:00 - 17:00
Laura Karantzis
Steady Saturated-Unsaturated Water Flow in a Sloping Domain and its Application to Landslides
Kentaro Okamoto
Zeta Function Associated with the Representation of the Braid Group
Yuichiro Yasui
Fast, Scalable, and Energy-Efficient Parallel Breadth-First Search

17:00 - 18:00 **Poster Session Voting**

October 28, 2015

Registration

Session Chair: Naoyuki Kamiyama

Fujitsu Session

9:45 - 10:35 **Manish Jain**
Armorway Inc.
Game-Theoretic Allocation of Limited Resources: Applications to Security Domains

10:40 - 11:15 **Kotaro Ohori**
FUJITSU LABORATORIES LTD.
The Role of Mathematical Technologies in Social System Design: An Interdisciplinary Approach to Complex Social Problems

11:20 - 11:55 **Akifumi Kira**
IMI, Kyushu University
Our Challenges to Real-Time Disaster-Recovery Scheduling

12:30 - 19:00 **Excursion with Lunch**

October 29, 2015

Registration

Session Chair: Yoshihiro Mizoguchi

9:45 - 10:35 **Marcel Jackson**
La Trobe University
Algebraic Foundations for Program Logics

10:40 - 11:25 **Kazushi Ahara**
Meiji University
Basic Research for the Patient-Specific Surgery Support System -An Identification Algorithm of the Mesentery Using 3D Medical Images-

COFFEE BREAK

Peter Höfner
NICTA and UNSW
11:45 - 12:30 **Using Process Algebra to Design Better Protocols**

LUNCH

Session Chair: Kirill Morozov

Arnab Roy
Fujitsu Laboratories of America
14:00 - 14:50 **Relational Hash**

Takenobu Seito
Bank Japan
14:55 - 15:40 **Cryptography and Financial Industry**

COFFEE BREAK

Pierluigi Cesana
Australia Branch, IMI, Kyushu University
16:00 - 16:35 **New Models for the Space-Time Evolution of Martensitic Microstructure**

16:40 - 17:00 **Poster Session Award Ceremony**
PHOTO

19:00 - 21:00 **Banquet**
AGORA

October 30, 2015

Registration

Session Chair: Konrad Polthier

9:45 - 10:35 **Rafael López**
Universidad de Granada
Capillary Surfaces Modeling Liquid Drops on Wetting Phenomena

10:40 - 11:25 **Taeyhun Rhee**
Victoria University of Wellington
Deformable Human Body Modeling from 3D Medical Image Scans

COFFEE BREAK

Marcio Gameiro
University of São Paulo
11:45 - 12:30 **Persistent Homology and Applications**

LUNCH

Session Chair: Philip Broadbridge

Graham Weir
Wellington, New Zealand
14:00 - 14:45 **The Mathematics describing Two Phase Geothermal Fluid Flows: Quantifying Industrial Applications and Innovations**

Dimetre Triadis
Australia Branch, IMI, Kyushu University
14:50 - 15:25 **Leveraging Progress in Analytic Groundwater Infiltration for New Solutions in Industrial Metal Solidification**

COFFEE BREAK

Matthew Simpson
Queensland University of Technology
15:45 - 16:30 **Why are Scratch Assays Difficult to Reproduce?**

16:30 **Closing**

Organizing Committee

Bob Anderssen (CSIRO, Australia)
Philip Broadbridge (La Trobe University)
Yasuhide Fukumoto (IMI), Chair
Naoyuki Kamiyama (IMI)
Yoshihiro Mizoguchi (IMI)
Konrad Polthier (Berlin Freie University)
Osamu Saeki (IMI)

Scientific Board

Hirokazu Anai (FUJITSU LABS. LTD.)
Yasuaki Hiraoka (Tohoku University)
Robert McKibbin (Massey University)
Ryuei Nishii (IMI)
Kanzo Okada (IMI)
Wil Schilders (TU Eindhoven)
Tsuyoshi Takagi (IMI)

Contents

Human Choice and Good Choice

Takayuki Osogami

Abstract The choice made by humans is known to depend on available alternatives in rather complex but systematic ways. There has been a significant amount of work on choice models for modeling such human choice. Most of the existing choice models, particularly those in the class of random utility models, however, cannot represent one of the typical phenomena of human choice, known as the attraction effect. Here, we review recent development of choice models that can be trained to learn the attraction effect and other typical phenomena of human choice from the data of the choice made by humans. We also discuss possible extensions of such work on choice models, which suggest potential directions of future research.

Keywords Choice models · Attraction effect · Boltzmann machines · Sequential decision making

1 Introduction

The choice made by humans is often biased or shows complex dependencies on the context where the choice is made. An illustrative example is given by Ariely [2], where students were asked to choose a way to subscribe a magazine from two or three options. Majority of the students liked the subscription to the online edition (Option Online) better than the (more expensive) subscription to both online and print editions (Option Both), when only these two options were available. However, when another option of the subscription to the print edition (Option Print) was also available at the same price as Option Both, the majority of the students liked Option Both the best among the three options.

The phenomenon shown by Ariely [2] is known as the attraction effect and is considered to be one of the phenomena that appear in human choice in a robust and significant manner [23]. In the example, the preference between Option Both and

T. Osogami (✉)
IBM Research—Tokyo, 19-21 Nihonbashi Hakozaki, Chuo-ku, Tokyo 103-8510, Japan
e-mail: osogami@jp.ibm.com

B. Anderssen et al. (eds.), *The Role and Importance of Mathematics in Innovation*, Mathematics for Industry 25,
DOI 10.1007/978-981-10-0962-4_1

Option Online depends on whether Option Print is also available as an alternative. In particular, Option Print acts as a decoy and increases the relative attractiveness of Option Both, because the online edition comes for free once the print edition is selected.

Although the attraction effect is well understood and is exploited by the practitioners of marketing, the research on modeling and learning such phenomena from the data of human choice is only recently addressed in the literature. In fact, despite the long history of research on choice models [13, 35], most of the choice models, in particular those fall into the class of random utility models, cannot represent the attraction effect [23]. In the literature of Psychology, researchers have proposed sequential sampling models, which mimic the cognitive process of the human making a choice [22]. Examples of the sequential sampling models include the decision field theory (DFT) [6, 24] and the leaky competing accumulator model [37, 38]. Roe et al. provide intuitive arguments as to why the DFT can represent the attraction effect and other typical phenomena of human choice, giving numerical examples where the DFT can represent these phenomena [24]. However, learning is not the focus of the research in Psychology, and there has been no significant study of learning the parameters of sequential sampling models from the data of human choice.

Here, we review recent development in the research on modeling and learning human choice, particularly those models that can represent and learn the attraction effect and other typical phenomena of human choice. We will also discuss how such choice models can be exploited for providing good services to humans or for making good decisions.

2 Modeling Human Choice

A model of choice is expected to give the probability of choosing each item from a given set of choices. Let $\mathscr{I}$ be the set of items that can constitute a choice set. Given a choice set, $\mathscr{X} \subseteq \mathscr{I}$, a choice model should give the probability of selecting each item in the choice set: $p(A|\mathscr{X})$ for each $A \in \mathscr{X}$.

An item is characterized by a vector of features. Let $\mathbf{v}_A$ be the feature vector of item $A \in \mathscr{I}$. Each element of a feature vector may be a raw attribute that is explicitly specified in the description of a product, such as the price and the speed. A feature vector may be a vector that is given by a nonlinear transformation of a vector of such raw attributes: $\mathbf{v}_A = \boldsymbol{\psi}(\mathbf{a}_A)$, where $\mathbf{a}_A$ is the vector of raw attributes of A.

The necessity of a nonlinear transformation can be understood by considering the standard choice model of the multinomial logit model (MLM), which is also known as the conditional logit model. In an MLM, the choice probability is given by

$$p(A|\mathscr{X}) = \frac{\exp(\mathbf{w}^\top \mathbf{v}_A)}{\sum_{X \in \mathscr{X}} \exp(\mathbf{w}^\top \mathbf{v}_X)}, \tag{1}$$

where $\mathbf{w}$ is a vector of the weights on the features. Here, $\mathbf{w}^\top \mathbf{v}_A$ can be understood as the attractiveness of A, and the choice probability is proportional to the exponentiated attractiveness. The nonlinear transformation, ψ, should be designed in a way that the attractiveness becomes linear with respect to the features. A popular practice is transformation to a binary vector of features, where each element of the binary vector denotes whether the value of a particular attribute is in a particular range [3, 7, 34]. Although such transformation to a binary vector is often hand-crafted, one could use systematic approaches such as locality sensitive hashing [1].

In the MLM, the ratio between the choice probabilities of two items, A and B, is independent of other alternatives. That is, we have

$$\frac{p(A|\mathscr{X})}{p(B|\mathscr{X})} = \frac{p(A|\mathscr{Y})}{p(B|\mathscr{Y})} \tag{2}$$

as long as $A, B \in \mathscr{X} \cap \mathscr{Y}$. Much of the research on choice models has been addressed to break this independence from irrelevant alternatives (IIA) [23], because the choice models having the property of IIA cannot represent the attraction effect and other typical phenomena of human choice.

Many of the choice models investigated in the literature fall into the class of random utility models [35]. Here, each item, A, is assumed to have a random utility, U_A. The probability of choosing A from $\mathscr{X}$ is given by the probability that U_A is larger than the random utility of any other items in $\mathscr{X}$:

$$p(A|\mathscr{X}) = \Pr(U_A > U_X, \forall X \in \mathscr{X} \setminus \{A\}). \tag{3}$$

The random utility model reduces to the MLM when $U_X = \mathbf{w}^\top \mathbf{v}_X + \varepsilon_X$ for each item X, where ε_X is an extreme value distribution that is independent of each other [15].

Random utility models must satisfy the principle of regularity [23], which states that the probability of choosing a particular item cannot be increased by adding alternatives:

$$p(A|\mathscr{X}) \geq p(A|\mathscr{Y}) \text{ if } \mathscr{X} \subset \mathscr{Y} \tag{4}$$

for $A \in \mathscr{X}$. This implies that the random utility models cannot represent the attraction effect, where adding a decoy, D, into a choice set, $\mathscr{X}$, increases the probability of choosing an item, A, that dominates D:

$$p(A|\mathscr{X}) < p(A|\mathscr{X} \cup \{D\}). \tag{5}$$

Although the majorities of the choice models studied in the literature [35] fall into the class of random utility models, researchers have studied the choice models outside this class. These include the decision field theory (DFT) [6, 24], the leaky competing accumulator model [37, 38], other sequential sampling models [22], and Bayesian choice models [27]. These models are not necessarily bounded by the regularity principle and have been shown numerically to represent the attraction effect and

other typical phenomena of human choice. The focus of this line of research is in the models and the interpretation of those models, and there has been little attempt to learn the parameters of those models from the data of human choice. For example, the hierarchical Bayesian model of Shenoy and Yu [27] might be trainable, but they do not discuss ways to train their model. In their experiments, they manually set the parameters of their model and show that the trained model represents three typical phenomena of human choice (the similarity effect, the compromise effect, and the attraction effect) in particular cases.

Recent work has investigated the choice models that can be trained to learn the typical phenomena of human choice from data [16, 17, 20, 32]. Osogami and Katsuki study a hierarchical Bayesian choice model with the concept of the visibility of items [16]. They also propose learning algorithms based on Gibbs sampling and approximate maximum a posteriori estimation and show that their choice model can be trained to learn the attraction effect using the experimental data of Ariely [2], which we have discussed in the beginning of this article. Takahashi and Morimura study a choice model of a Bayesian decision maker, who makes choices based on the posterior distribution of a Gaussian process after regression [32]. They estimate the parameters of a Gaussian process through convex optimization and show that their choice model can be trained to well represent the experimental data of Kivetz et al. [11], which involves the compromise effect (the phenomenon that humans tend to choose intermediate alternatives) [36]. In the following, we review the choice models studied by Otsuka and Osogami [17, 20].

Otsuka and Osogami propose a choice model, which they refer to as a deep choice model [17, 20]. The deep choice model is based on a restricted Boltzmann machine (RBM) and can be trained in a way that the log likelihood of given data is increased [12]. In [17], a special case of the deep choice model (an RBM choice model) is trained to learn the data of human choice about transportation means [4], which involves a phenomenon similar to the attraction effect. In [20], the deep choice model is trained based on an artificial dataset generated based on a scenario called a digit choice task, where a hypothetical agent chooses an image from a given set of images of handwritten digits from the MNIST dataset.[1] In the digit choice task of [20], the choice made by the hypothetical agent is designed to involve the attraction effect. That is, the deep choice model addresses two complexities in human choice. The first is the complex dependency of the human choice on available alternatives, and the second is the complex information that humans process to make a choice.

Figure 1 shows the architecture of the deep choice model studied in [20]. The deep choice model primarily consists of the stacked denoising autoencoder (SDA) [39] and the RBM [10, 29].

The SDA is a particular model of deep learning [26] and extracts features from images. An image from the MNIST dataset is in the gray scale and has the size of 28×28 bits, so that each image can be represented with a real vector of 784 dimensions. In [20], the SDA is used to extract real-valued features of 500 dimensions, which are then scaled and rounded into a binary vector of 500 dimensions. Let $\mathbf{v}_A$ be

[1] http://yann.lecun.com/exdb/mnist/index.html.

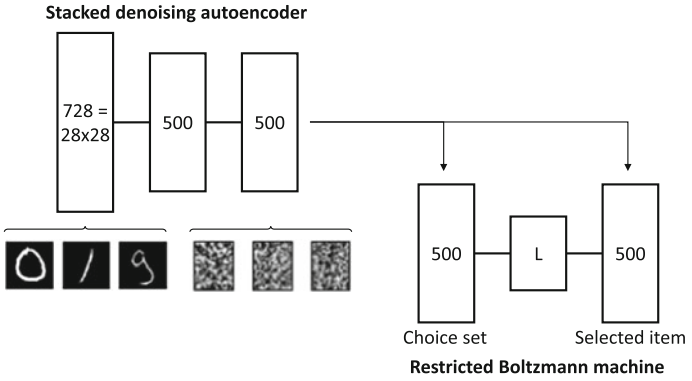


Fig. 1 A deep choice model studied in [20]. The integer in each rectangle represents the number of nodes in the corresponding layer, and L denotes the number of hidden nodes

the binary feature vector of an image, A. In [20], the feature vector of a choice set, $\mathscr{X}$, is defined as the average vector of the binary feature vectors of the items in $\mathscr{X}$:

$$\mathbf{v}_{\mathscr{X}} \equiv \frac{1}{|\mathscr{X}|} \sum_{X \in \mathscr{X}} \mathbf{v}_X, \tag{6}$$

which is not necessarily binary.

In [20], the RBM is used to define the choice probability on the basis of the feature vector of a choice set, $\mathscr{X}$, and the binary feature vector of each item, $X \in \mathscr{X}$. As is shown in Fig. 1, the RBM in the deep choice model consists of three parts of nodes. One part represents the choice set (choice-set nodes), and another part represents the selected item (selected-item nodes). These two parts constitute the visible layer of the RBM. Between the two parts of the visible layer, there is a layer of hidden nodes. The nodes in one layer are connected to the nodes in the other layer, but there are no connections within each layer. Let $\mathbf{W}$ be the matrix of the weight of the connections between the choice-set nodes and hidden nodes. Let $\mathbf{U}$ be the matrix of the weight between the selected-item nodes and hidden nodes. Let $\mathbf{b}_{\text{hid}}$ be the vector of the bias for the hidden nodes. Let $\mathbf{b}_{\text{out}}$ be the vector of the bias for the selected-item nodes.

The energy of the RBM in the deep choice model is then defined as

$$E(\mathbf{v}_{\mathscr{X}}, \mathbf{h}, \mathbf{v}_A) \equiv -\mathbf{v}_{\mathscr{X}}^{\top} \mathbf{W} \mathbf{h} - \mathbf{h}^{\top} \mathbf{U} \mathbf{v}_A - \mathbf{b}_{\text{hid}}^{\top} \mathbf{h} - \mathbf{b}_{\text{out}}^{\top} \mathbf{v}_A, \tag{7}$$

where $\mathbf{v}_{\mathscr{X}}$ and $\mathbf{v}_A$ are the values of the nodes in the visible layer, and $\mathbf{h}$ denotes the values of the hidden nodes.

In the deep choice model [20], the choice probability is then defined as

$$p(A|\mathscr{X}) = \frac{\exp(-F(\mathscr{X}, A))}{\sum_{X \in \mathscr{X}} \exp(-F(\mathscr{X}, X))}, \tag{8}$$

where

$$F(\mathscr{X}, A) \equiv -\log \sum_{\tilde{\mathbf{h}}} \exp(-E(\mathbf{v}_{\mathscr{X}}, \tilde{\mathbf{h}}, \mathbf{v}_A)) \tag{9}$$

denotes the corresponding free energy, where the summation with respect to $\tilde{\mathbf{h}}$ is over all of the possible configurations of the values of the hidden nodes.

Observe the similarities and the differences between the MLM (1) and the deep choice model (8). In the deep choice model, the feature vector of the choice set can affect what feature vector of an item makes the free energy of the RBM low, which in turn makes the relative choice probabilities of the items depend on the choice set. In particular, the deep choice model can break the regularity principle (4) and is shown to represent the attraction effect and other typical phenomena of human choice [17]. The hidden nodes of the deep choice model play the role of representing the particular dependency between the features of the items that are (un)likely to be selected.

3 Beyond Choice Models

Some of the choice models that we have seen allow us to make a prediction about what items are likely to be selected (or what features make an item more likely to be selected), depending on the choice set. The deep choice model illustrated in Fig. 1 motivates a more flexible model of giving choice probabilities over items, depending on various contexts (see Fig. 2). A choice set is an example of a context, but a context may be a sequence of the choice sets that are presented to a person, or a sequence of the choices that the person have made. The basic idea of the deep choice model can be used to model such choices that depend on the sequential context as long as the context is represented by a vector of a finite dimension.

When the sequential context is unbounded, however, a natural approach is to use a model that extends the Boltzmann machine to a stochastic process or a model of time-series data. The prior work has investigated such extensions of the Boltzmann machine in various ways. These include the spiking Boltzmann

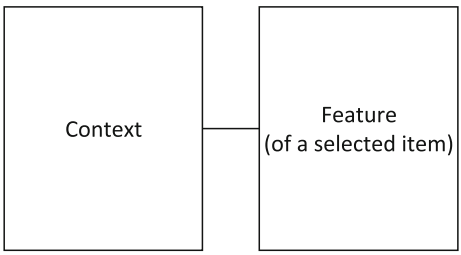


Fig. 2 A model of giving choice probabilities depending on the context

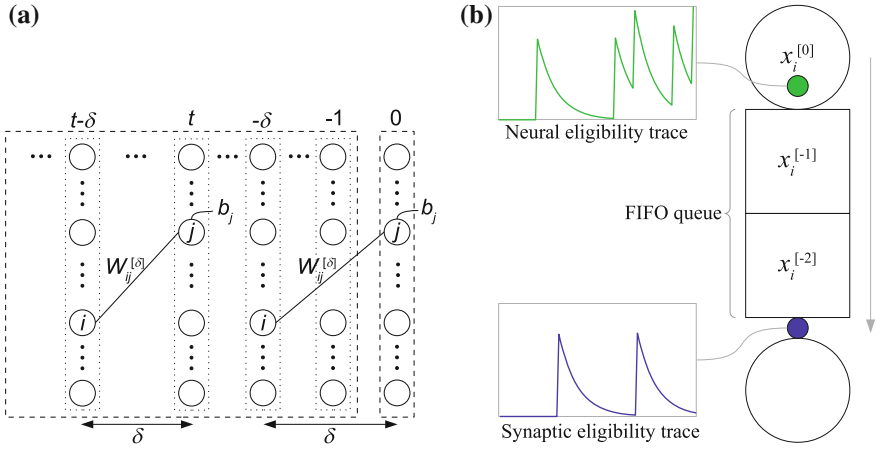


Fig. 3 **a** The dynamic Boltzmann machine unfolded through time [18]. **b** The dynamic Boltzmann machine [19]

machine [9], the temporal restricted Boltzmann machine [30], the temporal recurrent restricted Boltzmann machine [31], the factored conditional restricted Boltzmann machine [33], and the dynamic Boltzmann machine (DyBM) [18, 19].

Among these extensions of the Boltzmann machine, the DyBM has the particularly attractive property that its parameters can be trained to maximize the log likelihood of given sequential data through stochastic gradient methods under suitable assumptions. Figure 3a shows the DyBM when it is unfolded through time in a way that it corresponds the model of Fig. 2. In Fig. 3a, each column represents a pattern of time-series data at one moment. The right-most column represents the next pattern to be generated, and the probability distribution over the patterns depends on the preceding patterns (i.e., the context). It has been shown in [18] that the model in Fig. 3a is equivalent to the DyBM shown in Fig. 3b, where a node (or a neuron) is connected to another via a first-in-first-out (FIFO) queue, and each node holds the memory in the form of eligibility traces for storing some statistics of the preceding patterns (i.e., the context).

4 Toward Making Good Choices

While the objective of learning the parameters of the deep choice model [20] is to maximize the likelihood of given data of choice, the prior work has investigated ways to learn good policies of making decisions or choosing actions, depending on the context (i.e., the state or the history of prior actions and observations) for the models with the architecture shown in Fig. 2 [8, 21, 25]. In such prior work, the free energy is used to model the Q-function, or the state-action value function, which

represents the cumulative reward that can be obtained over a period of interest by taking a particular action from a particular state (or context). That is, their goal is to find the optimal policy for sequential decision- making, where an agent seeks to maximize the cumulative reward by sequentially choosing actions, depending on the outcomes of preceding actions. Besides the energy-based approaches of [8, 21, 25], the prior work has investigated various approaches of reinforcement learning [40] or planning [14] with (partially observable) Markov decision processes for sequential decision- making.

The techniques of sequential decision-making have been applied to actively learning the preferences of individuals, or preference elicitation [5], which are closely related to the parameters of a choice model. Indeed, when we interact with individuals in a sequential manner, we can learn the preferences of the individuals, or the parameters of their choice models, during the interaction to adapt our actions in consideration of their preferences. During interaction, we can take actions that are targeted primarily to collect informative observations or primarily to provide good services immediately or in the future. This interaction can be considered as sequential decision- making, where we select appropriate actions in sequence to achieve a long-term goal of providing good personalized services.

In our preliminary work, we apply an algorithm for sequential decision-making [28] to sequentially learn the parameters of a choice model for a person who we are interacting with. Specifically, our goal is to understand the preferences of a potential customer (here referred to as a user) about life-insurance. We ask questions to the user, where a question consists of a set of multiple life-insurance products. The user then answers the question by selecting one from the choice set. Here, we consider four attributes of a life-insurance product: the amount of monthly premium, the amount of total premium, the amount of coverage, and the amount of refund (in the case the user is alive after the period of coverage). We vary the values of these attributes to maximally elicit the preferences of the user.

We conducted an experiment of asking five questions to each user. The cost of responding to the five questions is sufficiently low, and 97 % of the users responded to all of the five questions. We evaluated the quality of the questions optimized through sequential decision making by predicting the users' response to the fifth question from the users' preferences that were estimated on the basis of the first four questions and responses to them. The area under the ROC curve was improved from 0.59 to 0.75 by our approach.

Acknowledgments A part of this research was supported by CREST, JST.

References

1. Andoni, A., Indyk, P.: Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. Commun. ACM **51**(1), 117–122 (2008)
2. Ariely, D.: Predictably Irrational: The Hidden Forces That Shape Our Decisions. Harper Perennial, revised and expanded edition (2010)

3. Arora, N., Huber, J.: Improving parameter estimates and model prediction by aggregate customization in choice experiments. Consum. Res. **28**, 273–283 (2001)
4. Bierlaire, M., Axhausen, K., Abay, G.: The acceptance of modal innovation: the case of swissmetro. In: Proceedings of the First Swiss Transportation Research Conference, March 2001
5. Boutilier, C.: A POMDP formulation of preference elicitation problems. In: Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI-02), pp. 239–246 (2002)
6. Busemeyer, J.R., Townsend, J.T.: Decision field theory: a dynamic cognition approach to decision making. Psychol. Rev. **100**, 432–459 (1993)
7. Chapelle, O., Harchaoui, Z.: A machine learning approach to conjoint analysis. In: Saul, L.K., Weiss, Y., Bottou, L. (eds.) Advances in Neural Information Processing Systems 17, pp. 257–264. MIT Press, Cambridge (2005)
8. Heess, N., Silver, D., Teh, Y.W.: Actor-critic reinforcement learning with energy-based policies. J Mach. Learn. Res. Workshop Conf. Proc. **24**, 45–58 (2012)
9. Hinton, G.E., Brown, A.D.: Spiking Boltzmann machines. In: Advances in Neural Information Processing Systems, pp. 122–128 (1999)
10. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **313**(5786), 504–507 (2006)
11. Kivetz, R., Netzer, O., Srinivasan, V.S.: Alternative models for capturing the compromise effect. J. Mark. Res. **41**(3), 237–257 (2004)
12. Larochelle, H., Bengio, Y.: Classification using discriminative restricted Boltzmann machines. In: Proceedings of the 25th International Conference on Machine Learning, pp. 536–543 (2008)
13. Luce, R.D.: Individual Choice Behavior: A Theoretical Analysis. Wiley, New York (1959)
14. Mausam, Kolobov, A.: Planning with Markov Decision Processes: An AI Perspective. Morgan & Claypool Publishers, San Rafael (2012)
15. McFadden, D.: Conditional logit analysis of qualitative choice behavior. In: Frontiers in Econometrics, pp. 105–142. Academic Press, New York (1974)
16. Osogami, T., Katsuki, T.: A Bayesian hierarchical choice model with visibility. In: Proceedings of the 22nd International Conference on Pattern Recognition, pp. 3618–3623, August 2014
17. Osogami, T., Otsuka, M.: Restricted Boltzmann machines modeling human choice. Adv. Neural Inf Process Syst **27**, 73–81 (2014)
18. Osogami, T., Otsuka, M.: Learning dynamic boltzmann machines with spike-timing dependent plasticity. Technical report RT0967, IBM Research (2015)
19. Osogami, T., Otsuka, M.: Seven neurons memorizing sequences of alphabetical images via spike-timing dependent plasticity. Sci Rep. **5**, 14149 (2015)
20. Otsuka, M., Osogami, T.: A deep choice model. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16) (2015)
21. Otsuka, M., Yoshimoto, J., Doya, K.: Free-energy-based reinforcement learning in a partially observable environment. In: Proceedings of European Symposium on Artificial Neural Networks—Computational Intelligence and Machine Learning, pp. 28–30, April 2010
22. Otter, T., Johnson, J., Rieskamp, J., Allenby, G.M., Brazell, J.D., Diederich, A., Hutchinson, J.W., MacEachern, S., Ruan, S., Townsend, J.: Sequential sampling models of choice: some recent advances. Mark. Lett. **19**(3–4), 255–267 (2008)
23. Rieskamp, J., Busemeyer, J.R., Mellers, B.A.: Extending the bounds of rationality: evidence and theories of preferential choice. J. Econ. Lit **44**, 631–661 (2006)
24. Roe, R.M., Busemeyer, J.R., Townsend, J.T.: Multialternative decision field theory: a dynamic connectionist model of decision making. Psychol. Rev. **108**(2), 370–392 (2001)
25. Sallans, B., Hinton, G.E.: Reinforcement learning with factored states and actions. J. Mach. Learn. Res. **5**, 1063–1088 (2004)
26. Schmidhuber, J.: Deep learning in neural networks: an overview. Neural Netw. **61**, 85–117 (2015)
27. Shenoy, P., Yu, A.J.: Rational preference shifts in multi-attribute choice: what is fair? In: Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci 2013), pp. 1300–1305 (2013)

28. Silver, D., Veness, J.: Monte-Carlo planning in large POMDPs. Adv. Neural Inf. Process. Syst. **23**, 2164–2172 (2010)
29. Smolensky, P.: Information processing in dynamical systems: foundations of harmony theory. In: Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations, vol. 1, Chap. 6, pp. 194–281. MIT Press (1986)
30. Sutskever, I., Hinton, G.E.: Learning multilevel distributed representations for high-dimensional sequences. In: International Conference on Artificial Intelligence and Statistics, pp. 548–555 (2007)
31. Sutskever, I., Hinton, G.E., Taylor, G.W.: The recurrent temporal restricted Boltzmann machine. In: Advances in Neural Information Processing Systems, pp. 1601–1608 (2008)
32. Takahashi, R., Morimura, T.: Predicting preference reversals via Gaussian process uncertainty aversion. In: Proceedings of the 18th International Conference on Artificial Intelligence and Statistics, May 2015
33. Taylor, G.W., Hinton, G.E.: Factored conditional restricted Boltzmann machines for modeling motion style. In: Proceedings of the 26th International Conference on Machine Learning (ICML 2009), pp. 1025–1032 (2009)
34. Toubia, O., Hauser, J.R., Simester, D.I.: Polyhedral methods for adaptive choice-based conjoint analysis. Mark. Res. **41**, 116–131 (2004)
35. Train, K.: Discrete Choice Methods with Simulation, 2nd edn. Cambridge University Press, Berkeley (2009)
36. Tversky, A., Simonson, I.: Context-dependent preferences. Manage. Sci. **39**(10), 1179–1189 (1993)
37. Usher, M., McClelland, J.L.: The time course of perceptual choice: the leaky, competing accumulator model. Psychol. Rev. **108**, 550–592 (2001)
38. Usher, M., McClelland, J.L.: Loss aversion and inhibition in dynamical models of multialternative choice. Psychol. Rev. **111**(3), 757–769 (2004)
39. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A.: Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn. Res. **11**, 3371–3408 (2010)
40. Wiering, M., van Otterlo, M. (eds.) Reinforcement Learning: State-of-the-Art. Springer (2012)

On 3D Scanning Technologies for Respiratory Mask Design

Dmitry Nikolayevich Znamenskiy

Abstract Within the Philips Research project a handheld, 3D face scanner has been developed to address the needs of CPAP mask design for Philips Respironics business unit. The scanner is based on the structure light technology proposed in [10], which is claimed to be motion robust, i.e. in typical conditions with shaky hands and moving objects, the scanner delivers sub-millimetre accurate 3D face models, suitable for the CPAP mask design applications. In this article we derive an analytic expression for the accuracy of the structured light scanner, where the lateral and axial measurement errors as a function of the hardware parameters and the object position and velocity. The analytic formulas can contribute to better understanding the motion invariant structured light technology and creates a room for the scanner specifications.

Keywords Apnoea · 3D scanning · CPAP

1 Introduction

Apnea Philips makes respiratory masks for patients with Obstructive Sleep Apnoea syndrome (OSA). OSA is a sleep disorder when people frequently stop breathing during the night due to closure of the upper airway, so people partially wake up many times during the night, which causes continual sleepiness during the day and other health disorders, see [1]. Sleep studies [2] show that 6–7% of western population suffer from at least a mild form of apnoea, where almost 85 % of the cases remains undiagnosed and untreated. Male gender, age, overweight, low muscle tone and snoring can increase the likelihood of apnoea up to 40 %.

Mask Design Since 1980, sleep apnoea is effectively treated [3] (but not cured) by providing positive air pressure which prevents the upper airway from obstruction. The positive air pressure (CPAP or Bi-PAP) is generated by a pump and delivered by means of a tube and a facial mask to a patient. As the patient is expected to sleep with

D.N. Znamenskiy (✉)
Philips Research, HTC36, 5656AE, Eindhoven, Netherlands
e-mail: Dmitry.Znamenskiy@philips.com

B. Anderssen et al. (eds.), *The Role and Importance of Mathematics in Innovation*, Mathematics for Industry 25,
DOI 10.1007/978-981-10-0962-4_2

the mask every might, the mask should ultimately fit the patient's face. If mask does not fit the face, a patient can get red marks where the mask contact is to tense, or air leaks where the mask contact is too loose. The air leaks reduce the efficiency of the therapy and, if the mask is leaking towards an eye, it can cause an eye inflammation. About 40 % of the patients stop with the treatment, due of problems with the mask [4].

Studying the facial dimensions of the average apnoea patient is a critical task in the making of small lightweight masks which would perfectly fit the patient.

While it is possible to find the average facial dimensions in various anthropometric surveys, e.g. like [5], the manual anthropometric measurements are often inaccurate, see [7], and the majority of studies does not address the specific subgroup and ethnicities of the OSA population. Moreover, the knowledge of the average facial dimensions is often not sufficient, as not only dimensions but also the face shape significantly varies per population group.

3D Scanning with structured light The respiratory mask design can considerably benefit form 3D scanning surveys, like [6], which collect data over the complete facial surface. While the 3D scanning technologies already became almost a commodity, it is hardly possible to find a commercial 3D camera for fast, robust and mobile data collection: a lightweight handheld device with acquisition time of less than a second, motion robust yet giving sub-millimetre accurate measurements. It was expected that the sub-millimetre scanner accuracy being comparable with the amplitude of the skin defects will result in sufficiently accurate average faces for meaningful mask design. Thus within Philips Research a new 3D face scanning technology was proposed, see [10] which could potentially meet the required specifications. Among different scanning methods [8], the so-called *structured light* was chosen.

Structured light method is based on the projection of a known pattern with a projector on a scene, and capturing of the resulting image with a camera of the scene. The camera is laterally displaced with respect to projector on the distance called the *baseline*. Similar to a stereo-camera, the system works on the basis of disparity as the camera captures laterally displaced codes where the amount of displacement depends on the distance to the object, see Fig. 1. Further the required depth resolution was

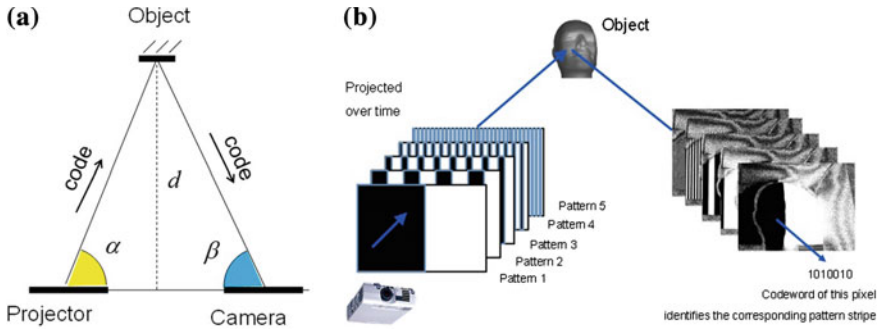


Fig. 1 **a** Triangulation principle, **b** a schematic of the structured light principle

gained using the sub-pixel accurate edge codes [9] and the motion invariance was achieved by alternating the polarity of the projected code images and using the edge addressing based on binary codes, where the position of decoded edges can be tracked from frame to frame, see [10]. The article is further structured as follows: In the next section we will present the main result. Then in Sect. 3 we will give an outline of the proof. Section 4 shows an application of the accuracy formulas to an example 3D scanner configuration. Section 5 describes embodiments of the structured light technology in Philips products. Appendix addresses the missing details.

2 Structured Light Accuracy

In order to better understand the potential of the technology and find the optimal scanner configuration, we invested time in derivation of an analytic expression for the accuracy of the scanner as a function of the hardware parameters and the object position and velocity. The scanner errors can be then translated to the accuracy of the average face computed over a collection of scans.

Consider the scanner and object optical system parameters: h-pixel size, Z_0—focusing distance, A—camera aperture size, f_c- camera focal length, N_r—sensor signal to noise ratio in percent of the camera dynamic range, F_s—frame rate, $C_{exp} \leq 1$—camera sensor image integration duty cycle, Z—distance to object, B—baseline between the camera and projector, V_x—relative lateral object velocity, V_z—relative object axial velocity, T_r—visible object texture intensity gradient in percent of dynamic range, see Fig. 2.

Proposition *The lateral scanner error E_X and the axial scanner error E_Z can be expressed as*

$$E_X \approx \frac{E \cdot Z}{f_c}, \qquad E_Z \approx \frac{E \cdot Z^2}{B \cdot f_c},$$

where E is the total system error which can be decomposed in the systematic and stochastic error

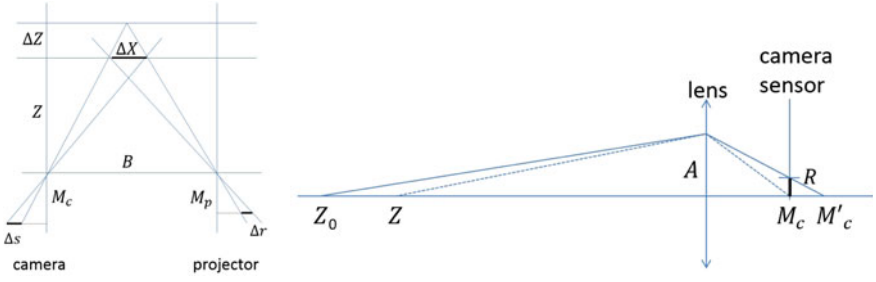


Fig. 2 Geometric parameters and camera blur

$$E = \bar{E} + E_n, \qquad \bar{E} \approx \frac{T_r \cdot V_X \cdot f_c}{2F_s \cdot Z} R, \qquad \bar{|E_n|} \leq \frac{N_r \cdot 3n}{2} R$$

where $n \sim N(0, 1)$ *is the Gaussian random variable and* R *is the system blur radius*

$$R = h + Af_c \frac{|Z - Z_0|}{Z \cdot Z_0} + \frac{|V_Z| \cdot B \cdot f_c \cdot C_{exp}}{F_s \cdot Z^2},$$

when the following two assumptions are satisfied:
(a) there are bounds on the relative lateral object speed V_X *and relative axial object speed* V_Z

$$|V_Z| \ll A \frac{|Z - Z_0| 2F_s \cdot Z}{Z_0 \cdot B}, \qquad |V_X| \ll A \frac{|Z - Z_0|}{Z_0} + \frac{|V_Z| \cdot B}{2F_s \cdot Z},$$

(b) the local texture gradient is not larger than the inverse of the system blur radius, i.e. $T_r \leq R^{-1}$.

3 Sketch of the Proof

The proof presented in this section is not rigorous/complete from the mathematical point of view, therefore we call it a 'sketch of the proof'. Below, we consider the projection and acquisition of the signal binary edge signal as the tracking of edges between the subsequent code frames is beyond the scope of the Proposition. The proof of the Proposition includes: (a) modelling of the optical system, (b) modelling of the camera signal, (c) modelling of the edge error and (d) application of the models.

Modelling of the optical system Below, we consider two Lemmas. The first one gives the relations between the real-world object localization error (object speed) versus the edge localization error (edge speed) on the camera sensor.

Lemma 1 *When* $\Delta Z \ll Z$ *and* $M_c \approx f_c$*, the object localization lateral error* E_X*, depth error* E_Z*, lateral object velocity* V_X *and axial object velocity* V_Z *are related to the edge localization error* E *and to the edge speed* v *on the camera sensor via the following equations:*

$$E_X \approx \frac{E \cdot Z}{f_c}, \qquad V_X \approx \frac{v_x \cdot Z}{f_c}, \qquad v_x \approx \frac{V_X \cdot f_c}{Z}, \qquad (a)$$

$$E_Z \approx \frac{E \cdot Z^2}{B \cdot f_c}, \qquad V_Z \approx \frac{v_z \cdot Z^2}{B \cdot f_c}, \qquad v_z \approx \frac{V_Z \cdot B \cdot f_c}{Z^2}, \qquad (b)$$

$$\Delta Z \approx \frac{\Delta r \cdot Z^2}{B \cdot M_p}. \qquad (c)$$

where v_x and v_z are the lateral edge speeds on the sensor caused respectively by the lateral and the axial object velocities.

The proof of lemma is given in appendix.

In the second lemma we give expression for the blur radius on the camera sensor as the function of the system parameters.

Lemma 2 *When $\Delta Z \ll Z$, $M_c \approx f_c$ and $f_c \ll Z_0$, the object appears blurred on the sensor with the blur radius*

$$R = h + Af_c \frac{|Z - Z_0|}{Z \cdot Z_0} + \frac{|V_Z| \cdot B \cdot f_c \cdot C_{exp}}{F_s \cdot Z^2},$$

Observe that the condition of the lemma is practically satisfied as for a typical small camera we have f_c equal to few mm, while Z and Z_0 are about tens of centimetres The proof of lemma is given in appendix.

Modelling of the edge signal Consider 1D camera sensor signal $I(s)$, $s \in \mathbb{R}$ which is sampled in the direction parallel to the baseline between the camera and the projector. We model $I(s)$ as the product of projected code signal $S(s)$ and the texture reflectivity $T(s)$, plus noise

$$I(s) = S(s) \cdot T(s) + \sigma \cdot n(s),$$

where $n(s) \sim N(0, 1)$ is the Gaussian random variable. Assume that the ideal edge position, at the absence of motion, texture crosstalk and noise is at $s = 0$ and that neighbouring pixel positions are at $s = h$, $s = -h$, where $2h$ is the distance between any two pixels on the sensor. At the absence of ambient light, object motion and sharply in the focus, the projected code signal can be modelled as

$$S(s) = \mathbb{1}(s \geq 0),$$

where $\mathbb{1}$ is the indicator function. In practical conditions, when the signal is superimposed over the ambient light S_0, and edge is acquired blurred with radius R, we model the signal $S(s)$ as

$$S(s) = a \cdot s + 0.5 + S_0, \qquad a = R^{-1}.$$

If object is moving in axial direction, the acquired signal appears displaced with speed v_z

$$S(s + v_z t) = a \cdot (s + v_z \cdot t) + 0.5 + S_0.$$

Note that the above model can be considered valid until

$$0 \leq S(s + v_z t) \leq 1. \tag{1}$$

We address the texture crosstalk effect with the simple linear model of the reflectivity

$$T(s) = c \cdot s + d,$$

where c, d are parameters such that $T(s) \in [0, 255]$. If an object is moving in lateral direction, the acquired texture appear displaced with speed v_x

$$T(s + v_x \cdot t) = c \cdot (s + v_x t) + d.$$

Combining the above edge and the camera model we have

$$\begin{aligned} I(s,t) &= S(s + v_z \cdot t) \cdot T(s + v_x \cdot t) + \sigma \cdot n(s). \\ &= \big(a \cdot (s + v_z \cdot t) + 0.5 + S_0\big) \cdot \big(c \cdot (s + v_x t) + d\big) + N_r d \cdot n(s), \end{aligned} \tag{2}$$

with $N_r = \sigma / d$.

Modelling of the edge error As we mentioned in the introduction we consider a coded light scheme where for every binary image pattern at time t we also project a negative of it at time $-t$. The edge signal in the positive-phase and negative-phase camera images can be modelled as

$$\begin{aligned} I_+(s,t) &= S(s - v_z \cdot t) \cdot T(s - v_x \cdot t) + \sigma \cdot n_+(s) \\ I_-(s,-t) &= S(-(s + v_z \cdot t)) \cdot T(s + v_x \cdot t) + \sigma \cdot n_-(s) \end{aligned} \tag{3}$$

where in the negative-phase camera images the edge signal is inverted while the texture signal stays the same. In order to cancel the effect of the unknown ambient light and minimize the influence of the local reflectivity on the exact position of the edge we normalize it as

$$I_n(s,t) = \frac{I_+(s,t) - I_-(s,-t)}{I_+(s,t) + I_-(s,-t)}. \tag{4}$$

The sub-pixel accurate observed edge position, i.e. edge error E, can be found by means of a linear interpolation of the normalized signal between the pixel positions $s = h$ and $s = -h$

$$E = h \cdot \frac{I_n(-h,t) + I_n(h,t)}{I_n(-h,t) - I_n(h,t)}, \tag{5}$$

see Fig. 3. Note, that the above approximation can be considered accurate until the estimated edge is located inside the pixel range

$$-h < E < h. \tag{6}$$

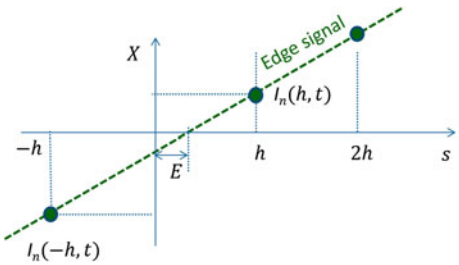


Fig. 3 Computation of the edge position E

Application of the models Observe that, due of Lemma 1, it is sufficient to prove under conditions of the Proposition that

$$\bar{E} \approx R\frac{c \cdot v_x \cdot t}{d}, \tag{7}$$

with $t = 1/(2F_s)$, and that

$$|E_n| = |E - \bar{E}| \leq R\frac{3N_r \cdot |n|}{2} \tag{8}$$

We prove first (7) and then (8). The first step towards this is the following lemma.

Lemma 3 *For E defined in (5) holds*

$$E = h \cdot \frac{I_+(-h,t) \cdot I_+(h,t) - I_-(-h,-t) \cdot I_-(h,-t)}{I_+(-h,t) \cdot I_-(h,-t) - I_-(-h,-t) \cdot I_+(h,t)}. \tag{9}$$

Proof The proof is quite straightforward. Substitute (4) in (5).

Let I_{pp}, I_{mp}, I_{pm} and I_{mm} denote $I_+(h,t), I_-(h,-t), I_+(-h,t)$ and $I_-(-h,-t)$ without noise, i.e.

$$\begin{aligned} I_{pp} &= S(h - v_z t)T(h - v_x t) \\ I_{pm} &= S(-h - v_z t)T(-h - v_x t) \\ I_{mp} &= S(-h - v_z t)T(h + v_x t) \\ I_{mm} &= S(h - v_z t)T(-h + v_x t) \end{aligned} \tag{10}$$

Below we define the systematic edge error which will be used to prove (7) and (8).

$$\bar{E} = h \cdot \frac{I_{pm}I_{pp} - I_{mm}I_{mp}}{I_{pm}I_{mp} - I_{mm}I_{pp}}. \tag{11}$$

Let us make some change of variables

$$\mu = v_z \cdot t, \qquad \nu = v_x \cdot t, \qquad \beta = d/c = T_r^{-1}.$$

The following lemma is then used to reduce $\bar{E}$

Lemma 4

$$(a)\quad I_{pm}I_{pp} - I_{mm}I_{mp} = 4d^2T_r\nu(1 - a^2h^2 - 2a\mu + a^2\mu^2)$$
$$(b)\quad I_{pm}I_{mp} - I_{mm}I_{pp} = h \cdot 4d^2\big(a - a^2\mu + T_r^2(\nu - ah^2 - 2a\mu\nu - a\nu^2 + a^2h^2(\mu + \nu) + a^2\mu^2\nu + a^2\mu\nu^2)\big) \tag{12}$$

Proof The proof is quite straightforward, but tediuos: use the suggested above change of variables and substitute (10) on the left-hand side of the lemma to get the expressions after simplification.

It follows from Lemma 4 that

$$\bar{E} = \frac{T_r\nu(1 - a^2h^2 - 2a\mu + a^2\mu^2)}{(a - a^2\mu + T_r^2(\nu - ah^2 - 2a\mu\nu - a\nu^2 + a^2h^2(\mu + \nu) + a^2\mu^2\nu + a^2\mu\nu^2)} \approx \frac{T_r\nu}{a} = R\frac{c \cdot v_x \cdot t}{d},$$

under assumptions of the Proposition which gives (7). In order to prove (8) we have to compare (9) and (11). The following Lemma compares respectively the nominators and denominators in (9) and (11).

Lemma 5 *Under assumptions of the Proposition we have*

$$\big(I_+(-h,t)\cdot I_+(h,t) - I_-(-h,-t)\cdot I_-(h,-t)\big) - \big(p_{pm}p_{pp} - p_{mm}p_{mp}\big) \approx 2d\sigma \cdot n_1,$$

and, similarly,

$$\big(I_+(-h,t)\cdot I_-(h,-t) - I_-(-h,-t)\cdot I_+(h,t)\big) - \big(p_{pm}p_{mp} - p_{mm}p_{pp}\big) \approx 2d\sigma \cdot n_2,$$

where n_1, n_2 are normal distributed random variables with variances at most 1.

The proof of Lemma 4 is given in the appendix. It follows then from Lemma 5 that

$$E \approx h\frac{4d^2T_r\nu + 2d\sigma \cdot n_1}{h \cdot 4d^2a + 2d\sigma \cdot n_2} = h\frac{2dT_r\nu + \sigma \cdot n_1}{h \cdot 2da + \sigma \cdot n_2}.$$

In order to proceed with proving (8) we need another lemma.

Lemma 6 *Let $n_1, n_2 \sim N(0,1)$ are normal distributed random variables. Under condition $n_1, n_2 \ll A, B$ and $A/B \le h$, there exist normal distributed random variable $n_3 \sim N(0,1)$ such that,*

$$\left|\frac{A + \sigma_1 \cdot n_1}{h \cdot B + \sigma_1 \cdot n_2} - \frac{A}{B}\right| \le \frac{3\sigma_1 \cdot |n_3|}{B}.$$

We apply Lemmas 4 and 6 with

$$A = 2dT_r v,$$

$$B = 2da,$$

to get (8):

$$|E_n| = |E - \bar{E}| \approx \left| h \cdot \frac{p_{pm}p_{pp} - p_{mm}p_{mp} + 2d\sigma \cdot n_1}{p_{pm}p_{mp} - p_{mm}p_{pp} + 2d\sigma \cdot n_2} - h \cdot \frac{p_{pm}p_{pp} - p_{mm}p_{mp}}{p_{pm}p_{mp} - p_{mm}p_{pp}} \right|$$

$$|E_n| \leq \frac{3\sigma \cdot |n_3|}{2d \cdot a} \approx R\frac{3N_r \cdot |n_3|}{2},$$

with $N_r = \sigma/d$, which completes the proof of the Proposition. □

4 Simulations

Consider, as example, a 3D scanner model with the following optical and system parameters. The camera sensor pixel size is 6×10^{-6} [m] which gives half of the distance between the pixels $h = 3 \times 10^{-6}$ [m]. The camera has the focal length $f_c = 0.008$ [m], F—number 2.5 and therefore the aperture $A = f_c/2.5 = 0.0032$ [m]. The camera is focused at distance $Z_0 = 0.4$ [m], and the active range of the scanner is $0.3\,[\text{m}] < Z < 0.7\,[\text{m}]$. The magnification of the camera $M_c = \frac{Z_0 \cdot f_c}{Z_0 - f_c} \approx f_c = 0.008$ [m]. Further we assume that the reflectivity change between the neighbouring pixels is 10 %. Hence $T_r/(2\text{h}) = 0.1$ and $T_r = 0.05/\text{h} \approx 1.7 \times 10^4$ [m]. The typical sensor SNR can be taken as $N_r = 1\,\%$ at the focusing distance $Z_0 = 0.4$ [m], and it grows as fourth power of distance to the object (a multiplication of the second power decay of projected intensity and second power decay of reflected intensity), which gives $N_r = (0.7/0.4)^4 1\,\% \approx 9.3\,\%$ at $Z = 0.7$ [m] and $N_r = (0.3/0.4)^4 1\,\% \approx 0.3\,\%$ at $Z = 0.3$ [m]. Take the baseline between the camera and projector $B = 0.07$ [m]. The camera frame rate is $F_s = 60\,[\text{s}^{-1}]$ and duty cycle 50 %, which gives the temporal image integration time of $120^{-1}\,[\text{s}^{-1}]$. Below we have applied the Proposition to make the level plot for the position error as a function of the lateral and axial object velocities. The set of first plots show the lateral and axial errors when the object is in focus at distance $Z_0 = 0.4$ [m]. One can see that the stochastic error variance is negligible (Fig. 4).

The second set of the plots shows the lateral and axial errors when the object is at distance $Z = 0.5$ [m]. The stochastic error variance for axial error is about millimetre and cannot be ignored anymore (Fig. 5).

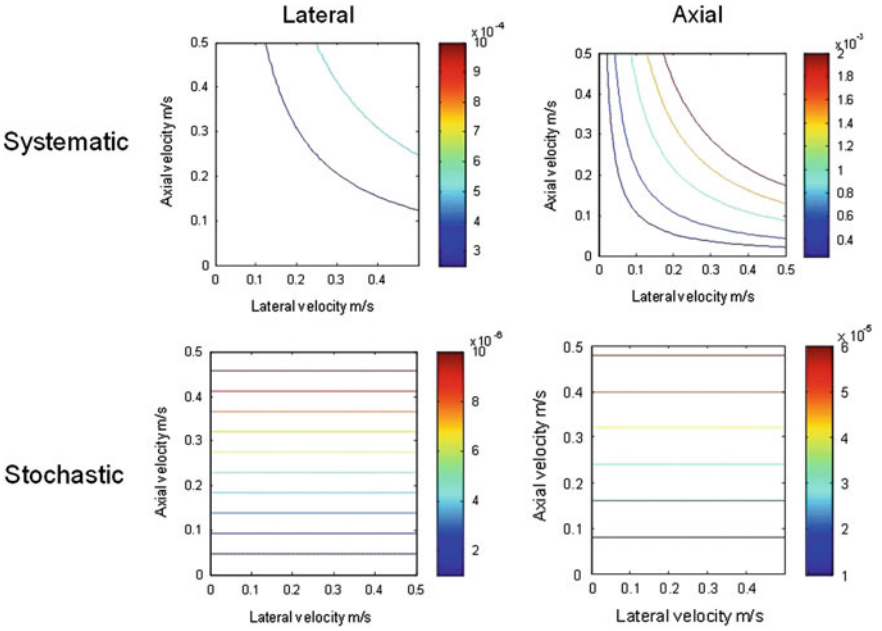


Fig. 4 Contour plots for scanning distance of 0.4 m

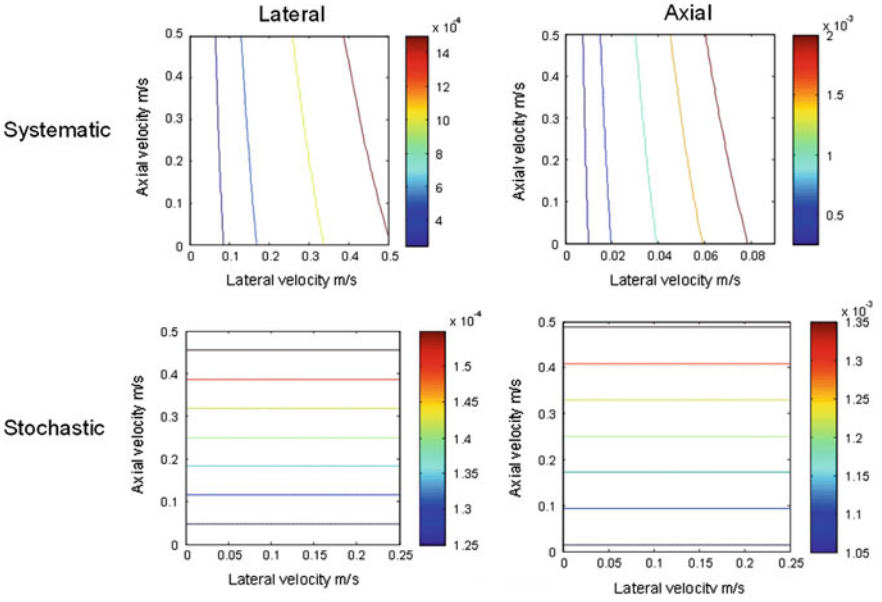


Fig. 5 Contour plots for scanning distance of 0.5 m

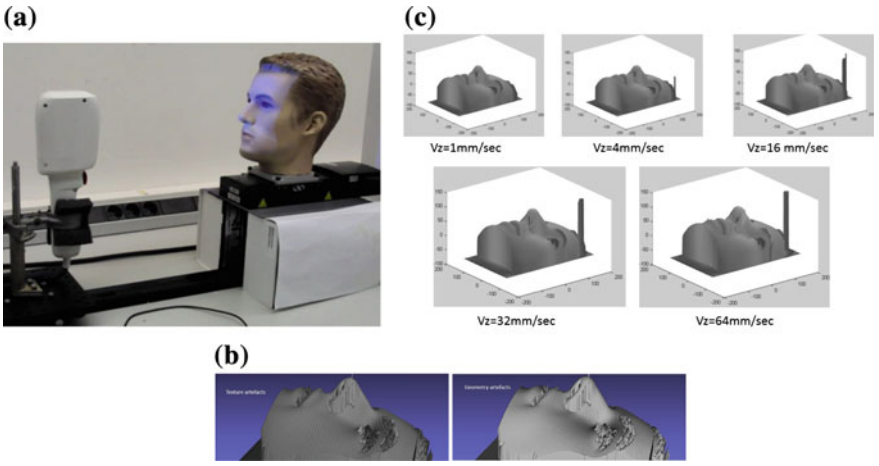


Fig. 6 **a** Scanner evaluation on the translation stage, **b** captured 3D image for different velocities, **c** artefacts at the maximal velocity

5 Realization in Products

On the practical side the new 3D scanning technologies were realized in a handheld 3D scanner Fig. 7a. The robustness of the scanner to the object motion was evaluated in a series of experiments where a manikin head was scanned while it was moving with certain velocity. The exact axial velocity was controlled by means of a translation stage, see Fig. 6. The 3D scans created with the scanner have sub-millimetre errors while the relative subject motion has the lateral and axial velocity of 5 and 20 cm/s respectively. Philips used 3D scanners to measure the OSA patients in Japan and design a special mask for Japanese OSA population, see Fig. 7b.

Acknowledgments The authors are grateful to Philips Respironics for providing a challenging topic of research, and to colleagues Ruud Vlutters and Karl van Bree who are co-authors of the motion invariant structured light principle [10].

Appendix

Proof of Lemma 1. Consider the left sketch on Fig. 2. From the similarity of triangles in the figure one can derive that

$$\frac{\Delta X}{Z} = \frac{\Delta s}{M_c} \tag{13}$$

$$\frac{\Delta Z}{\Delta X} = \frac{Z + \Delta Z}{B} \tag{14}$$

Fig. 7 **a** Philips internal 3D scanner, **b** a special version of Philips Wisp mask produced for Japanese market

The first one for $M_c \approx f_c$ implies

$$\Delta X = \frac{\Delta s \cdot Z}{M_c} \approx \frac{\Delta s \cdot Z}{f_c}, \tag{15}$$

and

$$\Delta s = \frac{\Delta X \cdot M_c}{Z} \approx \frac{X f_c}{Z} \tag{16}$$

Thus, if the object is displaced in the lateral direction we have point (a) of the Lemma:

The combination of the (14) and (15) implies

$$\Delta Z = \frac{\Delta s \cdot Z \cdot (Z + Z)}{B \cdot M_c} \tag{17}$$

If we assume that $\Delta Z \ll Z$, and $M_c \approx f_c$ we can approximate

$$\Delta Z \approx \frac{\Delta s \cdot Z^2}{B \cdot f_c}, \qquad \Delta s \approx \frac{\Delta Z \cdot B \cdot f_c}{Z^2}, \tag{18}$$

which implies points (b) and (c) of the Lemma. Thus, if the object is moving in the axial direction we have point (a) of the Lemma. The proof of point (c) can be obtained by flipping the camera and the projector sides. □

Proof of Lemma 2. We model the blur radius as the sum of the pixel blur, the optical blur and the motion blur.

$$R = R_h + R_o + R_m.$$

We assume the pixel blur equal to h, i.e. $R_h = h$. Consider first the optical blur. Suppose than the camera is focused at distance Z_0. Then we have from the lens equation

$$\frac{1}{Z_0} + \frac{1}{M_c} = \frac{1}{f_c}$$

Hence

$$M_c = \frac{Z_0 \cdot f_c}{Z_0 - f_c}.$$

If the object is located at distance Z, then the image is focused at distance

$$M_c^{'} = \frac{Z \cdot f_c}{Z - f_c}.$$

Then the object appears blurred on the sensor with the blur radius:

$$R_o = \frac{A}{M_c^{'}} |M_c^{'} - M_c| = A f_c \frac{|Z - Z_0|}{Z(Z_0 - f_c)} \approx A f_c \frac{|Z - Z_0|}{Z \cdot Z_0},$$

since $f_c \ll Z_0$. Consider the motion blur part. When the object is moving in the axial direction it causes the acquired edge move in laterally on the sensor, and the edge displacement R_m is equal to the absolute edge velocity $|v_z|$ times the exposure time T_{exp}:

$$R_m = v_z \cdot T_{exp} \approx \frac{|V_Z| \cdot B \cdot f_c \cdot C_{exp}}{Z^2 \cdot F_s},$$

where we apply Lemma 1, and where $T_{exp} = C_{exp}/F_s$ □

Proof of Lemma 5. It follows from the definitions of $I_+(h, t)$, $I_-(h, -t)$, $I_+(-h, t)$, $I_-(-h, -t)$ and I_{pp}, I_{mp}, I_{pm}, I_{mm}, and from the independence of $n_+(h)$, $n_+(-h)$, $n_-(h)n_-(-h)$ that

$$\begin{aligned}
&\big((I_+(-h, t) \cdot I_+(h, t) - I_-(-h, -t) \cdot I_-(h, -t)\big) - \big(I_{pm} I_{pp} - I_{mm} I_{mp}\big) \\
&= I_{pm}\sigma \cdot n_+(h) + I_{pp}\sigma \cdot n_+(-h) + I_{mm}\sigma \cdot n_-(h) - I_{mp}\sigma \cdot n_-(-h) \\
&= \sqrt{I_{pm}^2 + I_{pp}^2 + I_{mm}^2 + I_{mp}^2}\sigma \cdot n_1, \\
&\approx \sqrt{4a^2c^2R^2\beta^2}\sigma \cdot n_1 = 2d \cdot \sigma \cdot n_1,
\end{aligned}$$

for some $n_1 \sim N(0, 1)$. Similarly we get the second statement of the lemma. □

Proof of Lemma 6.

$$\left|\frac{A+\sigma_1\cdot n_1}{h\cdot B+\sigma_1\cdot n_2}-\frac{A}{B}\right|\approx\left|\frac{A}{B}\left(1+\frac{\sigma_1\cdot n_1}{A}\right)\left(1-\frac{\sigma_1\cdot n_2}{h\cdot B}\right)-\frac{A}{B}\right|$$

$$\approx\frac{A}{B}\left(\frac{\sigma_1\cdot n_1}{A}-\frac{\sigma_1\cdot n_2}{h\cdot B}\right)\leq\frac{2\sigma_1\cdot|n_3|}{B},$$

for some $n_3\sim N(0,1)$. □

References

1. Guilleminault, C., Tilkian, A., Dement, W.C.: The sleep apnea syndromes. Annu. Rev. Med. **27**, 465–484 (1976)
2. Young, T., Peppard, P.E., Gottlieb, D.J.: Epidemiology of obstructive sleep apnea: a population health perspective. Am. J. Respir. Crit. Care Med. **165**(9), 1217–1239 (2002)
3. Sullivan, C.E., Issa, F.G., Berthon-Jones, M., Eves, L.: Reversal of obstructive sleep apnoea by continuous positive airway pressure applied through the nares. Lancet **18**(1), 8225–8625 (1981)
4. Weigelt, L., Westbrook, P., Doshi, R.: Dissatisfaction with OSA management among CPAP rejecters and the role of the primary care physician. Sleep **33**, A159 (2010)
5. Yong J.W.: Head and Face Anthropometry of Adult U.S. Citizens, AD-A268 661 (1993). http://www.faa.gov/data_research/research/med_humanfacs/oamtechreports/1990s/media/am93-10.pdf
6. Zhuang Z., Bradtmiller B., Friess M.: A head and face anthropometry survey of U.S. respirator users, NIOSH NPPTL Anthrotech report, 28 May 2004. http://www.nap.edu/html/11815/Anthrotech_report.pdf
7. Farkas L.G.: Accuracy of anthropometric measurements: past, present, and future. Cleft Palate Craniofac J. **33**(1) 10–8; discussion 19–22 (1996)
8. Salvi, J., et al.: Pattern codification strategies in structured light systems. Pattern Recognit. **37**, 827–849 (2004)
9. Sato K.: Range imaging based on moving pattern light and spatio-temporal matched filter. In: Proceedings of the International Conference on Image Processing, 1996, vol. 1, pp. 33–36, 16–19 September 1996
10. Znamenskiy D.N., Vlutters R., van Bree K.C.: 3D Scanner using structured lighting, patent application US 20150204663, 23 July 2015

Mathematical Modeling for Break Down of Dynamical Equilibrium in Bone Metabolism

Takashi Suzuki, Keiko Itano, Rong Zou, Ryo Iwamoto and Eisuke Mekada

Abstract We study the break down of bone metabolism, using mathematical modeling. The principal part of this model is composed of two pathways of maturation, that is, from pre-osteoblast to osteoblast and from pre-osteoclast to osteoclast. There is also a pathway of acceleration to the formation of pre-osteoclast by pre-osteoblast. This pathway is evoked by a cytokine, called RANKL. Experimental data, on the other hand, suggest a differentiation annihilation factor to the maturation pathways above. Total mathematical modeling on these positive and negative feedback loops induces an insight, how the dynamical equilibrium of this metabolism breaks down, via mathematical analysis and numerical simulations. Then in vivo experiments are proposed to confirm actual existence of the above factor, together with the evaluation of medical manipulations.

Keywords Bone metabolism · Dynamical equilibrium · Mathematical oncology

T. Suzuki (✉) · K. Itano
Division of Mathematical Science, Department of Systems Innovation, Graduate School of Engineering Science, Osaka University, Machikaneyamacho 1-3, Toyonakashi 560-8531, Japan
e-mail: suzuki@sigmath.es.osaka-u.ac.jp

K. Itano
e-mail: itano@sigmath.es.osaka-u.ac.jp

R. Zou
Institute of Mathematics for Industry, Kyushu University, Motooka 744, Nishiku, Fukuokashi 819-0395, Japan
e-mail: zou@math.kyushu-u.ac.jp

R. Iwamoto · E. Mekada
Research Institute for Microbaial Diseases, Osaka University, Yamadagaoka 3-1, Suitashi 565-0871, Japan
e-mail: riwamoto@biken.osaka-u.ac.jp

E. Mekada
e-mail: emekada@biken.osaka-u.ac.jp

B. Anderssen et al. (eds.), *The Role and Importance of Mathematics in Innovation*, Mathematics for Industry 25,
DOI 10.1007/978-981-10-0962-4_3

1 Introduction

Biological phenomena are maintained through metabolism. Usually, this process is under the dynamical equilibrium. For example, bone metabolism is achieved under the balance of two kinds of cells, osteoblast and osteoclast, associated with bone formation and bone resorption, respectively. Break down of this balance, therefore, makes the individual unstable, and sometimes causes diseases, that is, osteopetrosis and osteoporosis if osteoblast dominates osteoclast and if osteoclast dominates osteoblast, respectively.

Maturation pathways of osteoblast and osteoclast are now recognized as follows [2, 6]. First, osteoblast and osteoclast are formed by differentiations of pre-osteoblast and pre-osteoclast, respectively. Here, hematopoiesis stem cell maturates to pre-cell of pre-osteoclast. Then there occurs proliferation of these pre-cells. There is also an acceleration in the differentiation of the above pre-cell to pre-osteoclast, from the pre-osteoblast through a cytokine, called RANKL. In the process of differentiation, finally, pre-osteoclasts form a cluster, called an MN osteoclast, and this cluster matures to a PN osteoclast. Recent experimental data, however, strongly suggest the production of differentiation annihilation factor (DAF) by MN osteoclast. This DAF annihilates both maturations, from pre-osteoblast to osteoblast and from MN osteoclast to PN osteoclast as illustrated in Fig. 1.

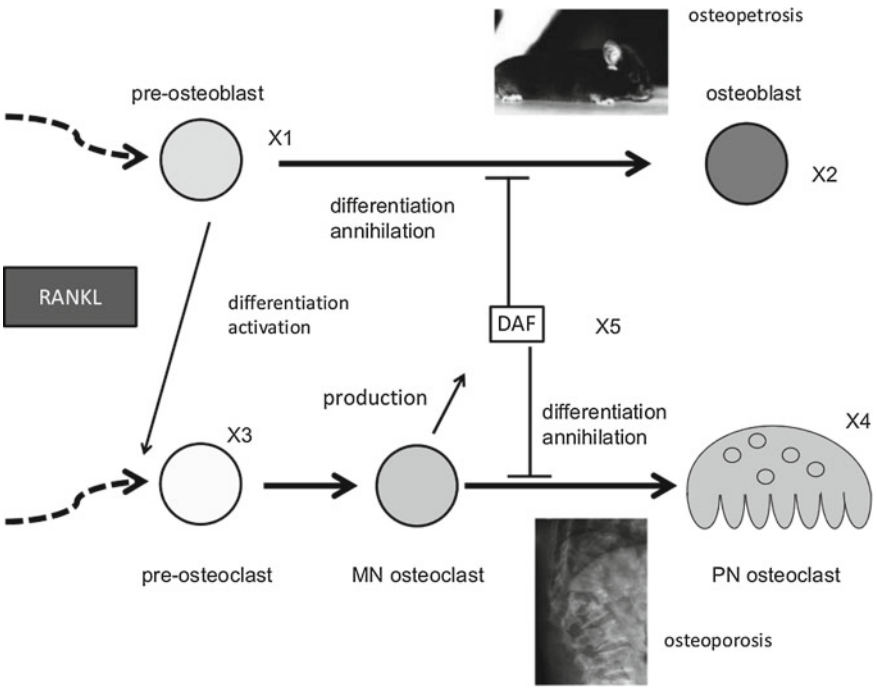


Fig. 1 Two pathways of maturation and three feedback loops

Mathematical models have been used to describe many biological phenomena [4, 5]. Here we examine the above hypothesis of DAF using mathematical modeling. We apply two methods for this purpose, that is, multi-scale modeling and break down of dynamical equilibrium, to predict what should be observed in experimental data and also to evaluate the drug effect.

2 Mathematical Analysis

Here we formulate the above feedback loops as a system of ordinary differential equations, pick up dynamical equilibria, and study their break down. First, we apply multi-scale modeling. The event is on the tissue level, where each cell is regarded as a point, and the densities of four kinds of cells are counted, that is, pre-osteoblast, osteoblast, pre-osteoclast, and osteoclast. Here we identify pre-osteoclast and MN osteoclast, while DAF is on the molecular level. We assume three functions of DAF, that is, production by MN osteoclast, decay by itself, and annihilation of two pathway of differentiation, from pre-osteoblast to osteoblast and pre-osteoclast to osteoclast. These effects on the molecular level are modeled as functional relations. Then dynamical equilibrium is formulated, and dependence on the parameters is examined in connection with its break down. Finally, transit to osteoporosis is suggested by mathematical analysis and numerical simulations at the occasion of break down of dynamical equilibrium.

2.1 Multi-scale Modeling

Densities of the four kinds of cells are defined on tissue level. Hence, pre-osteoblast, osteoblast, pre-osteoclast, and osteoclast are denoted by X_1, X_2, X_3, and X_4, respectively. Then it holds that

$$\frac{dX_1}{dt} = -\ell_1 X_1 + m_1 \tag{1}$$

$$\frac{dX_2}{dt} = \ell_1 X_1 \tag{2}$$

$$\frac{dX_3}{dt} = -\ell_2 X_3 + m_2 \tag{3}$$

$$\frac{dX_4}{dt} = \ell_2 X_3 \tag{4}$$

where m_1 and m_2 denote the amounts of supply per unit time of pre-osteoblast and pre-osteoclast, respectively, and ℓ_1 and ℓ_2 denote the rates of differentiations, from pre-osteoblast to osteoblast and from pre-osteoclast to osteoclast, respectively. These differentiations are annihilated by a factor, which we call DAF. It lies on the

molecular level, produced by MN osteoclast, identified with pre-osteoclast. Hence DAF density, denoted by X_5, is subject to

$$\frac{dX_5}{dt} = \gamma X_3 - \delta X_5 \tag{5}$$

where γ and δ denote the rates of production and self-inhibition, respectively, which are assumed to be positive constants. We call (1)–(5) the top-down model totally.

Positive and negative feedback loops, on the other hand, arise in the molecular level. Below, a, b, c, d, e, f, g, and h denote positive constants. First, pre-osteoblast accelerates the production of pre-osteoclast through the activation of RANKL, which is identified with the pre-osteoblast in this model. Thus we take

$$m_2 = m_2(X_1) = aX_1 + b. \tag{6}$$

Since DAF annihilates the maturations of osteoblast and osteoclast, we assume

$$\ell_1 = \ell_1(X_5) = \frac{c}{dX_5 + e} \tag{7}$$

$$\ell_2 = \ell_2(X_5) = \frac{f}{gX_5 + h}. \tag{8}$$

We call (6)–(8) the bottom-up model totally, under the agreement that m_1 is a positive constant. The precise forms of the bottom-up model, however, are not essential. For the moment it is sufficient to assume the strict convexity of the continuous mapping $x \in [0, \infty) \mapsto \varphi(x) \in (0, \infty)$, where

$$\varphi(x) = m_2\left(\frac{m_1}{\ell_1(x)}\right) \cdot \frac{1}{\ell_2(x)}. \tag{9}$$

In fact we have

$$\varphi(x) = \frac{1}{f}\left(\frac{am_1}{c}(dx + e) + b\right)(gx + h)$$

in the case of (6)–(8).

2.2 Dynamical Equilibrium and Break Down

From the above description, dynamical equilibrium in bone metabolism is formulated by

$$\frac{dX_1}{dt} = \frac{dX_3}{dt} = \frac{dX_5}{dt} = 0$$

which is equivalent to

$$\ell_1 X_1 = m_1, \quad \ell_2 X_3 = m_2, \quad \gamma X_3 = \delta X_5. \tag{10}$$

System of equations (10) is reduced to

$$\frac{\delta}{\gamma} X_5 = \varphi(X_5) \tag{11}$$

and then the other variables are determined by

$$X_1 = \frac{m_1}{\ell_1(X_5)}, \quad X_3 = \varphi(X_5). \tag{12}$$

From the strict convexity of $y = \varphi(x) > 0$, $x \geq 0$, there is a critical value $\overline{\lambda} > 0$ of $\lambda = \delta/\gamma$ concerning the number of solutions to (11). This number is acutally 2, 1, and 0, according to $\lambda > \overline{\lambda}$, $\lambda = \overline{\lambda}$, and $0 < \lambda < \overline{\lambda}$, respectively. Assume $\lambda > \overline{\lambda}$, let $X_5^+ = X_5^+(\lambda) > X_5^- = X_5^-(\lambda) > 0$ be the solutions to (11), and put

$$X_1^\pm = \frac{m_1}{\ell_1(X_5^\pm)}, \quad X_3^\pm = \varphi(X_5^\pm).$$

Then we obtain linearly nondegenerate equilibria of the system (1), (3), and (5), that is,

$$(X_1^\pm, X_3^\pm, X_5^\pm) = (X_1^\pm(\lambda), X_3^\pm(\lambda), X_5^\pm(\lambda)).$$

Dynamics of this system around $(X_1^\pm, X_3^\pm, X_5^\pm)$, on the other hand, is reduced to that of

$$\frac{dX_5}{dt} = \gamma X_3 - \delta X_5 \approx \gamma\varphi(X_5) - \delta X_5 \tag{13}$$

around $X_5 = X_5^\pm$.

By the strict convexity of $y = \varphi(x) > 0$, $x \geq 0$, therefore, the only stable dynamical equilibrium arises when $\lambda > \overline{\lambda}$, that is,

$$(X_1, X_3, X_5) = (X_1^-(\lambda), X_3^-(\lambda), X_5^-(\lambda)).$$

Then the other variables $(X_2, X_4) = (X_2(t), X_4(t))$ exhibit linear growth for t large.

This dynamical equilibrium $(X_1^-, X_3^-, X_5^-) = (X_1^-(\lambda), X_3^-(\lambda), X_5^-(\lambda))$ breaks down as λ decreases below $\overline{\lambda}$. At this occasion it arises the increase of the value $X_5^-(\lambda)$. Although $\lim_{\lambda \downarrow \overline{\lambda}} X_5^-(\lambda)$ exists, its increasing rate becomes extremely high if the malignancy proceeds on time.

2.3 Near from Dynamical Equilibrium

A natural question is what happens if the dynamical equilibrium breaks down. To approach this problem we introduce the notion of near from dynamical equilibrium. Here we assume that m_1, γ, and δ are positive constants, $m_2 = m_2(X_1), \ell_1 = \ell_1(X_5)$, and $\ell_2 = \ell_2(X_5)$, with the strict convexity of $y = \varphi(x) > 0$, $x \geq 0$, defined by (9). Then we say that the solution $(X_1, X_2, X_3, X_4, X_5)$ to (1)–(5) lies on near from dynamical equilibrium if it is in the region where the approximation

$$X_1 \approx \frac{m_1}{\ell_1(X_5)}, \quad X_3 \approx \varphi(X_5) \tag{14}$$

is valid, recalling (12).

This is the region where the dynamics of (X_1, X_2, X_3, X_4) is controlled by that of X_5. In particular, it holds that

$$\frac{dX_4}{dX_2} = \frac{\ell_2 X_3}{\ell_1 X_1} \approx \frac{\ell_2(X_5)}{m_1} \cdot \varphi(X_5) = \frac{1}{m_1} \cdot m_2 \left(\frac{m_1}{\ell_1(X_5)} \right)$$

and hence

$$\frac{d}{dt}\left(\frac{dX_4}{dX_2}\right) \approx \psi'(X_5)\frac{dX_5}{dt}, \quad \psi(x) = \frac{1}{m_1} \cdot m_2 \left(\frac{m_1}{\ell_1(x)} \right). \tag{15}$$

From the positive and negative feedback loops underlying this model, it holds that $\psi'(x) > 0, x \geq 0$. We can actually confirm this property for the case of (6)–(8). Relation (15) means that the break down of dynamical equilibrium, near from dynamical equilibrium, arises in accordance with the velocity of DAF density. More precisely, this break down leads to osteoporosis and osteopetrosis in the cases of $\frac{dX_5}{dt} > 0$ and $\frac{dX_5}{dt} < 0$, respectively.

Even in the case of $\lambda > \overline{\lambda}$, the orbit can stay around the unstable dynamical equilibrium $(X_1^+(\lambda), X_3^+(\lambda), X_5^+(\lambda))$ a relatively long time [1, 3]. Since this is the region near from dynamical equilibrium, such transient experience may cause serious damage to the individual, although the variables

$$(X_1, X_3, X_5) = (X_1(t), X_3(t), X_5(t))$$

eventually approach the stable dynamical equilibrium $(X_1^-(\lambda), X_3^-(\lambda), X_5^-(\lambda))$ as $t \uparrow +\infty$.

A numerical example is shown in Fig. 2. It deals with the bottom-up model (6)–(8) with $a = 1, b = 1, c = 1, d = 1, e = 3, f = 10, g = 1$, and $h = 1$. Here we take $\lambda = \frac{\delta}{\gamma} = 10$ to obtain two solutions to (11), denoted by $X_5^+(\lambda) = 4$ and $X_5^-(\lambda) = 1$. The unstable dynamical equilibrium is detected as $(X_1^+(\lambda), X_3^+(\lambda), X_5^+(\lambda)) = (7, 4, 4)$. We can see that the Morse index of this unstable equilibrium is 1, and hence it is

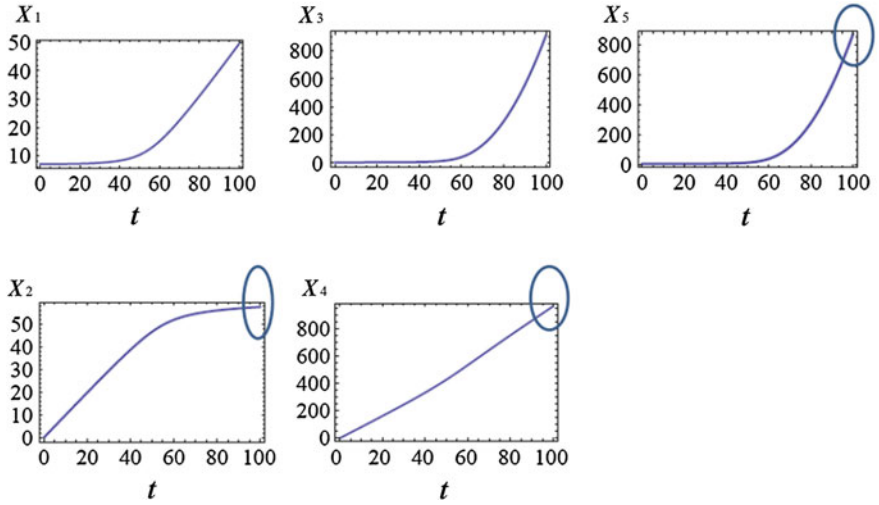


Fig. 2 Dynamics near unstable dynamical equilibrium

associated with a stable manifold of codimension 1. Consequently, generic orbit stays near this unstable equilibrium in a relatively long time.

In this example we take $X_1(0) = 7(1 + 0.01)$, $X_3(0) = 4(1 + 0.01)$, $X_5(0) = 4(1 + 0.01)$, $X_2(0) = 0$, $X_4(0) = 0$ as an initial value near from this unstable dynamical equilibrium. Numerical simulation shows that the orbit stays still around there at $t = 100$. Then $\frac{dX_5}{dt} > 0$ is kept, and consequently, we observe saturation of X_2 after $t \geq 60$, recalling (15). Here, increase of X_5 is due to that of the supply of X_3, which matures to X_4, and therefore, X_2 relatively saturates in spite of the two pathways of annihilation by X_5.

3 Biological Discussion

Having collected experimental data suggesting the existence of DAF, we are suspecting a protein to cast this factor. For in vivo experiments, here we examine more on this hypothesis from the theoretical point of view.

3.1 Near from Dynamical Equilibrium, Revisited

The orbit stays near from dynamical equilibrium at least initially if the initial value is taken there. Then, as we have mentioned, even transient saturation of either X_2 or X_4 takes an important clinical role.

At the break down of dynamical equilibrium the value X_5 takes $X_5^+(\overline{\lambda}) = X_5^-(\overline{\lambda})$, denoted by X_5^*. Since it holds that $\gamma\varphi(X_5^*) - \delta X_5^* > 0$ for $\lambda = \frac{\delta}{\gamma} < \overline{\lambda}$, we have $\frac{dX_5}{dt} > 0$ near this point, recalling (13). Hence break down of dynamical equilibrium occurs with the saturation of osteoblast by (15), which will be a driving force to osteoporosis.

Another near from dynamical equilibrium is achieved around the unstable dynamical equilibrium $(X_1^+(\lambda), X_3^+(\lambda), X_5^+(\lambda))$ for $\lambda > \overline{\lambda}$. Then the conditions $\frac{dX_5}{dt} > 0$ and $\frac{dX_5}{dt} < 0$ arise initially if $X_5(0) > X_5^+(\lambda)$ and $X_5(0) < X_5^+(\lambda)$, respectively. Such initial dynamics that $X_5(0)$ is close to $X_5^+(\lambda)$ may occur often if $0 < \lambda - \overline{\lambda} \ll 1$. In other words, if bone metabolism is close to the break down of dynamical equilibrium and the concentration of DAF becomes higher in some reason, saturation is induced to the production of either osteoblast or osteoclast, associated with increase and decrease of DAF, respectively.

3.2 DAF Knock Down

To identify the cell molecule casting DAF, knockdown technique may be used. Here we specify what should be observed under this operation, modifying the bottom-up model (6)–(8). We take three modifications, cutting the effects of annihilation, that is, annihilation of differentiation of pre-osteoblast to osteoblast, that of pre-osteoclast to osteoclast, and both. We thus obtain three bottom-up models, that is,

$$m_2 = aX_1 + b, \quad \ell_1 = c, \quad \ell_2 = \frac{f}{gX_5 + h}, \tag{16}$$

$$m_2 = aX_1 + b, \quad \ell_1 = \frac{c}{dX_5 + e}, \quad \ell_2 = f, \tag{17}$$

and

$$m_2 = aX_1 + b, \quad \ell_1 = c, \quad \ell_2 = f. \tag{18}$$

Their dynamical equilibria are reduced to

$$\lambda X_5 = \frac{1}{f}\left(\frac{am_1}{c} + b\right)(gX_5 + h), \tag{19}$$

$$\lambda X_5 = \frac{1}{f}\left(\frac{am_1}{c}(dX_5 + e) + b\right), \tag{20}$$

$$\lambda X_5 = \frac{1}{f}\left(\frac{am_1}{c} + b\right), \tag{21}$$

with $\lambda = \frac{\delta}{\gamma}$.

In the first and second cases of (19) and (20), there is $\overline{\lambda} > 0$ such that the unique solution $X_5 = X_5(\lambda)$ exists for $\lambda > \overline{\lambda}$, while there is no solution for $0 < \lambda \leq \overline{\lambda}$. This unique solution is stable and the break down of dynamical equilibrium arises with $\lim_{\lambda \downarrow \overline{\lambda}} X_5(\lambda) = +\infty$. This property has a strong contrast with that of the original model, (6)–(8). Then it is easy to suspect that saturation of X_2 and X_4 arises at this occasion in models (16) and (17), respectively. In the third case, finally, there is a unique stable dynamical equilibrium for any $\lambda > 0$.

Break down of dynamical equilibrium may be caused several ways. Increase of a or b in (6) may be achieved by injecting RANKL. This technique is standard in cell biology. Since (11) takes the form

$$\frac{\delta f}{\gamma} X_5 = \left(\frac{a m_1}{c} (d X_5 + e) + b \right) (g X_5 + h) \tag{22}$$

we can create a break down of dynamical equilibrium by making a or b large. Since X_2 saturates at this occasion, osteoporosis will be observed for wild type, which will not appear to the control, knockdown mice of DAF. In fact, the function $\psi(x)$ defined by (15) is constant and any solution will approach the unique dynamical equilibrium in the case of (21).

3.3 Medical Insights

We can present several medical insights based on the argument above. First, break down of dynamical equilibrium may be predicted by the increase of DAF. Second, the recovery of dynamical equilibrium plays an essential role against both diseases, osteopetrosis and osteoporosis. Since (22), good manipulations are increase of δ, f, c and decrease of γ, a, m_1, d, e, b, g, h. Generally, inhibition of DAF or that of differentiation to pre-osteoblast and pre-osteoclast are efficient to recover dynamical equilibrium. Third, rapid increase of DAF density can be a trigger of osteopetrosis or osteoporosis.

4 Conclusion

We have studied the role of DAF, supposed to be a key molecule in bone metabolism, using mathematical modeling. Two factors are adopted, multi-scale modeling and dynamical equilibrium. The notion of near from dynamical equilibrium is introduced, which is realized around the unstable dynamical equilibria and also at the moment of break down of dynamical equilibrium. There, saturation of one of the antagonists, osteoblast or osteoclast, arises, in accordance with the variation of the key molecule. Several suggestions, predictions, and evaluations are obtained theoretically to in

vivo experiments and medical manipulations. Consequently, mathematical methods applicable to other problems in cell biology are established.

Acknowledgments This work is supported in part by JSPS Grant-in-Aid Scientific Research (A) 26247013. (B) 15KT0016, and JSPS Core-to-Core Program, Advanced Research Networks.

References

1. Arnold, V.I.: Mathematical Methods of Classical Mechanics, 2nd edn. Springer, New York (1989)
2. Harada, S., Rodan, G.A.: Control of osteoblast function and regulation of bone mass. Nature **423**(6937), 349–355 (2003)
3. MacKay, R.S.: Stability of equilibria of Hamiltonian systems. In: Sarkar, S. (ed.) Nonlinear Phenomena and Chaos, pp. 254–270. Adam Hilger Ltd., Bristol (1986)
4. Murray, J.D.: Mathematical Biology, I: An Introduction, 3rd edn. Springer, New York (2003)
5. Murray, J.D.: Mathematical Biology, II: Spetial Models and Biomedical Applications, 3rd edn. Springer, New York (2003)
6. Teitelbaum, S.L., Ross, F.P.: Genetic regulation of osteoclast development and function. Nat. Rev. Genet. **4**(8), 638–649 (2003)

Why Do Hives Die? Using Mathematics to Solve the Problem of Honey Bee Colony Collapse

Mary R. Myerscough, David S. Khoury, Sean Ronzani and Andrew B. Barron

Abstract Honey bees are vital to the production of many foods which need to be pollinated by insects. Yet in many parts of the world honey bee colonies are in decline. A crucial contributor to hive well-being is the health, productivity and longevity of its foragers. When forager numbers are depleted due to stressors in the colony, such as disease or malnutrition, or in the environment, such as pesticides, this causes a reduction in the amount of food (nectar and pollen) that can be collected and a reduction of the colony's capacity to raise brood (eggs, larvae and pupae) to produce new adult bees. We use a set of differential equation models to explore the effect on the hive of high forager death rates. We track the population of brood; hive bees who work inside the hive; foragers who bring back food to the hive; and stored food. Using data from experimental research we devised functions that described the effect of the age that bees first become foragers on their success and lifespan as foragers. In particular, we examine what happens when bees become foragers at a comparatively young age and how this can lead to a sudden rapid decline of adult bees and the death of the colony.

Keywords Colony collapse disorder · Foraging · Pollination · Food security · *Apis mellifera*

M.R. Myerscough (✉) · S. Ronzani
School of Mathematics and Statistics, University of Sydney, Sydney, NSW 2006, Australia
e-mail: mary.myerscough@sydney.edu.au

D.S. Khoury
Kirby Institute, University of New South Wales, Sydney, NSW 2052, Australia
e-mail: dkhoury@kirby.unsw.edu.au

A.B. Barron
Department of Biological Sciences, Macquarie University, Sydney, NSW 2109, Australia
e-mail: andrew.barron@mq.edu.au

B. Anderssen et al. (eds.), *The Role and Importance of Mathematics in Innovation*, Mathematics for Industry 25,
DOI 10.1007/978-981-10-0962-4_4

1 Section Heading

Growing crops for food is arguably the oldest human industry. Food crops need to be pollinated to produce fruit or seed, either for consumption or to sow for the next crop. Cereals such as rice, wheat and maize are pollinated by wind, but almost all other crops require insects to transfer pollen from one flower to another to set the seed. Honey bees (*Apis mellifera*) are the most important pollinators of commercial crops and so the survival and health of honey bee colonies is vital for food security worldwide.

Individual honey bees only survive as part of a colony or hive, usually of several thousand bees. Bees in each colony collectively gather food and raise brood (that is, eggs, larvae and pupae) from eggs laid by a single queen [19]. Labour in the colony is self-organised in response to pheromonal and behavioural cues that are generated by interactions between adult bees, brood and food. In a commercial apiary, if all the bees in a hive die or abandon the colony, this leads to a financial loss to the apiarist who must either establish a new colony or suffer the loss of productive capacity.

In the last decade or more, the rate of honey bee colony failure has increased significantly and the number of commercial colonies has declined, particularly in the USA but also in Europe and Japan [11]. Colony failure includes overwintering losses in cold climates and loss due to diseases, pesticide exposure or the *Varroa* mite [11, 18]. One of the more puzzling types of colony loss is where previous healthy hives are found abandoned by adult bees, still containing stored food and dead brood, but with few or no adult bee, and few dead bees [16]. There may be no obvious cause for this very rapid depopulation of the hive. This phenomenon has become known as colony collapse disorder (CCD). It usually occurs in late spring or early summer, just as hives have emerged from their winter hibernation and are reaching their peak summer numbers.

There is now general agreement among bee biologists that there are many causes of CCD and these can include anything which puts a hive under stress, including diseases, parasites, pesticide use in the hive environment, apicultural practices such as keeping large number of hives close together or poor nutrition, due, for example to the hive foraging from a single type of plant including agricultural monocultures such as almond orchards or canola crops [2, 5]. All of these put stress on the hive but the presence of these stressors alone does not explain the mechanism of the sudden depopulation that hives experience in CCD.

In this paper, we use mathematical models for the population of adult bees, brood and food in the hive to explore the hypothesis that the mechanism that drives CCD is sustained high death rates of foragers which causes adult bee numbers to become depleted and leads to hive death.

Our hypothesis that high forager death rates drive hive collapse depends crucially on the population dynamics of the hive. All eggs in the hive are laid by a single queen bee. After 3 days a egg hatches into a larvae which is fed and cared for by worker bees in the hive. The larvae pupates after 9 days and 12 days later the adult bee emerges from pupation. Young adult bees remain in the hive, caring for brood

and for the queen, storing food that is brought back by foragers and doing other in-hive work. Older bees become foragers and leave the hive to gather nectar and pollen to supply the colony with food. Foragers are exposed to many hazards in the environment outside the hive and, as foraging itself is a metabolically expensive and risky activity, forager lifespan is generally less than 7 days from the time a bee starts to forage [17]. If a forager is diseased or malnourished, then her lifespan is likely to be shorter than that of a healthy forager in the same environment.

The transition from hive bee to forager is controlled by social feedback. If there are many foragers, then hive bees tend not to become foragers; if there are few foragers, however, then older hive bees will become foragers. This social inhibition is mediated by the pheromone ethyl oleate, which is produced by foragers [9]. Food shortages also stimulate hive bees to become foragers [14]. This, potentially enables a hive to survive a temporary shortage of food when food storage returns to normal levels within a few days.

It is well known that large hives raise a higher proportion of eggs to adulthood [1]. Larger populations of hive bees and foragers can give more care and find more food for larvae and other brood. When food is short, however, hive bees cannibalise some larvae and eggs to feed older larvae. Hence food shortages also reduce the number of brood raised to adults.

We will construct two differential equation models that represent the interactions between hive bees, foragers, food and brood. The first model can easily be analysed to show the bifurcation behaviour of the system and, in particular, what happens to the steady-state solutions as forager death rates increase. The second model takes into account the effect of the age that bees first become foragers on forager recruitment, survival and efficacy. This second model cannot be easily analysed, except numerically, but produces the rapid collapse of hive populations that is observed by apiarists.

2 A Basic Model for the Dynamics of Food, Brood, and Hive Bee and Forager Populations

We use a system of four differential equations to model the interactions illustrated in Fig. 1. In the model, we will only consider uncapped brood, that is, eggs and larvae. When a larva pupates, the hive bees cap the cell in the brood comb which the pupa occupies with a wax cap. When the adult bee emerges from pupation she chews through the cap. The pupae or capped brood are not represented explicitly in the model, but their presence is modelled by a delay between uncapped brood going into pupation and emerging as adults 12 days later.

The independent variable in the model is time t, measured in days. The dependent variables are f representing stored food within the hive, B, the number of uncapped brood items, H, the number of hive bees and F, the number of foragers. This model

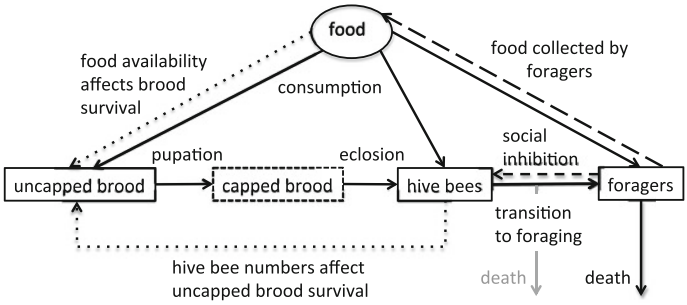


Fig. 1 A flow chart showing processes and interactions in the honey bee colony that are represented in the model. The *grey arrow* and labelling on the right-hand side of the chart describe death of bees during transition to foraging which is included in the extended model with age dependence but not in the basic model

was first presented in [8] and is based on a simpler model that represented foragers and hive bees only [7].

2.1 Model Equations

Food is collected by foragers and consumed by foragers, hive bees and brood. The difference between the rate of food collection and food consumption gives the rate of accumulation or depletion of stored food f:

$$\frac{df}{dt} = cF - (\gamma_B B + \gamma_H H + \gamma_F F). \tag{1}$$

Here c is the weight of food in grams collected per forager per day and γ_B, γ_H and γ_F are the average weight of food consumed per day by each brood item, hive bee or forager, respectively. For ease of analysis, we assume that all adult bees consume the same amount of food on average, so we set $\gamma_H = \gamma_F = \gamma_A$. Hence

$$\frac{df}{dt} = cF - \gamma_B B - \gamma_A (H + F). \tag{2}$$

In the model we include only brood that survives to adulthood. These brood items are a proportion of the total number of eggs laid by the queen and we model their rate of pupation as a linear rate proportional to the number of uncapped brood items. Hence

$$\frac{dB}{dt} = L\,S(H, f) - \phi B, \tag{3}$$

where L is the laying rate of the queen in number of eggs per day, ϕ is the rate of pupation in brood numbers per day where $1/\phi$ is the time that a brood item spends as uncapped brood, which is known to be 9 days hence $\phi = 1/9$. The function $S(H, f)$ gives the proportion of eggs that survive to become adult bees. This will depend on the number of hive bees and the amount of stored food, particularly when stored food levels are low. We model $S(H, f)$ as follows:

$$S(H, f) = \frac{f^2}{f^2 + b^2} \frac{H}{H + \nu} \tag{4}$$

where the parameters b and ν determine how quickly $S(H, f)$ approaches one as f and H increase, respectively. The function $S(H, f)$ saturates both with respect to food f and also with respect to hive bee numbers H. Clearly, the number of brood that is raised to adulthood cannot continue to increase indefinitely as food stocks and hive bee numbers increase. There is a maximum number of brood that a hive can raise and this maximum is determined by the queen's laying rate L. We assume that the proportion of brood raised increases linearly with the number of hive bees when H is low, but that the dependence on food stores is sigmoidal, due to the need for hive bees to spend more time finding food inside the hive when stored food f is very low.

Hive bees emerge τ days after they become pupae. Bees leave the hive bee class to become foragers. Generally speaking, the hive is a very safe environment with low adult bee mortality, so we do not include death of hive bees in the model. We model the change of hive bee population as

$$\frac{dH}{dt} = \phi B(t - \tau) - R(H, F, f)\, H \tag{5}$$

where the function $R(H, F, f)$ governs the recruitment rate of hive bees to the forager class. This rate is determined both by the proportion of foragers in the colony and by the availability of food stores. We model the recruitment function as

$$R(H, F, f) = \alpha_{\min} + \alpha_{\max} \left(\frac{b^2}{b^2 + f^2} \right) - \sigma \left(\frac{F}{F + H} \right) \tag{6}$$

where $\alpha_{\min}$ is the rate that hive bees become foragers when there are no foragers but plenty of food in the hive, $\alpha_{\max}$ determines the strength of the effect that low food stores have on forager recruitment and σ governs the effect of social inhibition on recruitment. If stored food is plentiful, then recruitment essentially does not depend on food at all, but only on the proportion of foragers to the total number of adult bees in the hive. If this proportion is high, bees are inhibited from becoming foragers and the rate of recruitment is low. Conversely, if there is a very low proportion of foragers among the adult bees of the hive, then inhibition is weak and recruitment is high. The constant $\alpha_{\min}$ is the rate of recruitment when there is plentiful stored food but no foragers. The constant $\alpha_{\max}$ governs the effect of food shortage on recruitment.

Foragers are recruited from the hive bee class and die at a linear rate:

$$\frac{dF}{dt} = R(H, F, f)H - \mu F. \tag{7}$$

Here the first term models recruitment and the second term represents forager death where μ is the forager death rate.

2.2 Results from the Basic Model

Figure 2 shows how the population of a model colony evolves for different forager death rates. When the death rate is low, with $\mu = 0.1$ (so 10 % of foragers are lost each day), the hive has a large population of both adult bees and brood, and food

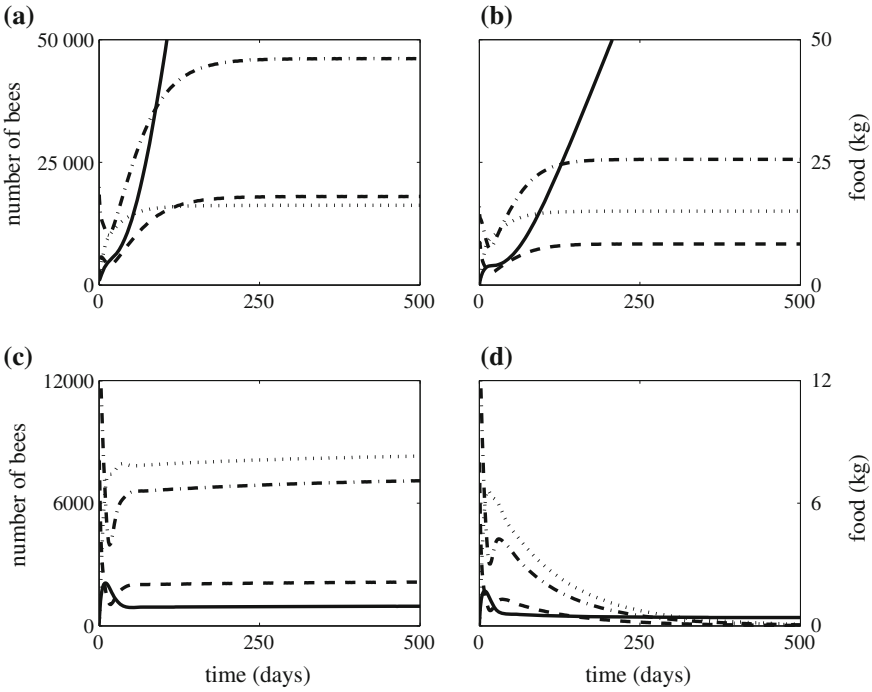


Fig. 2 Results from the basic model. Population and stored food as a function of time for different values of forager death rate μ. Food is represented by the *solid curve*, brood by the *dotted curve*, hive bees by the *dot-dash curve* and foragers by the *dashed curve*. Parameter values are $L = 2000$, $\phi = 1/9$, $\nu = 5000$, $\sigma = 0.75$, $\alpha_{\max} = 0.25$, $\alpha_{\min} = 0.25$, $b = 500$, $c = 0.1$, $\gamma_A = 0.007$, $\gamma_B = 0.018$ and $\tau = 12$. See [8] for a justification of these values. Initially all simulations had 20,000 hive bees, 10,000 foragers and no brood or food. For **a** $\mu = 0.1$, **b** $\mu = 0.2$, **c** $\mu = 0.43$ and **d** $\mu = 0.55$. Note that the vertical scale on **c** and **d** is different to that on **a** and **b**

stores increase without bound. Such a hive is in an ideal state for honey production as well as providing many pollinators to the surrounding area.

A higher death rate, $\mu = 0.2$, produces a hive with a smaller number of adult bees, although the brood population is not affected very much which suggests that the hive raises a similar number of adult bees, but that adult bees have a shorter lifespan. Food continues to increase but not as rapidly as when $\mu = 0.1$.

When death rates become highly elevated, then food stores in the model colony do not grow. When $\mu = 0.43$ the population of adult bees is very low compared to when $\mu = 0.1$ and, in fact, there are now fewer hive bees than uncapped brood. Nevertheless, the hive remains viable.

When μ is raised further, however, the hive collapses. The model shows an exponential decline in both adult bee and brood numbers and within 250 days, its population has dropped below 1000 adults, which is approximately the lowest number of bees a hive needs to survive. At the same time, about 500 g of stored food remains in the hive, even after the bees are nearly all dead.

Equations 2–7 can be solved analytically at steady state to obtain a bifurcation diagram with steady-state solutions as a function of forager death rate μ as shown in Fig. 3. When food grows unboundedly we ignore Eq. 2 and assume $f \to \infty$ in 3–7. The weight of residual stored food was calculated numerically when bee numbers are zero at steady state.

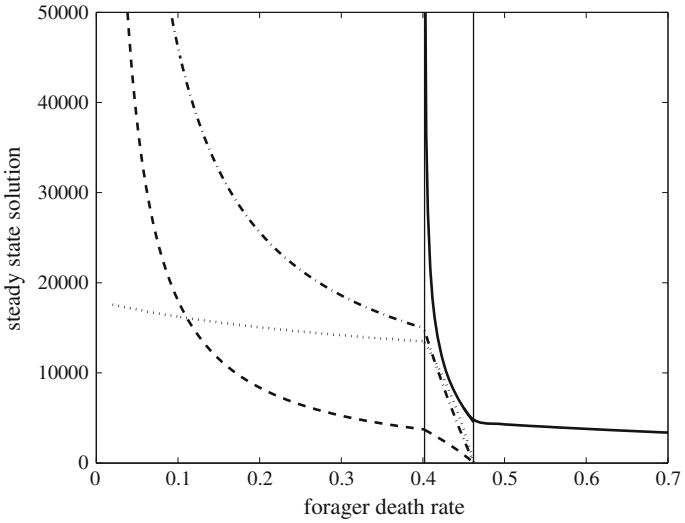


Fig. 3 Steady-state solutions as a function of forager death rate μ. Food is represented by the *solid curve*, brood by the *dotted curve*, hive bees by the *dot-dash curve* and foragers by the *dashed curve*. Parameter values are the same as in Fig. 2. The *vertical lines* at $\mu = 0.402$ and $\mu = 0.462$ separate the plot into regions with qualitatively different solution behaviour

When μ is low, that is $\mu < 0.402$ for this parameter set, food grows unboundedly. As μ increases within this range, the number of adult bees drops significantly while brood numbers are less sensitive to forager death rate.

For $0.402 < \mu < 0.462$, stored food does not grow unboundedly but has a finite steady state. At $\mu \approx 0.402$ the solution curves for adult bee and brood populations have a slope discontinuity in the bifurcation diagram which suggests that the processes that govern the steady-state values have changed. In particular, it suggests that stored food determines the bee populations when f remains finite. In this regime, increasing μ affects both adult bee and brood populations to the same degree, unlike in the case where f is unbounded where adult bee populations are much more sensitive to changes in μ than brood populations.

When $\mu > 0.462$ then bee populations go to zero as $t \to \infty$. However, by numerically solving Eqs. 2–7 it is evident that stored food is not zero even when bee populations are arbitrarily close to zero. This can be seen in Fig. 3. This result is in agreement with observations of hives that experience CCD as apiarists report that collapsed hives do have food stores remaining, so that starvation is clearly not the sole cause of CCD.

The results from the basic model demonstrate that high forager death rates, on their own can lead to the death of a colony. However, this model predicts that the decline of the colony will be slow, rather than the rapid decline observed in real colonies that experience CCD. This suggests that there is an important aspect of collapsing colony behaviour that is missing from the basic model.

3 Model Incorporating the Effects of Forager Age

Foraging is a task that is physically and cognitively demanding for the individual bee. Her wing muscles must be strong and well developed to carry her perhaps kilometres from the hive; she must be able to navigate accurately and successfully through the environment outside the hive; and she must be able to locate and recognise suitable sources of nectar and pollen and exploit them efficiently. When adult bees emerge from pupation, their brains and wing muscles are not yet mature and the time that each bee spends as a hive bee allows her to mature sufficiently to become a successful forager. It is well known that bees that become foragers when they are too young die sooner and do not forager as effectively as bees that become foragers at a later age [15, 20].

In the basic model, the length of time that a hive bee spends in the hive before she becomes a forager or, alternatively, her age at commencement of foraging a is the reciprocal the recruitment function $R(H, F, f)$:

$$a = \frac{1}{R(H, F, f)}. \tag{8}$$

Recent experiments by Perry and coworkers [12] have determined the effect of the age of commencement of foraging a on the number of foraging trips a forager makes each day and forager death rate. These researchers also observed that bees which started to become foragers at a young age, often did not make a successful transition to foraging. As part of the transition to foraging, hive bees undertake several exploratory flights outside the hive, for a total of approximately 30 min. Bees that started to make this transition too young were likely not to survive this exploratory phase and so successful transition to foraging was also age dependent.

We extended the model to include the effects of age of commencement of foraging [12]. In the extended model, forager death rate is given by $\mu = m_r M(a)$ where $M(a)$ is the forager death rate, dependent on a the age that a bee commences foraging and m_r is the ratio of the death rate in a stressed hive to the death rate in a healthy hive. We made the rate of food collection c dependent on a so that $c = c_T N(a)$ where c_T is the weight of food in grams that is collected in each foraging trip and $N(a)$ is the average number of foraging trips made per day by a bee who commences foraging at age a. We also introduced function $T(a)$ which gave the proportion of hive bees that successfully made the transition to foraging, so that bees left the hive bee class at a rate $R(H, F, f)\, H$ but arrived in the foraging class at a rate $T(a)\, R(H, F, f)\, H$.

The extended model consists of four differential equations and one algebraic equation:

$$\frac{df}{dt} = c_T N(a) F - (\gamma_B B + \gamma_H H + \gamma_F F) \tag{9}$$

$$\frac{dB}{dt} = L \frac{f^2}{f^2 + b^2} \frac{H}{H + \nu} - \phi B \tag{10}$$

$$\frac{dH}{dt} = \phi B(t - \tau) - R(H, F, f)\, H \tag{11}$$

$$\frac{dF}{dt} = T(a) R(H, F, f) H - m_r M(a) F \tag{12}$$

$$a = \frac{1}{R(H, F, f)} \tag{13}$$

where $R(H, F, f)$ is given by Eq. 6 as before. Plots of the three functions $N(a)$, $M(a)$ and $T(a)$ are shown in Fig. 4. The functions $N(a)$ and $M(a)$ model data points which are shown in Fig. 4 (although they are not statistically fitted to this data), but $T(a)$ is modelled based on qualitative observations. Note that these equations are essentially the same as Eqs. 2–6, except that they include the functions $N(a)$, $M(a)$ and $T(a)$. Fundamentally, these functions are dependent on the recruitment function $R(H, F, f) = 1/a$ and so the differential equations above could be written as functions of f, B, H and F only. However, the age a that foragers commence foraging is an important biological quantity. Writing these equations in terms of a and including Eq. 13 is not only neater, but also conveys the biological importance of these additional functions, $N(a)$, $M(a)$ and $T(a)$ more clearly.

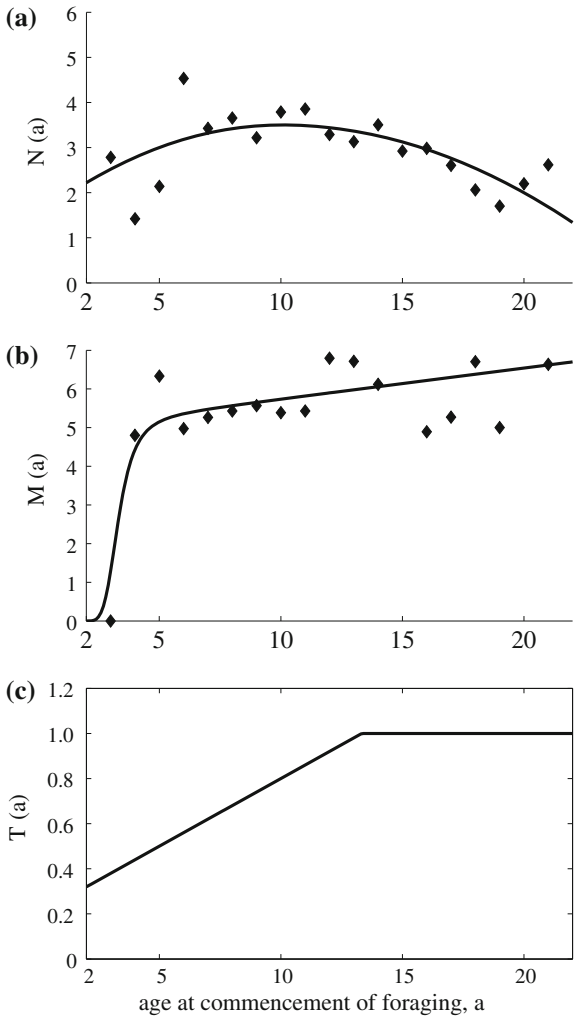


Fig. 4 Plots of the functions dependent on age of commencement of foraging a in the extended model. **a** $N(a)$, the average numbers of trips per day; **b** $M(a)$ the death rate as a function of a; **c** $T(a)$ the proportion of hive bees that successfully make the transition to foragers

3.1 Results of the Extended Model

Figure 5 shows how model populations and stored food change as a function of time for various values of m_r in the extended model.

For sufficiently low values of m_r, food increases unboundedly. The age of commencement of foraging is more than 20 days for a healthy hive and drops as m_r declines. As in the basic model, as death rates increase with increasing m_r the number of adult bees declines, but brood numbers are not significantly affected.

When $m_r = 1.91$, stored food does not increase without bound, although the populations of brood, hive bees and foragers appear to stabilise and there is no evident

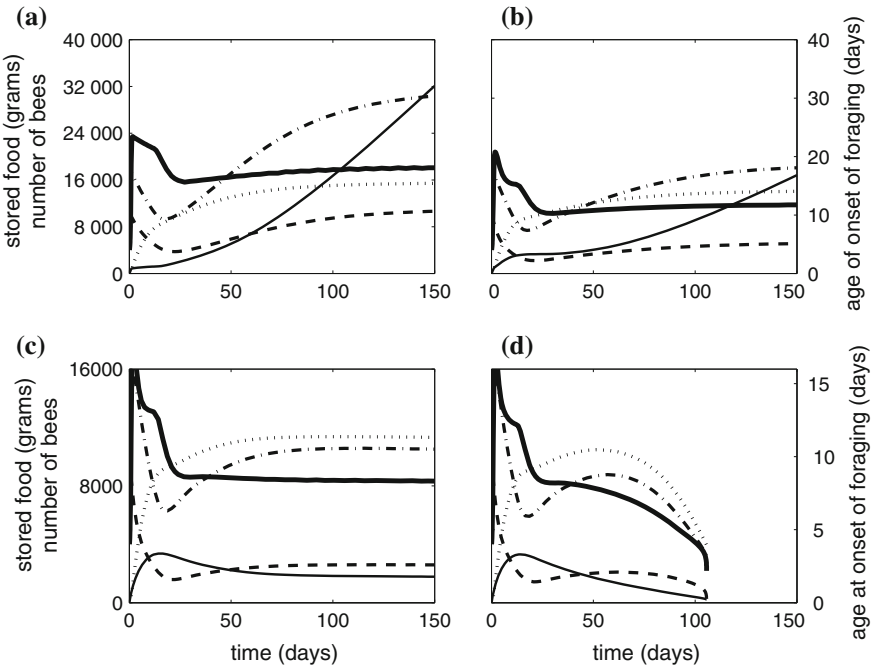


Fig. 5 Solutions of the extended model showing the changes in brood, hive bee and forager populations, stored food and the age of commencement of foraging as a function of time for different values of m_r. Food is represented by the *thin solid curve*, brood by the *dotted curve*, hive bees by the *dot-dash curve*, foragers by the *dashed curve* and age of commencement of foraging is represented by the *thick solid line*. Parameter values and initial conditions are the same as in Fig. 2. The functions $N(a)$, $M(a)$ and $T(a)$ are as shown in Fig. 4. **a** $m_r = 1$; **b** $m_r = 1.6$; **c** $m_r = 1.91$; **d** $m_r = 2.0$. Note that the vertical scale on **c** and **d** is different to that on **a** and **b**

collapse. However, when $m_r = 2$, the populations, which initially look similar to those when $m_r = 1.91$ for $t < 70$ days collapse rapidly in 30 days from about 12,000 adult bees to no foragers and only hive bees that are less than 2 days old and so unable to make the transition to foragers. Stored food declines monotonically after about $t = 20$.

The effect of increasing death rates on the age of adult bees can be seen explicitly in the extended model. As m_r increases from 1 to 1.6, to 1.91 the age at steady state when hive bees become foragers decreases from 23 to 13 to 9 days. Bees will have a shorter lifespan as foragers when m_r is higher, as well as starting to forage at a younger age.

These results suggest that the dependence of forager lifespan and efficacy on the age that a bee first starts to forage is crucial in CCD. As foragers die, social inhibition is reduced and so hive bees become foragers more rapidly and hence at a younger age. Younger foragers have shorter lifespans and bees who become foragers when they are younger than 10 days old, make fewer foraging trips on average as their

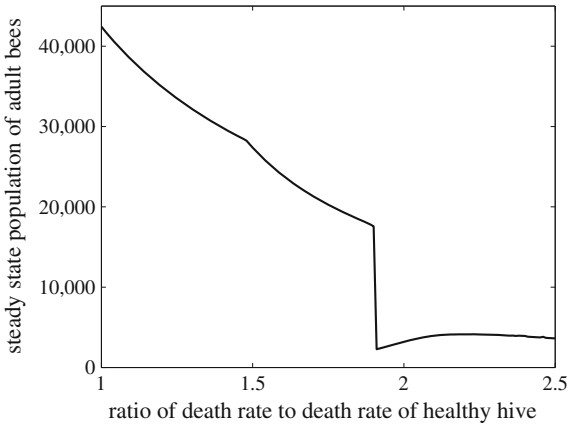


Fig. 6 Plot showing the approximate steady-state population of adult bees as a function of m_r. The steady-state population presented is calculated numerically by solving the extended model Eqs. 10–13 with the same initial conditions and parameter values used in Fig. 2 and taking the population values at $t = 1000$ as an approximation for the steady-state value. The slope discontinuity at $m_r \approx 1.4$ is due to the slope discontinuity in $T(a)$ which is illustrated in Fig. 4

age of commencement of foraging declines. This poor rate of food collection has an effect on stored food and as this decreases, it starts to have an impact on brood raising via $S(H, f)$ (Eq. 4) and also promotes an increased rate of recruitment via $R(H, F, f)$ (Eq. 6). Eventually, stimulated by very low food stores, hive bees are becoming foragers and leaving the hive as soon as possible and dying rapidly either during or after transition to foraging as they do not have the maturity to survive long outside the hive. This leads to a colony with some residual food stores, a low level of brood and comparatively few remaining adult bees which is what is observed in CCD.

Numerical solutions to the extended model show that forager age and adult bee populations at the steady state that corresponds to a viable colony, decline as m_r gets larger (Fig. 6). When m_r reaches a critical value, the solution for a viable colony population no longer exists and the colony will collapse as $t \to \infty$. The numerical solution shown in Fig. 6 suggests that there is a bifurcation where the steady state corresponding to a viable population is lost.

Figure 5 suggests that when this collapse occurs it is driven by a slow decline in food stores which slowly increases the rate of recruitment, and consequent loss, of foragers as the recruitment rate increases as stored food becomes scarce. To further explore the effect of stored food on the steady state, we took f as fixed and analytically calculated the steady-state values of B, F, H and a as a function of m_r for different fixed values of f. This effectively produced a bifurcation diagram with a series of bifurcation curves shown in Fig. 7. When f is very low, the solution curves have a fold bifurcation with two stable steady states, one that corresponds to a viable population

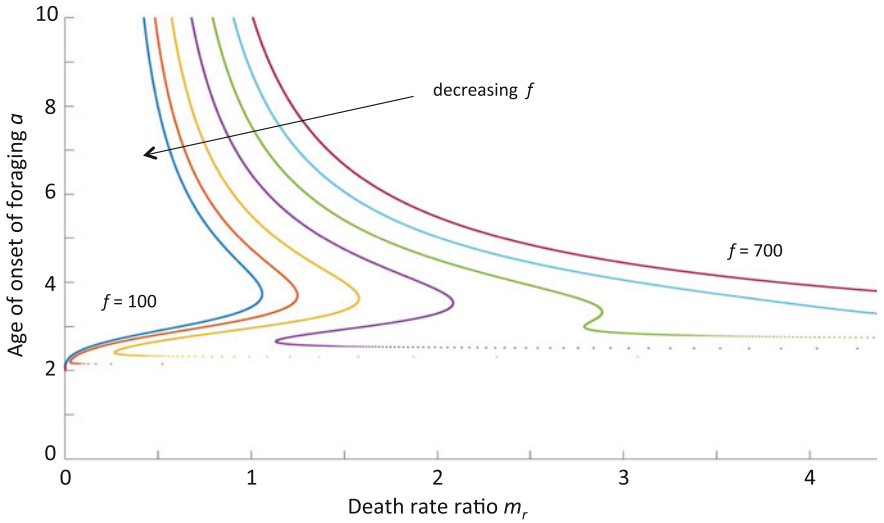


Fig. 7 A bifurcation diagram showing a series of solution curves, each for a different fixed value of f, between $f = 100$ and $f = 700$. The bifurcation parameter is m_r and the age of commencement of foraging is used as the measure of the solution

and one that corresponds to a collapsed colony. Once f is greater than about 550, then there is no fold and the only steady state corresponds to a viable population.

For $m_r = m_r^*$, a fixed value of m_r, both the age of commencement of foraging and the populations of adult bees will decrease at steady state as f decreases. There is a critical value of f, f_{crit} such that when $f = f_{crit}$, the upper bifurcation point on the solution curve is at $m_r = m_r^*$. When $f > f_{crit}$ then the steady state that corresponds to a viable colony exists, but when $f < f_{crit}$ then there is only one steady state and this corresponds to a collapsed colony. Thus if f is declining very slowly then the colony populations will decline slowly until $f = f_{crit}$ when the population will collapse.

This brief, qualitative discussion suggests the possibility of a need for a more comprehensive and rigorous analysis to determine the factors that govern the existence and timing of population collapses in the extended model.

4 Conclusions and Opportunities from Mathematical Models for CCD

The models presented here support the hypothesis that sustained high forager death rates, along with the decreased forager survival and efficacy with decreasing age, leads to CCD. This has yet to be tested in the field, but forager loss is a mechanism that is consistent with our understanding that CCD has many causes, each of which produce stress in the hive [3]. Pesticide use, for example, clearly causes increased

forager mortality directly but also indirectly as sublethal effects reduce the quality of the foraging force [6] and increase susceptibility to disease [13]. Diseases in the hive, also contribute to forager mortality, both directly through death due to disease and because foragers that are diseased will tend to have less energy and may be cognitively compromised and so will make fewer foraging trips and may become lost more easily [10]. Likewise, malnourished foragers are likely to collect less food and have a shorter lifespan than healthy bees. Therefore, although the mechanism of CCD proposed in this model is quite specific, the underlying causes of CCD can be very diverse and still produce the same sustained high forager death rate that we hypothesise leads to CCD [3].

Once we understand the mechanism of CCD then we can explore possible solutions. The model suggests different data from hives that can be monitored to identify colonies that are vulnerable to collapse. One clear result from both models is that adult bee numbers are much more sensitive to forager loss than brood numbers and so apiarists should monitor adult bee numbers rather than brood numbers to determine colony health and vulnerability to CCD.

Models also allow us to explore different ways to prevent vulnerable colonies from collapsing. One very simple potential approach to rescue collapsing colonies is in-hive feeding where food is put directly into the hive. This is easy to incorporate into the models. With in-hive feeding, the equation for f in the extended model becomes

$$\frac{df}{dt} = c_T N(a)F - (\gamma_B B + \gamma_H H + \gamma_F F) + C \tag{14}$$

where C is the rate, in grams/day, that food is put into the hive. Figure 8 shows how feeding the colony even quite small amounts (60 g of food per day) enables it to survive. In an apicultural setting, hives can be fed for a short time, until forager mortality is reduced and the hive is able to maintain a viable population without feeding. Although there are costs associated with in-hive feeding these are likely to be less than the costs of replacing collapsed colonies.

Finally, demographic models for honey bee colonies, such as these, can be used as platforms to develop more complicated models that explicitly incorporate the effect of disease [4]. Many diseases and parasites including the *Varroa* mite affect bees differently at different stages of their life cycle. Demographic models for honey bee colonies, offer the potential to explore disease dynamics and prevention strategies more closely.

Mathematical modelling will never replace field studies of honey bee health and behaviour. However, models are a valuable complement to experimental studies, particularly for exploring hypotheses and predicting the outcomes of complex interactions within honey bee colonies. Models can show what are the likely consequences of colony behaviours and help experimental researchers to focus their efforts on the most important processes in colony dynamics. Together modellers and experimental researchers can ensure that research into honey bee health, which profoundly impacts food production and therefore food security, is both timely and effective.

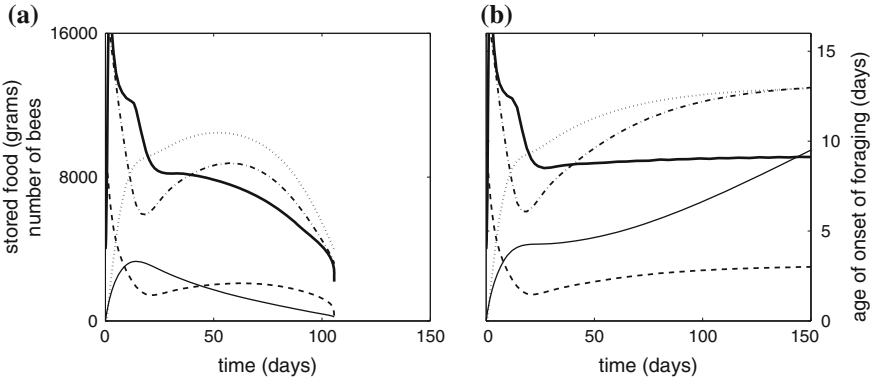


Fig. 8 Plots showing the effect of in-hive feeding. Food is represented by the *thin solid curve*, brood by the *dotted curve*, hive bees by the *dot-dash curve*, foragers by the *dashed curve* and age of commencement of foraging is represented by the *thick solid line*. For both plots $m_r = 2$. Plot **a** is the same simulation as in Fig. 5d showing hive collapse. Plot **b** is exactly the same simulation but with in-hive feeding (Eq. 14), with $C = 60$. All parameter values and initial conditions are the same as Fig. 5

References

1. Allen, M.D., Jeffree, E.P.: The influence of stored pollen and of colony size on the brood rearing of honeybees. Ann. Appl. Biol. **44**, 649–656 (1956)
2. Archer, C.R., Pirk, C.W.W., Wright, G.A., Nicolson, S.W.: Nutrition affects survival in African honeybees exposed to interacting stressors. Funct. Ecol. **28**, 913–923 (2014)
3. Barron, A.B.: Death of the bee hive: understanding the failure of an insect society. Curr. Opin. Insect Sci. **10**, 45–50 (2015)
4. Betti, M.I., Wahl, L.M., Zamir, M.: Effects of infection on honey bee population dynamics: a model. PLOS ONE **9**, e110237 (2014)
5. Cornman, R.S., Tarpy, D.S., Chen, Y., Jeffreys, L., Lopez, D., Pettis, J.S., vanEngelsdorp, D., Evans, J.D.: Pathogen webs in collapsing honey bee colonies. PLOS ONE **7**, e43562 (2012)
6. Desneux, N., Decourtye, A., Delpuech, J.M.: The sublethal effects of pesticides on beneficial arthropods. Ann. Rev. Entomol. **52**, 81–106 (2007)
7. Khoury, D.S., Myerscough, M.R., Barron, A.B.: A quantitative model of honey bee population dynamics. PLOS ONE **6**, e18491 (2011)
8. Khoury, D.S., Barron, A.B., Myerscough, M.R.: Modelling food and population dynamics in honey bee colonies. PLOS ONE **8**, e59084 (2013)
9. Leoncini, I., Le Conte, Y., Costagliola, G., Plettner, E., Toth, A.L., Wang, M.W., Huang, Z., Becard, J.M., Crauser, D., Slessor, K.N., Robinson, G.E.: Regulation of behavioral by a primer pheromone produced by adult worker honey bees. PNAS **101**, 17559–17564 (2004)
10. Li, Z., Chen, Y., Zhang, S., Chen, S., Li, W., Yan, L., Shi, L., Wu, L., Sohr, A., Su, S.: Viral infection affects sucrose responsiveness and homing ability of forager honey bees Apis mellifera L. PLOS ONE **8**, e77354 (2013)
11. Neumann, P., Carreck, N.L.: Honey bee colony losses. J. Apic. Res. **49**, 1–6 (2010)
12. Perry, C.J., Sovik, E., Myerscough, M.R., Barron, A.B.: Rapid behavioral maturation accelerates failure of stressed honey bee colonies. PNAS **112**, 3427–3432 (2015)
13. Pettis, J.S., Lichtenberg, E.M., Andree, M., Stitzinger, J., Rose, R., vanEnglesdorp, D.: Crop pollination exposes honey bees to pesticides which alters their susceptibility to the gut pathogen Nosema ceranae. PLOS ONE **8**, e70182 (2013)

14. Toth, A.L., Robinson, G.E.: Worker nutrition and division of labour in honeybees. Anim. Behav. **69**, 427–435 (2005)
15. Vance, J.T., Williams, J.B., Elekonich, M.M., Roberts, S.P.: The effects of age and behavioral development on honey bee (Apis mellifera) flight performance. J. Exp. Biol. **212**, 26042611 (2009)
16. van Engelsdorp, D. Evans, J.D., Saegerman, C., Mullin, C., Haubruge, E., Nguyen, B.K., Frazier, M., Frazier, J., Cox-Foster, D., Chen, Y.P., Underwood, R., Tarpy, D.R., Pettis, J.S.: Colony collapse disorder: a descriptive study. PLOS ONE **4**, e6481 (2009)
17. Visscher, P.K., Dukas, R.: Survivorship of foraging honey bees. Insectes Sociaux **44**, 1–5 (1997)
18. Williams, G.R., Tarpy, D.R., vanEnglelsdorp, D., Chauzat, M.-P., Cox-Foster, D.L., Delaplane, K.S., Neumann, P., Pettis, J.S., Rogers, R.E.L., Shutler, D.: Colony collapse disorder in context. BioEssays **32**, 845–846 (2010)
19. Winston, M.L.: The Biology of the Honeybee. Harvard University Press, Cambridge (1987)
20. Woyciechowski, M., Moron, D.: Life expectancy and onset of foraging in the honeybee (Apis mellifera). Insectes Sociaux **56**, 193–201 (2009)

Zeta Function Associated with the Representation of the Braid Group

Kentaro Okamoto

Abstract There is a well-known zeta function of the $\mathbb{Z}$-dynamical system generated by an element of the symmetric group. By considering this zeta function as a model, we construct a new zeta function of an element of the braid group. In this article, we show that the Alexander polynomial which is the most classical polynomial invariant of knots can be expressed in terms of this braid zeta function. Furthermore, we show that the zeta function associated with the tensor product representation $\beta_{n,q}^{\otimes r}$ can be expressed by some braid zeta function for the case of special braids whose closures are isotopic to certain torus knots. Moreover, we introduce the zeta function associated with the Jones representation which is defined by using the R-matrix satisfying the Yang–Baxter equation. Then, we calculate this zeta function for $n = 3$ and show the relation between the Alexander polynomial and the Jones polynomial.

Keywords Zeta function · Braid group · Representation theory · Knot theory

1 Introduction

Let S_n be the symmetric group acting on the finite set $X := \{1, 2, \ldots, n\}$. Then, for any permutation $\sigma \in \mathrm{S}_n$, the $\mathbb{Z}$-*dynamical zeta function* of σ is defined as

$$\zeta(s, \sigma) := \exp\left\{\sum_{m=1}^{\infty} \frac{|\mathrm{Fix}(\sigma^m)|}{m} s^m\right\}, \tag{1}$$

where $\mathrm{Fix}(\sigma^m)$ is the set of fixed points defined as follows:

$$\mathrm{Fix}(\sigma^m) := \{x \in X \mid \sigma^m x = x\}. \tag{2}$$

K. Okamoto (✉)
Kyushu University, Nishi-ku, Fukuoka 819-0395, Japan
e-mail: k-okamoto@math.kyushu-u.ac.jp

B. Anderssen et al. (eds.), *The Role and Importance of Mathematics in Innovation*, Mathematics for Industry 25,
DOI 10.1007/978-981-10-0962-4_5

Example 1 When $\sigma = (1)(2,3) \in \mathrm{S}_3$, we can compute the set $\mathrm{Fix}(\sigma^m)$ as follows.

$$\mathrm{Fix}(\sigma^m) = \begin{cases} \{1,2,3\} & (m \equiv 0 \mod 2), \\ \{1\} & (m \equiv 1 \mod 2). \end{cases}$$

Then, we have

$$\begin{aligned} \zeta(s,\sigma) &= \exp\left\{ \sum_{m=1}^{\infty} \frac{3}{2m} s^{2m} + \sum_{m=1}^{\infty} \frac{1}{2m-1} s^{2m-1} \right\} \\ &= \exp\left\{ \sum_{m=1}^{\infty} \frac{2}{2m} s^{2m} + \sum_{k=1}^{\infty} \frac{1}{k} s^{k} \right\} \\ &= \exp\left\{ \log(1-s^2)^{-1} + \log(1-s)^{-1} \right\} \\ &= \frac{1}{(1-s^2)(1-s)}. \end{aligned}$$

In [9], the following interesting properties are shown by Kim, Koyama and Kurokawa.

Proposition 1 ([9, Proposition 1]) *For any permutation $\sigma \in \mathrm{S}_n$, the $\mathbb{Z}$-dynamical zeta function $\zeta(s,\sigma)$ has the following properties.*

(1) Let $\mathrm{Cycle}(\sigma)$ *be the set of primitive cycles of σ, and $l(P)$ be the length of cycle $P \in \mathrm{Cycle}(\sigma)$. Then, $\zeta(s,\sigma)$ has the Euler product:*

$$\zeta(s,\sigma) = \prod_{P \in \mathrm{Cycle}(\sigma)} \frac{1}{1 - s^{l(P)}}. \tag{3}$$

(2) Let $p_n : \mathrm{S}_n \longrightarrow \mathrm{GL}_n(\mathbb{Z})$ be the permutation representation. Then, $\zeta(s,\sigma)$ has the determinant expression:

$$\zeta(s,\sigma) = \det(I_n - p_n(\sigma)s)^{-1}. \tag{4}$$

(3) $\zeta(s,\sigma)$ satisfies the functional equation:

$$\zeta(s,\sigma) = \mathrm{sgn}(\sigma)(-s)^{-n}\zeta(1/s,\sigma), \tag{5}$$

where $\mathrm{sgn} : \mathrm{S}_n \longrightarrow \{\pm 1\}$ *is the signature of the permutation.*

(4) $\zeta(e^{-s},\sigma)$ satisfies an analogue of the **Riemann hypothesis***: i.e. all poles of $\zeta(e^{-s},\sigma)$ satisfy*

$$\mathrm{Re}(s) = 0. \tag{6}$$

Example 2 When $\sigma = (1)(2,3) \in \mathrm{S}_3$, the set of primitive cycles of σ is $\{(1), (2,3)\}$, and the length of (1) and $(2,3)$ are 1 and 2 respectively. Then we have the Euler product of $\zeta(s,\sigma)$ as follow.

$$\zeta(s,\sigma) = \frac{1}{(1-s^2)(1-s)}. \tag{7}$$

From (7), we obtain the following functional equation.

$$\zeta(1/s,\sigma) = \frac{1}{(1-1/s)(1-1/s^2)} = \frac{s^3}{(1-s)(1-s^2)} = s^3\zeta(s,\sigma).$$

On the other hand, using the following permutation matrix

$$p_3((1)(2,3)) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix},$$

we have the following determinant expression.

$$\det(I_3 - p_3((1)(2,3))s) = (1-s)(1-s^2) = \zeta(s,\sigma)^{-1}.$$

Furthermore, all poles of $\zeta(e^{-s},\sigma)$ satisfy

$$|e^{-s}| = e^{-\mathrm{Re}(s)} = 1.$$

Then $\mathrm{Re}(s) = 0$ and this is the analogue of the Riemann hypothesis.

Remark that the residue of $\zeta(s,\sigma)$ at $s = 1$ gives us only the information of the length of σ. If σ is a primitive cycle in itself, we can calculate the residue as follows:

$$\operatorname*{Res}_{s=1}\zeta(s,\sigma) = \lim_{s\to 1}(s-1)(1-s^n)^{-1} = -\frac{1}{n}. \tag{8}$$

Our goal is to generalize such properties to the case of the braid group. Consequently, we generalize $\zeta(s,\sigma)$ to the zeta function of a braid by using the *Burau representation* of the braid group. As an application the (nonnormalized) Alexander polynomial $\Delta_K(q)$ which is the famous knot invariant can be expressed by the residue of this new zeta function. This is analogous to the fact that the residue of the Dedekind zeta function at $s = 1$ has invariants of an algebraic field such as the class number, discriminant and regulator.

2 Preliminary

We recall the notations and settings on the braid group briefly. We refer to [3, 8] and [13] for more details. Let B_n be the n-string Artin braid group. It is known that B_n has the following presentation:

$$\mathrm{B}_n := \langle \sigma_i \ (1 \leq i \leq n-1) \mid \begin{array}{c} \sigma_i\sigma_j = \sigma_j\sigma_i \ (|i-j| \geq 2), \\ \sigma_i\sigma_{i+1}\sigma_i = \sigma_{i+1}\sigma_i\sigma_{i+1} \ (1 \leq i \leq n-2) \end{array} \rangle.$$

The generator σ_i can be identified with the crossing between the ith and $(i+1)$-st strands as Fig. 1 (see [3, Theorem 1.8]), and the multiplication implies that the braid obtained by attaching the generators from the top to the bottom.

Moreover we define the *closure* of the braid by connecting upper ends and lower ends (Fig. 2). The closure of σ is denoted by $\widehat{\sigma}$.

Next, we define the *torus type braid*. For a coprime pair $(n, m) \in \mathbb{N} \times \mathbb{Z}$, we denote

$$\sigma_{n,m} := (\sigma_1\sigma_2\cdots\sigma_{n-1})^m \in \mathrm{B}_n. \tag{9}$$

Then the closure of $\sigma_{n,m}$ is isotopic to the torus knot $T(n, m)$. The torus knot is a special kind of knot that lies on the surface of an (unknotted) torus in $\mathbb{R}^3$. Since B_1 is trivial, we assume that the number of strands n is larger than 1. For example, $\widehat{\sigma_1^3} = T(2, 3)$ is known as the trefoil knot (Fig. 3).

Let $\beta_{n,q}$ be the *Burau representation*, which is defined by

$$\beta_{n,q}(\sigma_i) := I_{i-1} \oplus \begin{pmatrix} 1-q & 1 \\ q & 0 \end{pmatrix} \oplus I_{n-i-1} \in \mathrm{GL}_n(\Lambda). \tag{10}$$

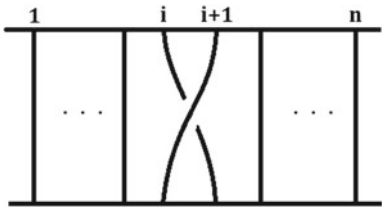


Fig. 1 Generator σ_i

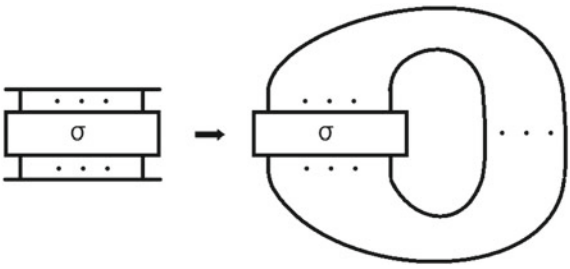


Fig. 2 The closure of a braid σ

Fig. 3 Trefoil knot $T(2,3) = \widehat{\sigma_1^3}$

Here $\Lambda := \mathbb{Z}[q^{\pm 1}]$ (the ring of Laurent polynomials over $\mathbb{Z}$), and q is a complex parameter. Hence, we can define the braid zeta function of $\sigma \in \mathrm{B}_n$ by the determinant expression.

Definition 1 We define the zeta function of $\sigma \in \mathrm{B}_n$ as below:

$$\zeta(s, \sigma; \beta_{n,q}) = \det(I_n - \beta_{n,q}(\sigma)s)^{-1}. \tag{11}$$

3 Main Results

We now can state the main results.

Theorem 1 (1) *For $\sigma \in \mathrm{B}_n$, the zeta function associated with the Burau representation $\zeta(s, \sigma; \beta_{n,q})$ satisfies the functional equation:*

$$\zeta(s, \sigma; \beta_{n,q}) = \mathrm{sgn}_q(\sigma^{-1})(-s)^{-n}\zeta(1/s, \sigma^{-1}; \beta_{n,q}). \tag{12}$$

Here $\mathrm{sgn}_q(\sigma) := \det(\beta_{n,q}(\sigma))$.
(2) *If $\widehat{\sigma}$ is a knot, the residue of $\zeta(s, \sigma; \beta_{n,q})$ at $s = 1$ is given as follows:*

$$\operatorname*{Res}_{s=1} \zeta(s, \sigma; \beta_{n,q}) = -\frac{1}{[n]_q} \Delta_{\widehat{\sigma}}(q)^{-1}. \tag{13}$$

Here $\Delta_{\widehat{\sigma}}(q)$ is the **Alexander polynomial** *of a knot $\hat{\sigma}$, and $[n]_q$ is the q-integer defined by*

$$[n]_q := \frac{1 - q^n}{1 - q}. \tag{14}$$

(3) *Assume that q is a point of the unit circle on the complex plane, in other words, q is expressed by $e^{i\theta}$ ($\theta \in \mathbb{R}$), and that the argument of q satisfies $|\theta| < 2\pi/n$. Then for any $\sigma \in \mathrm{B}_n$, the braid zeta function of σ satisfies an analogue of* **Riemann hypothesis***: all poles of $\zeta(e^{-s}, \sigma; \beta_{n,q})$ satisfy*

$$\mathrm{Re}(s) = 0. \tag{15}$$

Example 3 We compute the zeta function of the braid $\sigma = \sigma_1\sigma_2^{-1}\sigma_1\sigma_2^{-1} \in \mathrm{B}_3$. From the definition, we have

$$\zeta(s, \sigma; \beta_{3,q}) = \frac{1}{(1-s)(1-(q^2-2q+1-2q^{-1}+q^{-2})s+s^2)} = \zeta(s, \sigma^{-1}; \beta_{3,q}).$$

Then we can calculate the functional equation as follows.

$$\begin{aligned}\zeta(1/s, \sigma; \beta_{3,q}) &= \frac{1}{(1-1/s)(1-(q^2-2q+1-2q^{-1}+q^{-2})1/s+1/s^2)} \\ &= \frac{-s^3}{(1-s)(1-(q^2-2q+1-2q^{-1}+q^{-2})s+s^2)} \\ &= (-s^3)\zeta(s, \sigma^{-1}; \beta_{3,q}).\end{aligned}$$

The residue of $\zeta(s, \sigma; \beta_{3,q})$ at $s = 1$ can be written by

$$\operatorname*{Res}_{s=1} \zeta(s, \sigma; \beta_{3,q}) = -\frac{1}{1-q^2+2q+2q^{-1}-q^{-2}} = -\frac{1}{(1+q+q^2)(-1+3q^{-1}-q^{-2})}.$$

Here $-1+3q^{-1}-q^{-2}$ is the (nonnormalized) Alexander polynomial of $\widehat{\sigma}$. The nontrivial poles of $\zeta(e^{-s}, \sigma; \beta_{3,q})$ satisfy

$$1-(q^2-2q+1-2q^{-1}+q^{-2})e^{-s}+e^{-2s} = 0.$$

Then we have the following equation:

$$e^s + e^{-s} = q^2 - 2q + 1 - 2q^{-1} + q^{-2}.$$

Since $q = e^{i\theta}$, we have

$$\frac{e^s+e^{-s}}{2} = \cos 2\theta - 2\cos\theta + \frac{1}{2}. \tag{16}$$

If θ satisfy the condition $-\frac{2\pi}{3} < \theta < \frac{2\pi}{3}$, the right-hand side of (16) belongs to the interval $(-1, 1)$. Then, e^s is the element of $\{z \in \mathbb{C} \mid |z| = 1\}$. In other words, the real part of s is equal to 0.

Remark that $\zeta(s, \sigma; \beta_{n,q})$ does not have the Euler product expression, however, Theorem 1 is analogous to Proposition 1.

If σ is the torus type braid, we can calculate the zeta function explicitly. Moreover, we obtain the formula of the zeta function associated with the tensor product representation $\beta_{n,q}^{\otimes r}$ of the torus type braid.

Theorem 2 (1) *For a coprime pair (n, m), we obtain the following formula:*

$$\zeta(s, \sigma_{n,m}; \beta_{n,q}) = \frac{1 - q^m s}{(1-s)(1-q^{nm}s^n)}. \tag{17}$$

(2) *We choose $n \in \mathbb{Z}_{\geq 2}$, $m \in \mathbb{Z}$, $r \in \mathbb{N}$ such that the pair $(n, r! \cdot m)$ is coprime. Then we have*

$$\zeta(s, \sigma_{n,m}; \beta_{n,q}^{\otimes r}) = K_{n,m,r}(s, q) \cdot \prod_{l=1}^{r} \zeta(s, \sigma_{n,lm}; \beta_{n,q})^{b_{r,l,n}}. \tag{18}$$

Here

$$\zeta(s, \sigma; \beta_{n,q}^{\otimes r}) := \det(I_{nr} - \beta_{n,q}^{\otimes r}(\sigma)s)^{-1}, \tag{19}$$

$$K_{n,m,r}(s, q) := \prod_{l=1}^{r} \left(\frac{1-s}{1-q^{lm}s} \right)^{a_{r,l}+b_{r,l,n}}, \tag{20}$$

and $a_{r,l} := \binom{r}{l}(-1)^l$, $b_{r,l,n} := \binom{r}{l} \frac{(n-1)^l - (-1)^l}{n}$.

Example 4 The zeta function of the torus type braid $\sigma_{3,1} = \sigma_1\sigma_2 \in \mathrm{B}_3$ can be calculated as

$$\zeta(s, \sigma_{3,1}; \beta_{3,q}) = \frac{1 - qs}{(1-s)(1-q^3s^3)}.$$

By computing the Kronecker tensor product $\beta_{3,q}(\sigma_{3,1}) \otimes \beta_{3,q}(\sigma_{3,1})$, we calculate $\zeta(s, \sigma_{3,1}; \beta_{3,q}^{\otimes 2})$ as follows.

$$\begin{aligned}
\zeta(s, \sigma_{3,1}; \beta_{3,q}^{\otimes 2}) &= \frac{1}{(1-s)(1-q^2s)^2(1+qs+q^2s^2)^2(1+q^2s+q^4s^2)} \\
&= \frac{(1-s)^2}{(1-q^2s)^2} \cdot \frac{(1-qs)^2}{(1-s)^2(1-q^3s^3)^2} \cdot \frac{1-q^2s}{(1-s)(1-q^6s^3)} \\
&= K_{3,1,2}(s, q) \cdot \zeta(s, \sigma_{3,1}; \beta_{3,q})^2 \cdot \zeta(s, \sigma_{3,2}; \beta_{3,q}).
\end{aligned}$$

By Theorem 2, we obtain the following corollary.

Corollary 1 (1) *The Alexander polynomial of the torus knot $T(n, m)$ is given by*

$$\Delta_{T(n,m)}(q) = \frac{(1-q)(1-q^{nm})}{(1-q^n)(1-q^m)} = \frac{[nm]_q}{[n]_q[m]_q}. \tag{21}$$

(2) *We choose $n \in \mathbb{Z}_{\geq 2}$, $m \in \mathbb{Z}$, $r \in \mathbb{N}$ such that the pair $(n, r! \cdot m)$ is coprime. Then we have*

$$\operatorname*{Res}_{s=1}\zeta(s,\sigma_{n,m};\beta_{n,q}^{\otimes r}) = -\frac{1}{[n]_q^{n^{r-1}}}\prod_{l=1}^{r}\frac{1}{(1-q^{lm})^{a_{r,l}+b_{r,l,n}}\Delta_{T(n,lm)}(q)^{b_{r,l,n}}}. \tag{22}$$

Proof (1) By using the formula (17), we can calculate the residue of $\zeta(s,\sigma_{n,m};\beta_{n,q})$ as follows.

$$\operatorname*{Res}_{s=1}\zeta(s,\sigma_{n,m};\beta_{n,q}) = -\lim_{s\to 1}\frac{1-q^m s}{1-q^{nm}s} = -\frac{[m]_q}{[nm]_q}.$$

Then, from the Theorem 1 (2), we have the formula (21).
(2) The number of $[n]_q$ is calculated by

$$\begin{aligned}\sum_{l=1}^{r} b_{r,l,n} = \sum_{l=0}^{r} b_{r,l,n} &= \frac{1}{n}\sum_{l=0}^{r}\binom{r}{l}(n-1)^l - \frac{1}{n}\sum_{l=0}^{r}\binom{r}{l}(-1)^l \\ &= \frac{1}{n}(1+(n-1))^r \\ &= n^{r-1}.\end{aligned}$$

Hence the formula (22) follows from Theorem 1 (2).

4 Zeta Function and Jones Polynomial

In this section, we introduce the relation between the zeta function and the Jones polynomial. Let $\chi_{n,q}$ be the *Jones representation*, which is defined by

$$\chi_{n,q}(\sigma_i) := I_2^{\otimes(i-1)} \otimes R \otimes I_2^{\otimes(n-i-1)} \in \mathrm{GL}_{2^n}(\Lambda). \tag{23}$$

Here R is

$$R := \begin{pmatrix} 1 & & & \\ & 0 & q & \\ & 1 & 1-q & \\ & & & 1 \end{pmatrix}. \tag{24}$$

R is called *R-matrix* and is one of the solutions of *Yang-Baxter equation*.

Definition 2 For $\sigma \in \mathrm{B}_n$ we define the *zeta function associated with the Jones representation* as below.

$$\zeta(s,\sigma;\chi_{n,q}) := \det(I_{2^n} - \chi_{n,q}(\sigma)s)^{-1}. \tag{25}$$

Moreover we define the *deformation zeta function* as follows.

$$\zeta_t(s,\sigma;\chi_{n,q}) := \det(I_{2^n} - \chi_{n,q}(\sigma)\cdot\mu_t^{\otimes n}s)^{-1}. \tag{26}$$

Here μ_t is

$$\mu_t := \begin{pmatrix} 1 & 0 \\ 0 & t \end{pmatrix}. \tag{27}$$

Then, the Jones polynomial $J_{\widehat{\sigma}}(q)$ can be expressed by using the zeta function as follows.

$$\frac{d}{ds} \log \zeta_q(s, \sigma; \chi_{n,q}) \Big|_{s=0} = \mathrm{tr}(\chi_{n,q}(\sigma) \cdot \mu_q^{\otimes n}) = q^{\frac{1}{2}(n-\varepsilon(\sigma)-1)}(1+q) J_{\widehat{\sigma}}(q). \tag{28}$$

Here $\varepsilon(\sigma)$ is the *exponent sum* of σ defined by

$$\varepsilon(\sigma) = \sum_{k=1}^{r} e_k,$$

if σ can be expressed by $\sigma = \sigma_{i_1}^{e_1} \cdots \sigma_{i_r}^{e_r}$. For instance, for any 3-braid $\sigma \in \mathrm{B}_3$, the deformation zeta function (26) is expressed by the zeta function associated with the Burau representation.

$$\zeta_t(s, \sigma; \chi_{3,q}) = \frac{\zeta(ts, \sigma; \beta_{3,q}) \zeta(t^2 s, \sigma; \beta_{3,q})}{(1-s)(1-t^3 s)}. \tag{29}$$

Moreover, under the condition of Theorem 1 (3), we obtain the following generating function expression which converges when the absolute value of s is smaller than 1.

$$\zeta(s, \sigma; \beta_{n,q}) = \exp\left\{ \sum_{m=1}^{\infty} \frac{\mathrm{tr}\beta_{n,q}(\sigma^m)}{m} s^m \right\}. \tag{30}$$

Then, computing the logarithmic derivation of (30), we have

$$\frac{d}{ds} \log \zeta(s, \sigma; \beta_{n,q}) \Big|_{s=0} = \mathrm{tr}\beta_{n,q}(\sigma). \tag{31}$$

From (29) and (31), we obtain the following.

$$\frac{d}{ds} \log \zeta_t(s, \sigma; \chi_{3,q}) \Big|_{s=0} = 1 + t^3 + (t + t^2)\, \mathrm{tr}\, \beta_{3,q}(\sigma). \tag{32}$$

By (28) and (32), for any 3-braid $\sigma \in \mathrm{B}_3$ we have

$$J_{\widehat{\sigma}}(q) = q^{\frac{1}{2}\varepsilon(\sigma)}(q^{-1} + q - 1 + \mathrm{tr}\, \beta_{3,q}(\sigma)). \tag{33}$$

Furthermore if $\hat{\sigma}$ is a knot, calculating the residue of $\zeta_q(s, \sigma; \chi_{3,q})$ at $s = 1/q$ and $s = 1/q^2$, we also have the following relation by using Theorem 1 (2) and (29).

$$\operatorname*{Res}_{s=1/q} \zeta_q(s,\sigma;\chi_{3,q}) = -\frac{\zeta(q,\sigma;\beta_{3,q})}{q[3]_q \Delta_{\widehat{\sigma}}(q)(1-1/q)(1-q^2)} = \frac{\zeta(q,\sigma;\beta_{3,q})}{(1-q)^2[3]_q!\Delta_{\widehat{\sigma}}(q)}, \tag{34}$$

$$\operatorname*{Res}_{s=1/q^2} \zeta_q(s,\sigma;\chi_{3,q}) = -\frac{\zeta(1/q,\sigma;\beta_{3,q})}{q^2(1-1/q^2)(1-q)[3]_q \Delta_{\widehat{\sigma}}(q)} = \frac{\zeta(1/q,\sigma;\beta_{3,q})}{(1-q)^2[3]_q!\Delta_{\widehat{\sigma}}(q)}. \tag{35}$$

Here $[n]_q!$ is known as the *q-factorial* by $[n]_q! := [n]_q \cdot [n-1]_q \cdots [1]_q$.

Unfortunately, for $n > 3$, we have no simplifications of formula of the deformation zeta function (26). It is important and interesting problem to compute (26) in terms of topological representation such as the Burau representation. Furthermore, it is also meaningful to express the formula of colored Jones polynomial ([14]) by using the zeta functions. Thus, the formula of the Jones polynomial which is written by the Burau representation is expected to bring a new prospect for the zeta function and the volume conjecture([12]).

References

1. Alexander, J.W.: Topological invariants of knots and links. Trans. Amer. Math. Soc. **20**, 275–306 (1928)
2. Artin, E.: Theory of braids. Ann. Math. **48**(2), 101–126 (1947)
3. Birman, J. S.: Braids, Links, and Mapping Class Groups. Ann. Math. Studies, vol. 82. Princeton University Press, Princeton (1974)
4. Birman, J.S.: On the Jones polynomial of closed 3-braids. Inventiones mathematicae **81**, 287–294 (1985)
5. Blanchet, C., Marin, I.: Cabling Burau representation. preprint, (2006). arXiv:math/0701189
6. Deitmar, A., Koyama, S., Kurokawa, N.: Absolute zeta functions. Proc. Japan Acad. Ser. A Math. Sci. **84**(8), 138–142 (2008)
7. Karacuha, S., Lomp, C.: Integral calculus on quantum exteriora algebras. Int. J. Geom. Methods Mod. Phys. **11**1450026, 20 (2014)
8. Kassel, C., Turaev, V.: Braid Groups. Springer, Berlin (2008)
9. Kim, S., Koyama, S., Kurokawa, N.: The Riemann hypothesis and functional equations for zeta functions over $\mathbb{F}_1$. Proc. Japan Acad. Ser. A Math. Sci. **85**(6), 75–80 (2009)
10. Koyama, S., Nakajima, S.: Zeta functions of generalized permutations with application to their factorization formulas. Proc. Japan Acad. Ser. A Math. Sci. **88**(8), 115–120 (2012)
11. Kurpita, B.I., Murasugi, K.: A Study of Braids. Mathematics and its Applications, vol. 484. Kluwer Academic Publishers, Dordrecht (1999)
12. Murakami, H.: An introduction to the volume conjecture and its generalizations. Acta Math. Vietnam. **33**(3), 219–253 (2008). MR MR2501844
13. Ohtsuki, T.: Quantum Invariants—A Study of Knots, 3-Manifolds, and their Sets, Series on Knots and Everything, vol. 29. World Scientific Publishing Co., Inc., Singapore (2002)
14. Rozansky, L.: The universal R-matrix, Burau representation, and the Melvin-Morton expansion of the colored Jones polynomial. Adv. Math. **134**, 1–31 (1998)
15. Squier, C.C.: The Burau representation is unitary. In: Proceedings of the Amer. Math. Soc. vol. 90(2), pp. 199–202 (1984). MR 85b:20056
16. Yamamoto, T.: Inversion formulas for tridiagonal matrices with applications to boundary value problems. Numer. Funct. Anal. Optim. **22**, 357–385 (2001)

Fast, Scalable, and Energy-Efficient Parallel Breadth-First Search

Yuichiro Yasui and Katsuki Fujisawa

Abstract The breadth-first search (BFS) is one of the most centric processing in graph theory. In this paper, we presented a fast, scalable, and energy-efficient BFS for a nonuniform memory access (NUMA)-based system, in which the NUMA architecture was carefully considered. Our implementation achieved performance rates of 175 billion edges per second for Kronecker graph with 2^{33} vertices and 2^{37} edges on two racks of a SGI UV 2000 system with 1,280 threads and the fastest entries for a shared-memory system in the June 2014 and November 2014 Graph500 lists. It also produced the most energy-efficient entries in the first and second (small data category) and third, fourth, fifth, and sixth (big data category) Green Graph500 lists on a 4-socket Intel Xeon E5-4640 system.

Keywords Graph algorithm · NUMA-aware computing · Graph500 benchmark

1 Introduction

The breadth-first search (BFS) is one of the most important and fundamental graph algorithm. It can be used to obtain certain properties about the connections between the nodes in a given graph. BFS is not only used as a stand-alone, but also works as a subroutine in applications that determine the maximum flow [6, 7], connected components [5], graph centrality [3, 8], clustering [10]. Theoretically, the well-known algorithm of BFS [5] that uses the FIFO (First-first First-in) queue, has a linear complexity of $O(n + m)$, where $n = |V|$ is the number of vertices and $m = |E|$ is

Y. Yasui (✉)
Center for Co-evolutional Social Systems, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan
e-mail: y-yasui@imi.kyushu-u.ac.jp

K. Fujisawa
Institute of Mathematics for Industry, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan
e-mail: fujisawa@imi.kyushu-u.ac.jp

B. Anderssen et al. (eds.), *The Role and Importance of Mathematics in Innovation*, Mathematics for Industry 25,
DOI 10.1007/978-981-10-0962-4_6

the number of edges in a given graph $G = (V, E)$. This is optimal for theoretical purposes, but there is an actual need for efficient graph processing for large-scale real-world networks. Theoretical complexity analysis alone is not sufficient, because large-scale BFS computations require a significant amount of memory to enable multiple memory accesses over a wide memory space. In this paper, we discuss efficient graph traversal algorithms, in which the nonuniform memory access (NUMA) architecture was carefully considered.

Table 1 lists the BFS performances of related work and our implementations in terms of traversed edges per second (TEPS) and TEPS per watt (TEPS/W). Our latest implementation achieves a performance of over 40 GTEPS for Kronecker graph with SCALE 27 on a 4-socket NUMA system (such as with four CPU sockets). An alternative implementation achieves 174 GTEPS for Kronecker graph with SCALE 34 on a shared-memory SGI UV 2000 supercomputer based on a cache coherent (cc)-NUMA architecture with 1,280 threads (two UV 2000 racks), shown in Table 2. In addition, our implementations achieved the fastest entries for a shared-memory

Table 1 TEPS and TEPS/W scores on 4-socket Intel Xeon servers

Year	Reference	CPU	#Cores	SCALE	EF	TEPS (G)	TEPS/W (M)
2010	Graph500 [17]	Intel Xeon E5-4640 × 4	32	27	16	0.1	0.20
2010	Agarwal et al. [1]	Intel Xeon 7500 × 4	32	20	64	1.3	–
2012	Beamer et al. [2]	Intel Xeon E7-8870 × 4	40	28	16	5.1	–
2013	Yasui et al. [21]	Intel Xeon E5-4640 × 4	32	26	16	11.2	17.39
2014	Yasui et al. [22]	Intel Xeon E5-4640 × 4	32	27	16	29.0	45.43
2015	Yasui et al. [23]	Intel Xeon E5-4640 × 4	32	27	16	41.8	87.12

Table 2 TEPS score and scalability (weak scaling) on SGI UV2000

SCALE	EF	#CPUs	#Cores	TEPS (G)	Speedup
26	16	1	10	7.7	1.00×
27	16	2	20	15.3	1.98×
28	16	4	40	24.2	3.13×
29	16	8	80	42.1	5.46×
30	16	16	160	59.4	7.69×
31	16	32	320	94.8	12.28×
32	16	64	640	131.4	17.02×
33	16	128	1280	174.7	22.63×

Fig. 1 Our achievements of the Green Graph500 benchmark. **a** 1st (June 2013). **b** 2nd (Nov. 2013). **c** 3rd (June 2014). **d** 4th (Nov. 2014). **e** 5th (July 2015). **f** 6th (Nov. 2015)

single-node system in the June 2014, November 2014, July 2015, and November 2015 Graph500 lists.

Figure 1 show certificates that our implementations are highly energy-efficient, achieving first position in the small data category of the first and second and the big data category of the third, fourth, and fifth, and sixth Green Graph500 lists.

2 Preliminaries

2.1 Graph500 and Green Graph500 Benchmarks

The Graph500 benchmark[1] is designed to measure computational performance for applications that require an irregular memory access pattern. It is based on a score of the traversed edges per second (TEPS), which is computed by a generated edge list and an output of BFS [17]. The Green Graph500 benchmark[2] is designed to measure the energy efficiency of a computer in terms of TEPS per Watt [11]. These lists have been updated biannually since their introduction in 2010. Both benchmarks must perform the following steps:

1. **Generation**. This step generates the edge list of the Kronecker graph [16] with 2^{scale} vertices and $2^{scale} \cdot edgefactor$ edges by *scale* times the Kronecker prod-

[1] Graph500 benchmark: http://www.graph500.org.

[2] Green Graph500 benchmark: http://green.graph500.org.

ucts of an initiator matrix $\begin{pmatrix} 0.57 & 0.19 \\ 0.19 & 0.10 \end{pmatrix}$, where the *scale* and the *edgefactor* are input parameters.

2. **Construction (timed)**. This step constructs the graph representation from the edge list obtained in Step 1.
3. **BFS iterations (timed)**. This step executes 64 BFSs from different source vertices and computes the median TEPS score from 64 TEPS scores.

2.2 Parallel Breadth-First Search

At first, we assume that the input of a BFS is a graph $G = (V, E)$ consisting of a set of vertices V and a set of edges E. A BFS explores the various edges spanning all other vertices $v \in V \backslash \{s\}$ from the source vertex $s \in V$ in a given graph G, and outputs the *predecessor map* π, which is a map from each vertex v to its parent. When the predecessor map $\pi(v)$ points to only one parent for each vertex $v \in V$, it represents a tree with the root vertex $s \in V$. In addition, the predecessor map of source vertex $\pi(s)$ points itself to s.

The well-known textbook algorithm for breadth-first search is not suitable for parallelism, which uses the FIFO (First-first First-in) queue. Therefore, we use Algorithm 1, utilizes two queues: *current queue* CQ and *next queue* NQ, called level-synchronized breadth-first search. In this algorithm, we assume that an input graph $G = (V, A^F)$ based on an adjacency vertex list A^F represents a directed graph, where an adjacency list $A^F(v)$ contains the adjacency vertices w of outgoing edges $(v, w) \in E$ for each vertex $v \in V$. If an input graph is undirected, it uses (v, w) and (w, v) edges instead of $(v, w) \in E$ edges. This algorithm starts with the current queue

Table 3 Number of edges traversed by each traversal direction in a BFS of Kronecker graph with SCALE 26 and edgefactor 16

Level	Top–down m_F	Bottom–up m_B	Direction-optimizing (best case) $\min(m_F, m_B)$
0	2	2,103,840,895	2
1	66,206	1,766,587,029	66,206
2	346,918,235	52,677,691	52,677,691
3	1,727,195,615	12,820,854	12,820,854
4	29,557,400	103,184	103,184
5	82,357	21,467	21,467
6	221	21,240	227
Total	2,103,820,036	3,936,072,360	65,689,631
Ratio	100.00 %	187.09 %	3.12 %

CQ as the source s. At each level k, this algorithm finds unvisited adjacency vertices $A^F(v)$, $v \in$ CQ that are connected to the current queue CQ, and appends them to the next frontier NQ for level $k+1$. After the edge traversal, NQ becomes the current queue CQ for the next level. When the frontier is empty, the algorithm terminates. Finally, this algorithm requires atomic operations for a consistency in the most deeply loop, which called the same number of edges. It will be the performance bottleneck, because of it guaranteed high cost in general.

Algorithm 1: Level-synchronized Breadth-first search

```
   Input  : Digraph G = (V, A^F), vertex s.
   Data   : current queue CQ, next queue NQ,
            and visited vertices VS.
   Output: predecessor π(v), ∀v ∈ V.
1  π(v) ←⊥, ∀v ∈ V\{s}
2  π(s) ← s
3  VS ← {s}
4  CQ ← {s}
5  NQ ← ∅;
6  while CQ ≠ ∅ do
7  |  NQ ← Top–down(G, CQ, VS, π)
8  |  swap(CQ, NQ)
9  Procedure Top–down(G, CQ, VS, π)
10 |  NQ ← ∅
11 |  for v ∈ CQ in parallel do
12 |  |  for w ∈ A^F(v) do
13 |  |  |  if w ∉ VS atomic then
14 |  |  |  |  π(w) ← v
15 |  |  |  |  VS ← VS ∪ {w}
16 |  |  |  |  NQ ← NQ ∪ {w}
17 |  return NQ
```

Beamer et al. [2] proposed a direction-optimizing algorithm for breadth-first search (Algorithm 2) that reduces the number of edges explored. Similar to Algorithm 1, this algorithm performs a traversal procedure (lines 7–10) and swaps of NQ and CQ (line 7) at each level. This algorithm has two different traversal directions: *Top–down* and *Bottom–up*, chooses one from these directions by the size of current queue $\left|\deg_G v, v \in \text{CQ}\right|$. The former traverses the next queue NQ from the current queue CQ, whereas the latter finds the *frontier* CQ from all unvisited vertices $V \backslash \text{VS}$ as candidate neighbors. Table 3 describes how the traversal direction is determined for the top–down and bottom–up approaches (line 7). The traversal direction moves from the top–down to the bottom–up in the *growing* phase $|\text{CQ}| < |\text{NQ}|$, and returns from the bottom–up to the top–down in the *shrinking* phase $|\text{CQ}| \geq |\text{NQ}|$. The computational complexities are $O(m)$ for the top–down direction and $O(m \cdot \text{diam}_G)$ for the bottom–up direction, where m is the number of edges and diam_G is the diameter of the given graph. The direction-optimizing algorithm that combines these algorithms has $O(m \cdot \text{diam}_G)$ complexity, however, it works well experimentally.

Algorithm 2: Direction-optimizing Breadth-first search

Input : Digraph $G = (V, A^F, A^B)$, vertex s.
Data : frontier queue CQ, next queue NQ, and visited vertices VS.
Output: Predecessor map $\pi(v), \forall v \in V$.

```
1  π(v) ←⊥, ∀v ∈ V\{s}
2  π(s) ← s
3  VS ← {s}
4  CQ ← {s}
5  NQ ← ∅
6  while CQ ≠ ∅ do
7      if use_TopDown(G, CQ, NQ, VS) then
8          NQ ← Top–down(G, CQ, VS, π)
9      else
10         NQ ← Bottom–up(G, CQ, VS, π)
11     swap(CQ, NQ)
12 Procedure Bottom–up(G, CQ, VS, π)
13     NQ ← ∅
14     for w ∈ V \ VS in parallel do
15         for v ∈ A^B(w) do
16             if v ∈ CQ then
17                 π(w) ← v
18                 VS ← VS ∪ {w}
19                 NQ ← NQ ∪ {w}
20                 break
21     return NQ
```

3 NUMA-Aware Programming with ULIBC

Current systems are designed based on the NUMA architecture. On such NUMA and cc-NUMA systems, each processor has a local memory, and these connect to one another via an interconnect, such as the Intel QPI, AMD HyperTransport, or SGI NUMAlink 6, shown in Fig. 3b as an example. A thread running on a processor core can access a local memory faster than access remote (nonlocal) memory on a NUMA system. The performance of BFS depends on the speed of memory access, because the complexity of memory accesses is greater than that of computation. Therefore, the NUMA-aware speedups effect to improve the performance of breadth-first search. Table 4 lists used NUMA systems in this paper.

We investigate the characteristics of NUMA system in terms of memory bandwidth using the STREAM benchmark and the Thread Affinity Interface of Intel compiler. If a binary is built by the Intel compiler, it controls the pinning of running threads to logical cores using `KMP_AFFINITY` environment value. We called this the CPU affinity. Two affinity strategies *scatter* and *compact* are available on Intel Thread Affinity Interface. Each thread is assigned by the affinity strategy and the

Table 4 NUMA systems

System	CPU name (LLC size)	Sockets × Cores × SMT	RAM	Compiler
SB4	Xeon E5-4640 (20.0 MB)	4 × 16 × 2	512.0 GB	ICC-15.0.1
UV2000	Xeon E5-4650 v2 (25.0 MB)	256 × 10 × 1	64.0 TB	ICC-14.0.0

Table 5 Number of "Sockets × Physical-Cores × SMT" for each CPU affinity on SB4 (4-socket 8-core server

CPU Affinity	Number of threads						
	1	2	4	8	**16**	32	64
Compact	1 × 1 × 1	1 × 1 × 2	1 × 2 × 2	1 × 4 × 2	1 × 8 × 2	2 × 8 × 2	4 × 8 × 2
Scatter	1 × 1 × 1	2 × 1 × 1	4 × 1 × 1	4 × 2 × 1	4 × 4 × 1	4 × 8 × 1	4 × 8 × 2

number of threads, shown in Table 5. In this table, we describe the thread placement on SB4 in terms of the number of sockets, physical-cores on a socket, and SMT (Simultaneous Multithreading) on a physical-core. *Compact* assigns the thread $n + 1$ to a free thread context as close as possible to the thread context where the n thread was placed. *Scatter* distributes the threads as evenly as possible across the entire system.

Figure 2 plots a memory bandwidth for each affinity type, versus number of threads using the TRIAD operation of the STREAM benchmark. The TRIAD operation computes $\mathbf{a} \leftarrow \mathbf{b} + r \cdot \mathbf{c}$ using three vectors $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathbf{R}^n$ with n elements, whose element holds double precision floating point number (8 bytes). This figure shows the boxplot of memory bandwidth in GB/s (giga bytes per second) for the array with range 1GB to 8 GB that indicates (minimum, first quartile, median, third quartile, and maximum). First, this figure shows that scatter and compact are more stable than the default caused the performance degradation at four threads and larger. If the affinity settings are not enabled, a parallel computation cannot elicit high performance on a NUMA system. Second, the scatter is larger than the compact in case of same number of threads used. It means the memory bandwidth depends on the number of sockets used. Therefore, the local memory access can access using a memory bus independently, because a NUMA system consists of pairs of CPU sockets and local memory. Finally, we observed that the bandwidth also depends on the number of physical-cores in the compact affinity. However, it is not as strong as the number of sockets,

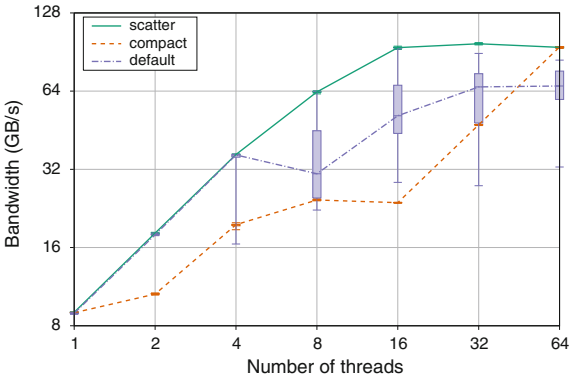


Fig. 2 Box plot of bandwidths GB/s by using the TRIAD operation of the STREAM benchmark with the array size as between 1 to 8 GB, versus number of threads for each CPU affinity. Each line connects median GB/s score

because it is small for the performance gap between 8 threads (4 physical-cores) and 16 threads (8 physical-cores). This shows that a single-thread implementation cannot elicit memory bandwidth. From these results, we considered the placements of running threads and referenced data to improve the performance on a NUMA system.

Figure 3a shows bandwidths between two NUMA nodes on SB4 using the `numactl` command. We specified the placement of running threads and referenced data by the `--cpunodebind` option for CPU binding and the `--membind` option for Memory binding of `numactl` command. From this result, the local memory access (24 GB/s) is approximately 8 times faster than the remote memory access (2.9–3.4 GB/s). In addition, we described the obtained bandwidths in a punch figure of SB4, shown in Fig. 3a.

In this section, we propose a general management approach for processor and memory affinities on a NUMA system. Previous work includes; the Portable Hardware Locality (hwloc) [4], the Likwid [18], Thread Affinity Interface of Intel compiler [13], and OpenUH compiler [12]. However, we cannot find a library for obtaining the position of each running thread, such as the CPU socket index, physical-core index in each CPU socket, or thread index in each physical-core. We have developed a management library for processor and memory affinities, called ULIBC. ULIBC supports many operating systems (although we have only confirmed Linux, Solaris, and AIX) and is available at

https://bitbucket.org/yuichiro_yasui/ulibc.

ULIBC provides an "MPI rank"-like index, starting at zero, for each CPU socket, each physical core in each CPU socket, and each thread in each physical core, which are available for the corresponding process, respectively. We have already applied ULIBC to graph algorithms for the shortest path problem [20], BFS [14, 15, 21–23], and mathematical optimization problems [9].

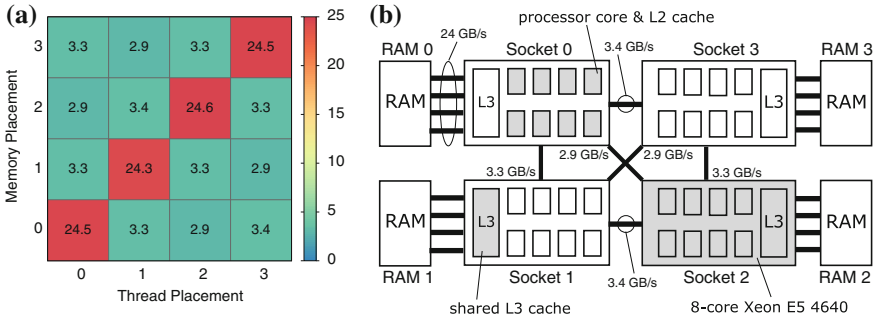


Fig. 3 Bandwidth score GB/s between arbitrary NUMA nodes by using the TRIAD operation of the STREAM benchmark on 4-socket server. **a** Bandwidths between NUMA nodes. **b** SB4 (4-socket Intel Xeon E5-4640) system

4 NUMA-Aware Breadth-First Search

We proposed some efficient algorithms and implementations in terms of two strategies; (a) reducing the number of scanning edges like as the direction-optimizing algorithm and (b) improving the locality of memory access. Table 6 shows our algorithms and implementations BD13 [21], ISC14 [22], HPCS15 [23] based on their. The TEPS scores obtained for Kronecker graph with SCALE 27 on SB4. Reference is published at the Graph500 Benchmark. Dir. Opt. means the result that are obtained by Beamer et al. The entire speedup is 8.33 ($=\frac{42.5}{5.1}$) times faster than that of the original direction-optimizing algorithm. We focused the NUMA-aware partitioned graph representation and the sorting for adjacency list and vertex index.

4.1 NUMA-Aware Partitioned Graph Representation

Our NUMA-optimized algorithms that are based on Beamer et al.'s direction optimizing algorithm [2], use the 1-D partitioning for sets of vertices and edges to improve the access to local memory [21–23]. Each sets of partial vertices V_k and edges E_k on the kth NUMA node is defined by

$$\begin{aligned} V_k &= \left\{ v_j \mid j \in \left[\frac{n}{\ell} \cdot k, \frac{n}{\ell} \cdot (k+1) \right) \right\}, \\ E_k &= \{(v,w) \mid ((v,w) \in E) \wedge (v \in V) \wedge (w \in V_k)\} \end{aligned} \tag{1}$$

where n is the number of vertices and the divisor ℓ is set to the number of NUMA nodes (CPU sockets). Our implementations used partial adjacency lists $A_k^F(v), v \in V$ for the top–down direction and $A_k^B(w), w \in V_k$ for the bottom–up direction on the k-th NUMA node as follows:

Table 6 Implementation and algorithms and data structure used

Algorithms reference	Reference [17]	NUMA-BFS [1]	Dir. Opt. [2]	BD13 [21]	ISC14 [22]	HPCS15 [23]
Level sync. BFS	×	×				
Direction optimizing			×	×	×	×
(a) NUMA-aware Graph		×		×	×	×
(a) Adjacency list sorting					×	×
(b) Vertex index sorting						×

$$A_k^F(v) = \{w \mid (v, w) \in E_k\}, v \in V,$$
$$A_k^B(w) = \{v \mid (v, w) \in E_k\}, w \in V_k. \tag{2}$$

As before, the working spaces NQ, VS, and π in Algorithms 1 and 2 are partitioned into NQ_k, VS_k, and π_k by using corresponding partial vertices V_k and are allocated to the local memory on the k-th NUMA node with the memory pinned. In contrast with the NQ, VS, and π, the current queue CQ are duplicated into CQ_k, which are allocated to the local memory on the k-th NUMA node. Therefore, the swap procedure in Algorithms 1 and 2 constructs each CQ_k from all partial NQ_k. This operation requires all-to-all communication the same as the all-gather.

4.2 Adjacency List Sorting

The bottom–up procedure checks that each unvisited vertex connects to the frontier vertices that are included in the current queue. Therefore, the number of loops at the bottom–up procedure depends on the order of each adjacency list for each unvisited vertex. It is difficult to obtain the optimal ordering for the adjacency vertex list, and we use the heuristic that constructs an adjacency vertex list $A(v)$ of each vertex $v \in V$, which is sorted by by the out-degree. Table 7 shows a comparison of the number of traversed edges for each level in the top–down and the bottom–up for each order; *Descending order* and *Ascending order*. The table shows that most traversed edges were concentrated in Level-2 and that the number of traversed edges is affected by each order.

We investigate the breakdown of the vertex finding for each adjacency order shown in Table 8. First, the table shows that the number of zero-degree vertices is half the total number of vertices, and almost of vertices are found in the bottom–up excluding them. Second, the difference of the order effects a position in adjacency

Table 7 Number of traversed edges in a BFS for Kronecker Graph with SCALE 27

Level	Descending order			Ascending order	
	Top–down	Bottom–up	Hybrid	Bottom–up	Hybrid
0	22	4,223,250,243	22	4,223,039,317	22
1	239,930	3,258,645,723	239,930	4,063,345,725	239,930
2	1,040,268,126	**83,878,899**	**83,878,899**	**848,743,124**	**848,743,124**
3	3,145,608,885	19,616,130	19,616,130	19,935,737	19,935,737
4	37,007,608	139,606	139,606	139,868	139,868
5	98,339	41,846	41,846	41,846	41,846
6	260	41,586	260	41,586	260
Total	4,223,223,170	7,585,614,033	103,916,693	9,155,287,203	869,100,787
%	100 %	179.6 %	2.5 %	216.8 %	20.6 %

Table 8 Breakdown of the vertex finding in a BFS of Kronecker graph with SCALE 27

	Descending order		Ascending order	
Isolated vertices	71,140,085	53.00 %	71,140,085	53.00 %
Top–down	215,070	0.16 %	215,070	0.16 %
Bottom–up (1st)	**60,462,127**	**45.05 %**	**25,489,401**	**18.99 %**
Bottom–up (2nd or later)	2,358,918	1.76 %	37,331,644	27.81 %

list. The number of vertices that are found in the first position using the descending order, is more than twice larger compared with the ascending order. In particular, when using descending order, we can estimate that a BFS requires the almost of CPU time in bottom–up at the first vertex of adjacency list. The descending order improved the performance of the direction-optimizing algorithm as 2.66 ($=\frac{29.0}{10.9}$ from Table 6) times for Kronecker graph with SCALE 27, reducing the unnecessary edge traversals.

4.3 Vertex Index Sorting

We propose a vertex index sorting technique to improve the locality of accessing working spaces VS in the Top–down and CQ in the Bottom–up (mainly), which is similar to that in [19]. The current queue CQ at the Bottom–up is implemented by the bitmap structure, which represents the set for each vertex as one bit. If the corresponding vertices are in a set, the bit is 0; otherwise, the bit is 1. Because the number of accesses of each element is equal to the in-degree $\deg_G^{in}$ of corresponding vertex $v \in V$, if the elements frequently used are located closely in a memory, a cache memory works well. At the first of the graph construction, our technique constructs new vertex indices $\{0, 1, ..., n-1\}$ as follows:

$$\deg_G^{\text{in}}(v_0) \geq \deg_G^{\text{in}}(v_1) \geq ... \geq \deg_G^{\text{in}}(v_{n-1}). \tag{3}$$

Kronecker graphs, which are used in the Graph500 benchmark, have a power-law degree distribution. Figure 4a, b represent the numbers of accesses for the visited vertices VS in Top–down and the current queue CQ in Bottom–up of each BFS required by Direction-optimizing [2] and our implementation applied vertex index sorting. These obtained by the average of 64 BFSs for Kronecker graph with SCALE 20. Our new algorithm drastically improves the locality of vertex access from irregular memory accesses in BFS. Applying our vertex index sorting, the access frequency distribution for each vertex in a BFS is similar to the degree distribution of input graph. Table 6 shows that our new algorithm is 1.47 ($=\frac{42.5}{29.0}$) times faster than our previous work.

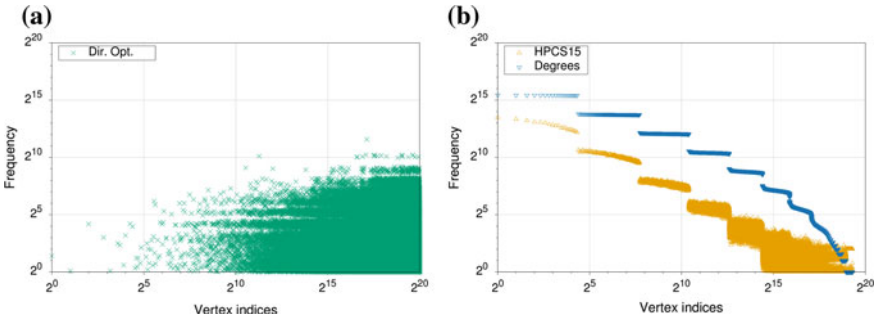


Fig. 4 Access frequency of vertex traversal for Kronecker graph with SCALE 20. **a** Previous work. **b** This paper and degrees

4.4 Numerical Results on 4-Socket Xeon System

Figure 5a compares the BFS performance of our three implementations for Kronecker graph on the SB4 system with 64 threads. These results obtained by varied with problem size SCALE and constant edgefactor 16, which determine graph size as 2^{scale} vertices and $2^{\text{scale}} \cdot$ edgefactor edges. All implementations can solve up to SCALE 29. They reaches the peak at SCALE 26 or 27 and suffers performance degradation for large SCALEs 28 and 29. Figure 5b compares the strong scaling performance of our three implementations for Kronecker graph with SCALE 27 on the SB4 system. These results obtained by varied with the number of threads. All implementations obtained approximately same performance in terms of the parallelization efficiency.

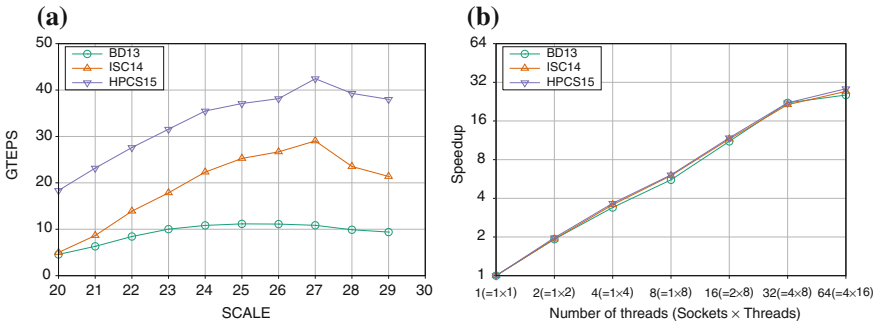


Fig. 5 BFS performance of our implementations for Kronecker graph on SB4. **a** GTEPS with 64 threads. **b** Strong scaling performance for SCALE 27

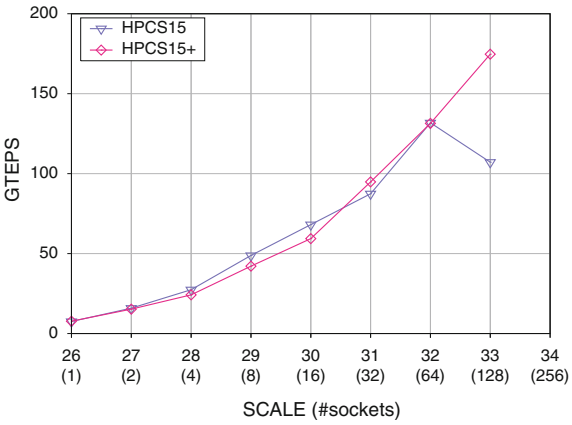


Fig. 6 Weak scaling with SCALE 26 per CPU socket on UV 2000

4.5 Numerical Results on SGI UV 2000

Figure 6 shown the weak scaling performance of our HPCS15 implementation, which collects TEPS scores with fixed problem size as SCALE 26 per socket. From this figure, the HPCS15 reaches the peak at SCALE 32 on one UV2000 rack with 640 cores, and suffers performance degradation for 2 racks. Therefore, we developed HPCS15+ to improve the scalability of HPCS15, which uses the backward graph A^B instead of the forward graph A^F to reduce an inefficient all-gather operation with small elements (and also a memory requirement) in swap procedure at Top–down likes as [1]. The HPCS15+ scales up to 1,280 threads and achieves 174.704 GTEPS for SCALE 34 shown in this figure.

5 Conclusion

In this paper, we described efficient breadth-first algorithms for the large scale networks on a single NUMA system, which adapted by NUMA-aware graph representation [21], and the sorting techniques for adjacency list and vertex index [22, 23]. Our HPCS15 achieved over 40 GTEPS on the 4-way Intel Xeon server. This result is the most power-efficient entries in the June 2014, November 2014, July 2015, and November 2015 Green Graph500 lists. Finally, we showed that our HPCS15+ scales up to 1,280 threads, and achieves 174 GTEPS on two racks of the UV 2000 system with 1,280 threads. This result is the fastest entry for a shared-memory single-node system in the November 2014 and July 2015, and November 2015 list of the Graph500 benchmark.

Acknowledgments This research was supported by the Core Research for Evolutional Science and Technology (CREST) and the Center of Innovation (COI) programs of the Japan Science and Technology Agency (JST), the Institute of Statistical Mathematics (ISM), and Silicon Graphics International (SGI) Corp.

References

1. Agarwal, V., Petrini, F., Pasetto, D., Bader, D.A.: *Scalable graph exploration on multicore processors*. In: Proceedings of the ACM/IEEE International Conference on High Performance Computing, Networking, Storage and Analysis (SC10), IEEE Computer Society (2010)
2. Beamer, S., Asanović, K., Patterson, D.A.: *Direction-optimizing breadth-first search*. In: Proceedings of the ACM/IEEE International Conference on High Performance Computing, Networking, Storage and Analysis (SC12), IEEE Computer Society, p 12 (2012)
3. Brandes, U.: A faster algorithm for betweenness centrality. J. Math. Sociol. **25**(2), 163–177 (2001)
4. Broquedis, F., Clet-Ortega, J., Moreaud, S., Furmento, N., Goglin, B., Mercier, G., Thibault, S., Namyst, R.: hwloc: a generic framework for managing hardware affinities in HPC applications. In: Proceedings of the 18th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP2010) (2010)
5. Cormen, T., Leiserson, C., Rivest, R.: Introduction To Algorithms. MIT Press, Cambridge (1990)
6. Dinic, E.A.: Algorithm for solution of a problem of maximum flow in a network with power estimation. Sov. Math. Dokl. **11**, 1277–1280 (1970)
7. Edmonds, J., Karp, R.M.: Theoretical improvements in algorithmic efficiency for network flow problems. J. ACM **19**(2), 248–264 (1972)
8. Frasca, M., Madduri, K., Raghavan, P.: *NUMA-aware graph mining techniques for performance and energy efficiency*. In: Proceedings of the ACM/IEEE International Conference on High Performance Computing, Networking, Storage and Analysis (SC12), IEEE Computer Society (2012)
9. Fujisawa, K., Endo, T., Yasui, Y., Sato, H., Matsuzawa, N., Matsuoka, S., Waki, H.: *Petascale general solver for semidefinite programming problems with over two million constraints*. In: Proceedings of the IEEE International Symposium on Parallel and Distributed Processing (IPDPS 14), IEEE Computer Society (2014)
10. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA **99**, 7821–7826 (2002)
11. Hoefler, T.: *GreenGraph500 Submission Rules*. http://green.graph500.org/greengraph500rules.pdf
12. Huang, L., Jin, H., Yi, L., Chapman, B.: Enabling locality-aware computations in OpenMP. Exploring Languages for Expressing Medium to Massive On-Chip Parallelism archive J. Sci. Program. **18**(3–4), 169–181 (2010)
13. Intel(R) C++ Compiler XE 13.1 User and Reference Guide, Intel Thread Affinity Interface
14. Iwabuchi, K., Sato, H., Yasui, Y., Fujisawa, K.: *Hybrid BFS approach using semi-external memory*. In: Proceedings of the International Workshop on High Performance Data Intensive Computing (HPDIC2014) (2014)
15. Iwabuchi, K., Sato, H., Yasui, Y., Fujisawa, K., Matsuoka, S.: *NVM-based hybrid BFS with memory efficient data structure*. In: Proceedings of the IEEE International Conference on BigData, IEEE Computer Society (2014)
16. Leskovec, J., Chakrabarti, D., Kleinberg, J., Faloutsos, C., Ghahramani, Z.: Kronecker graphs: An approach to modeling networks. J. Mach. Learning Res. **11**, 985–1042 (2010)
17. Murphy, R.C., Wheeler, K.B., Barrett, B.W., Ang, J.A.: *Introducing the Graph500*, In: Proceedings of the Cray User Group (2010)

18. Treibig, J., Hager, G., Wellein, G.: LIKWID: A lightweight performance-oriented tool suite for x86 multicore environments. PSTI2010 (2010)
19. Ueno, K., Suzumura, T.: *Highly scalable graph search for the Graph500 Benchmark*. In: Proceedings of the 21st International ACM Symposium on High-Performance Parallel and Distributed Computing (HPDC'12), pp. 149–160, ACM (2012)
20. Yasui, Y., Fujisawa, K., Goto, K., Kamiyama, N., Takamatsu, M.: NETAL: High-performance implementation of network analysis library considering computer memory hierarchy. J. Oper. Res. Soc. Japan **54**(4), 259–280 (2011)
21. Yasui, Y., Fujisawa, K., Goto., K.: *NUMA-optimized parallel breadth-first search on multicore single-node system*. In: Proceedings of the IEEE International Conference on BigData, IEEE Computer Society (2013)
22. Yasui, Y., Fujisawa, K., Sato, Y.: *Fast and energy-efficient breadth-first search on a single NUMA system*. In: Kunkel, J.M., Ludwig, T., Meuer, H. (eds.) Supercomputing, Lecture Notes in Computer Science. vol. 8488, pp. 365–381. Springer, Berlin (2014)
23. Yasui, Y., Fujisawa, K.: *Fast and scalable NUMA-based thread parallel breadth-first search*. In: Proceedings of the HPCS 2015 (The 2015 International Conference on High Performance Computing & Simulation), ACM, IEEE, IFIP, Amsterdam, the Netherlands (2015)

Basic Research for the Patient-Specific Surgery Support System—An Identification Algorithm of the Mesentery Using 3D Medical Images

Kazushi Ahara, Munemura Suzuki, Yoshitaka Masutani and Takuya Ueda

Abstract Currently, laparoscopic surgery has widely been accepted as a major option for the treatment of abdominal diseases such as gallbladder stone, gastric cancer, and colon cancer. Preoperative anatomical assessment is crucial for laparoscopic surgery. Anatomically, the digestive tract is covered with a bilayered thin-membranous tissue called 'mesentery' and the arteries and veins courses within the mesentery. Although preoperative anatomical information of mesentery will add clinical information, current imaging modalities cannot visualize mesentery because of its limits of spatial resolution of medical imaging. In this study, combining the anatomical features of the mesentery and data such as an interrelation between the mesentery and the surrounding organs obtained with empirical laws of physicians, we propose technology to reproduce geometrically a shape of the mesentery, of which describing is impossible from 3D medical images. Rendering and visualizing the structure of mesentery from the data of preoperative medical images is expected in clinical medicine.

Keywords 3D medical images · Mesentery · Interpolation

K. Ahara (✉)
Meiji University, 4-21-1 Nakano, Nakano-ku, Tokyo 164-8525, Japan
e-mail: ahara@meiji.ac.jp

M. Suzuki
Suzuki Medical Imaging Lab., 217-6-205, Taniyamachuo,
Kagoshima city 891-0104, Japan
e-mail: munemura.surf@gmail.com

Y. Masutani
Hiroshima City University, Hiroshima, Japan
e-mail: masutani@hiroshima-cu.ac.jp

T. Ueda
Chiba Medical Center, Chiba, Japan
e-mail: takuedarad@gmail.com

B. Anderssen et al. (eds.), *The Role and Importance of Mathematics in Innovation*, Mathematics for Industry 25,
DOI 10.1007/978-981-10-0962-4_7

1 Introduction

Recently mathematicians have contributed to medical researchers in Japan. As importance of mathematical approach in medical fields is recognized, many mathematical collaborative projects with medical researches are adopted. Such mathematical approach in medical field is called medical mathematics. One of the main themes in medical mathematics is the research of modeling and simulation of PDE (diffusion equation). Another theme is analyzing medical images and visualization of anatomical structures of organs.

The purpose of this study is to establish the method to visualize 3 dimensional morphology of the mesentery, a membrane in the abdomen, which is invisible with current imaging modalities.

We propose a method to estimate the normal vector field which meets the various requirements for anatomical constraint. Then we approximate and render the image of the smooth surface of membrane using interpolation technique.

2 Anatomical Features of the Mesentery

The mesentery is a membranous anatomical structure, that covers aero digastric tract such as the stomach, the colon, and the small intestine. The colon and the small intestine do not float within the abdominal cavity. Figure 1 shows the abdominal cavity, cutting most of the small intestine to explain the structure. The small intestine is wrapped and hanging from the dorsal wall of the abdominal cavity by a membranous tissue called mesentery (Fig. 2left).

The left upper figure of Fig. 2 shows the cross sectional image of the abdomen. The colon is also wrapped by the membrane. The ascending (* in Fig. 1) and descending

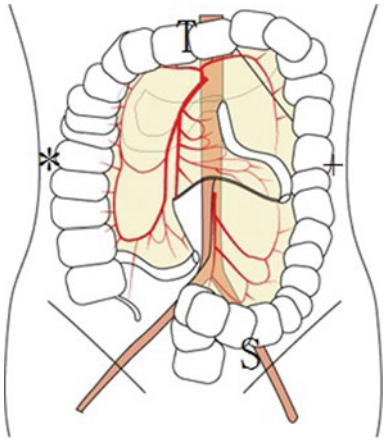


Fig. 1 The scheme of abdominal cavity, ascending colon (*), transverse colon (T), descending colon (+), and sigmoid colon (S)

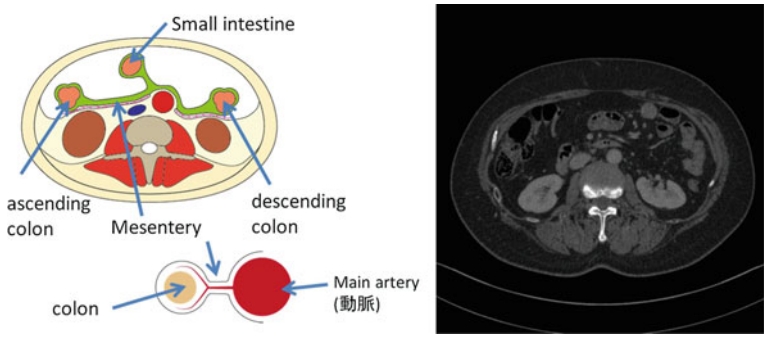


Fig. 2 The cross section of the abdominal cavity and the illustration of the relationship between colon, artery, and mesentery (*left*), and CT scan image (*right*)

(+ in Fig. 1) colon attached to the posterior wall of the abdomen (Fig. 2left). But the transverse colon (T in Fig. 1) and sigmoid colon (S in Fig. 1) are hung like the small intestine (the membranous structure that hangs transverse colon and sigmoid colon are called transverse mesocolon and sigmoid mesocolon, respectively).

Arteries and veins that perfuse to the colon and intestine run inside the mesentery. Although the membranous structure of the mesentery is hard to be visualized on medical imaging such as X-ray computed tomography (CT) because of its limit of spatial resolution and contrast resolution, the arteries and veins that runs inside the mesentery is usually visualized on medical imaging. Although the shape of mesentery is not directly visualized on medical imaging, experienced medical doctors recognized the shape of the mesentery using these anatomical constrains in their images.

3 Importance of 3D Rendering of the Mesentery

In recent years, much interest is focused on minimally invasive surgeries with an advance in medical technology. Traditionally, a surgeon widely opens the abdomen for the operation of colon cancer. The surgeons directly see the organs and recognize the anatomical structures. However, this open surgery leaves large surgical scars after surgery.

The laparoscopic surgery is developed as one of minimally invasive surgeries. In contrast to traditional open surgery, it only requires small incision and leaves small scars (Fig. 3). The surgeons insert devices through the hole in order to perform surgical operations. As they observe the abdominal organs through the inserted camera (Fig. 3 laparoscope), their vision is limited. Moreover, as they perform operation indirectly using dedicated devices, they follow the procedure with low haptic feedback [1, 2]. Therefore the laparoscopic surgery requires much experience and training. If we obtain the whole shape of the mesentery before surgery, it helps planning for safe operation and it is useful to education for less-experienced surgeons.

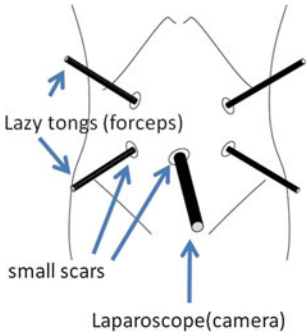


Fig. 3 Surgical apparatus configuration at a laparoscopic surgery

In this study, combining the anatomical features of the mesentery and data such as interrelation between the mesentery and the surrounding organs obtained with empirical laws of physicians, we propose technology to reproduce geometrically a shape of the mesentery, which is hard to reconstruct from the medical images.

4 Identification of Mesentery from Point Group

The idea of our work is very simple. If a veteran knows the position of the mesentery, we may ask him where it is in the medical images. We prepare a viewer of CT scan images. We ask a doctor to look at images and to make dots (or lines) at positions where the mesentery is. If we have big mass of coordinate data of points on the mesentery, we may determine it as a surface and we can render its image (Fig. 4).

It is not a brand new technology how to construct a surface in the 3D-space $\mathbf{R}^3$ from a family of discrete point data. For a given point group in the 3D-space, we can connect a point with its neighbors by edges. If it looks like a polyhedron, it is a rough position of the answer surface. This method sounds rather good, but if points are scattered randomly, we cannot suppose any surface at all. Though the mesentery is a thin tissue, it may not be easy to suppose the shape of the surface. Indeed, the mesentery may bend or may be folded complicatedly and we need more *structural data* in a point group.

Let $\{\mathbf{x}_i\}$ $(n = 1, \ldots, N)$ be the point group that a doctor gives. If we have a set of unit normal vectors $\{\mathbf{n}_i\}$ $(n = 1, \ldots, N)$ of the mesentery at each point, then it get easier to suppose a shape of it. That is, we want to find a function f such that the zero point set of f contains each x_i and that the gradient vector of f is parallel to n_i at x_i. Using Radial Basis Functions (RBF) method, we explain the way we obtain f from data $\{\mathbf{x}_i\}$ and $\{\mathbf{n}_i\}$. Here RBF method is a representation method of a surface interpolator. See [3, 4]. This method is also applied in visualization of curved thin slab. See [5],

We fix a small constant $\varepsilon > 0$. Let $\mathbf{x}_{N+i} = \mathbf{x}_i + \varepsilon\mathbf{n}_i$ and $\mathbf{x}_{2N+i} = \mathbf{x}_i - \varepsilon\mathbf{n}_i$ for $n = 1, 2, \ldots, N$. Let the coordinate of $\mathbf{x}_i$ be $(x_i \ y_i \ z_i)^T$ $(n = 1, 2, \ldots, 3N)$. Let

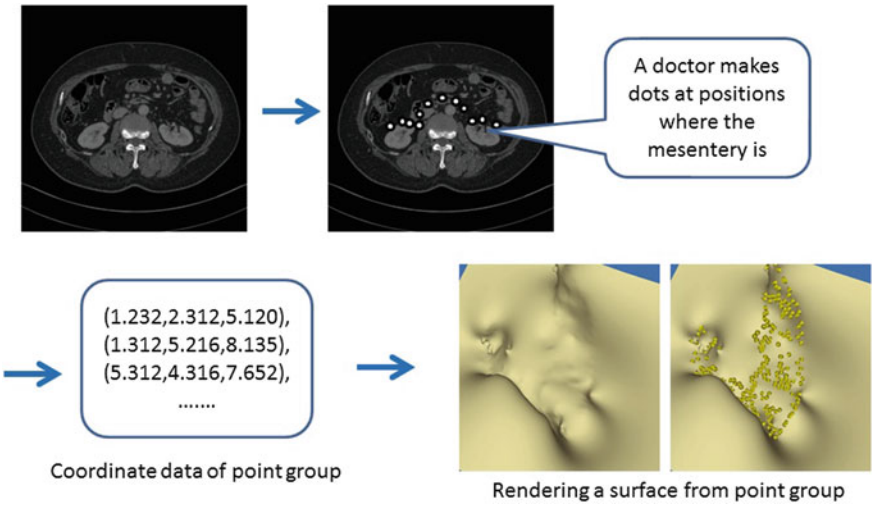


Fig. 4 A pipeline for the mesentery shape inference. Point group by manual input yields smooth surface

s_i be scalars given by $s_i = 0, s_{N+i} = 1, s_{2N+i} = -1$ for $n = 1, 2, \ldots, N$ (Fig. 5). Assume that a function $f(\mathbf{x}) = f(x, y, z)$ is given by the following:

$$f(\mathbf{x}) = \sum_{i=1}^{3N} \lambda_i \phi(|\mathbf{x} - \mathbf{x}_i|) + c_1 + c_2 x + c_3 y + c_4 z, \tag{1}$$

where λ_i's and c_j's are constants and we want to determine these constants such that $f(\mathbf{x}_i) = s_i$. The function $\phi(r)$ is called a basis function, and in our case we set $\phi(r) = r$. This is a system of a linear equations for λ_i's and c_j's, but the solution is not unique. We add conditions as follows. This additional condition guarantees that $f(\mathbf{x})$ belongs to the Beppo-Levi space on $\mathbf{R}^3$.

$$\sum_{i=1}^{3N} \lambda_i = \sum_{i=1}^{3N} \lambda_i x_i = \sum_{i=1}^{3N} \lambda_i y_i = \sum_{i=1}^{3N} \lambda_i z_i = 0. \tag{2}$$

Adding the system of the Eq. (2) to $f(\mathbf{x}_i) = s_i$, we have a unique solution for λ_i's and c_j's (in case that $\mathbf{x}_i$ are in general positions).

The remaining problem is how to obtain the unit normal vectors $\{\mathbf{n}_i\}$. Here we introduce a mathematical formulation of our work. We need transform anatomical knowledge of the mesentery into a mathematical structure of a surface. We do not need any advanced mathematics, but we have to know both of mathematics and anatomy. The local structure of the mesentery consists of two thin membranes, (that is, twofolded films pasted together). Here we assume that the mesentery is one

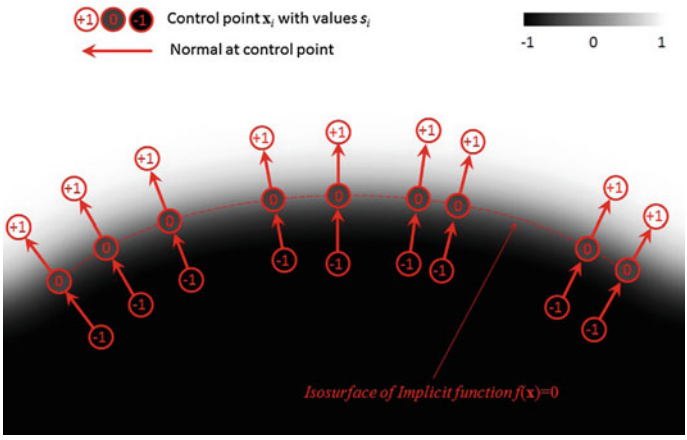


Fig. 5 Control point configuration for RBF reconstruction of implicit function f(x) representing a surface f(x) = 0. A point with a normal yields 3 control points for f(x) = +1, 0 and −1

smoothly embedded surface in 3D-space. This is the first assumption. We suppose that the mesentery is homeomorphic to a 2-disk. That is, the surface has no genus in it. Of course we cannot prove mathematically that the mesentery has no genus, but observing establishing process of the mesentery through one's childhood, we know that most of the mesentery is a simple disk. This is the second assumption. If the mesentery is a disk, its boundary is an embedded circle. Half of the boundary meets colon. This is the third assumption. The remaining part of the boundary meets arteries, a big blood vessel. This is the fourth assumption. There are some vessels which are put between two thin membranes and we can see these vessels in medical images. We may refer the shape of such vessels. (In the sequel we omit this assumption for simplicity.)

Figure 6 is a conceptual diagram of the mesentery, under the above conditions from 1 to 4. Of course the colon must be a tube-shaped organ and the arteries are 3D lines and the mesentery is 3D surface. This is a figure in which we simplify the situations.

Figure 7 is a sample of point groups. The outline polylines with dots are in the colon. From this image, we recognize that these points are decorated in a certain

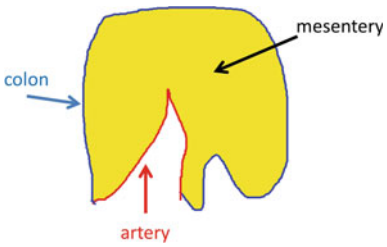


Fig. 6 Mesentery shape example with surrounding anatomical structures. The mesentery ends at the colon and the blood vessels

Fig. 7 A point group example. Several points in the same category are connected

surface in 3D-space. Point group is a set of dots and a medical specialist put dots on the CT scan images. We ask the doctor to classify these dots by the following criterion. It is easy for specialists make this classification. Type 0 points are points on the center line of the colon. From the assumption, these points are on the exterior boundary of the mesentery. Type 1 points are points near the colon. In the neighbors of the colon, there are some characteristic arteries named the marginal artery, and we assume that the points of this type is on the arteries. Type 2 and 3 points are on (and near) the main artery. (The superior mesenteric artery to the ileocolic artery, and the superior rectal artery to the inferior mesenteric artery.) These are on the interior boundary of the mesentery. Type 4 points are other points on the mesentery. See Fig. 8.

If we have the classification in the point data, we add edges (from a point to a point) with a geometric restriction which is derived from the anatomical criterion. Indeed, points of type 0 are arranged in a row (on a curve in the 3D-space) and we make a polyline. See Fig. 7. In the same way, points of type 1, 2, 3 are configured on

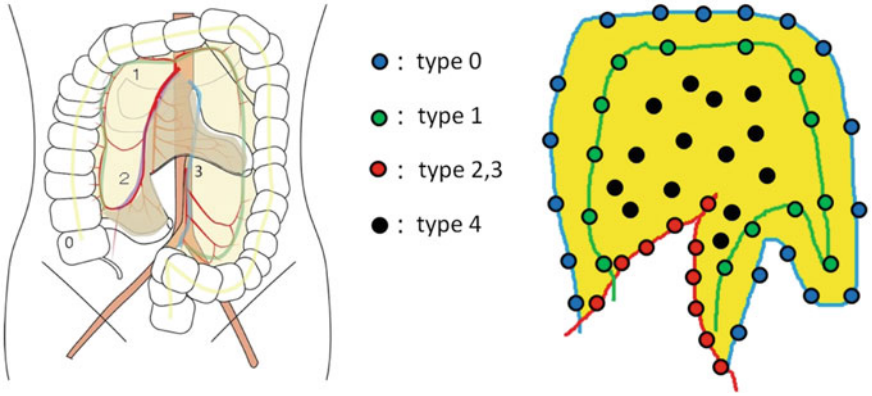


Fig. 8 Anatomical classification of point group into 5 categories

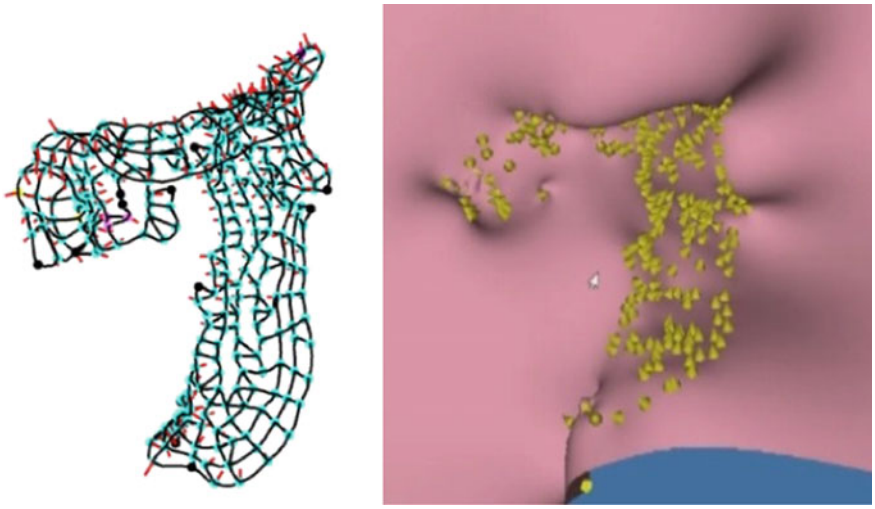


Fig. 9 Graph structure and inferred normals from point group (*left*), and reconstructed surface with control points and normals as cones (*right*)

a curve in the 3D-space respectively. Points of type 4 are distributed almost over a surface, and we connect points with their neighbors. In general we have a piecewise linear embedded surface in the 3D-space. Two points contained in different types may be near to each other. We connect these points with consideration for geometrical criterion. Using these estimated edges, we determine the normal vectors for each point. See Fig. 9.

5 Observation and Conclusion

Hereafter, we make observation about the method.

First, we explain difference between our approach and the one in [5]. In our method, there is no way of using the gradient of intensity in the medical images to determine the normal vectors of the surface. Because the mesentery is a membrane, we have to configure the surface from data of point group (or line group) using interpolation method. Moreover, the mesentery is homeomorphic to a disk with a boundary, we need another scheme in order to cut out the outside part of the mesentery.

It is interesting to discuss which basis function $\phi(r)$ is the best one in RBF method. There is no explicit data on this problem, but according to our observation there is less difference in resulting images if we use various basis functions. The precise observation about the relationship between the basis function and the resulting images is a future problem.

We have a problem what is a 'good' result. That is, it is a problem what is a measure which is provided in order to test the quality for the rendered surface. This is an important but difficult problem. Because goodness of the rendered image may not be precision of its shape but usefulness for surgeons. Under the circumstances visual assessment is one of the most realistic solution. This is visual estimation by doctors from the anatomical viewpoint. This is a future problem.

We have a mathematical problem of what is the minimum number of points in order to provide a good result. Generally if we have many points data on the mesentery, it is easy to interpolate the surface. This is an open problem.

In this paper, we do not provide a concrete method to estimate the normal vector field. There are some ways of estimation on the normal vectors from the data of neighborhood points. The authors are preparing another paper in which they discuss the estimation in the view of mathematics and information technology.

As seen in Fig. 8, there looks some arteries in the mesentery. If we get the position of these arteries automatically from the medical images, it is very helpful to determine the position of the mesentery. This also is a future problem.

We have other problems in interpolation. We assume that the points of type 0, 1, 2, and 3 are configured in a row, that is, a 3D curve. But indeed, the colon and the arteries may have complicated shapes and the point set of each type is not in a row. See an example in Fig. 10. In this situation we need an optimization scheme for identification of a space curve. After we obtain the normal vectors, it is not straightforward to determine the orientation of the surface of the mesentery. Here we have another orientation optimization problem.

In the viewpoint of user interface, we have other problems. We have to take care of a human error within the input data of the point group. We suppose that a doctor uses viewer software to draw dots on medical images. Clicking mouse button is influenced by hand trembling. We need to develop a perturbation optimization of identification of the surface. There is a problem of the density optimization of the point group. If the shape of the mesentery is complicated, for example, is folded or bends, then the density of points must be thick. But medical specialists do not know the proper density to determine a shape of the surface. Thus we need to estimate the density properness of point group. We are developing a system with user interface which allows medical specialists to know this properness of the point group.

In a plan for the future, we will make an investigation into the usefulness of this method in the cites of clinical medicine. Rendering and visualizing the structure of

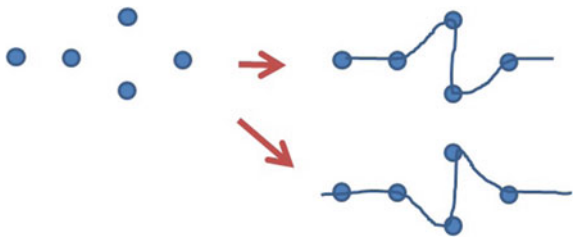


Fig. 10 Ambiguity in shape identification from point group. Four points may yield several shapes including these two

the mesentery from the data of preoperative medical images is expected in clinical medicine, in order to make a plan in preparation of operations for surgeons and to train themselves for the technique of surgery.

References

1. Emad, H.: Aly: Laparoscopic colorectal surgery, summary of the current evidence. Ann. R. Coll. Surg. Engl. **91**, 541–544 (2009)
2. Andreas, M.K.: Evolution and future of laparoscopic colorectal surgery. World J. Gastroenterol. **7; 20**(41), 15119–15124 (2014)
3. Carr, C.J., et al.: Reconstruction and Representation of 3D Objects with Radial Basis Functions, pp. 67–76. ACM SIGGRAPH2001 (2001)
4. Buhmann, M.D.: Radial Basis Functions. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, Cambridge (2003)
5. Masutani, Y.: RBF-based representation of volumetric data: application in visualization and segmentation. Proc. MICCAI **2002**, 300–307 (2002)

Using Process Algebra to Design Better Protocols

Peter Höfner

Abstract Protocol design, development and standardisation still follow the lines of *rough consensus and running code*. This approach yields fast and impressive results in a sense that protocols are actually implemented and shipped, but comes at a price: protocol specifications, which are mainly written in natural languages without presenting a formal specification, are (excessively) long, ambiguous, underspecified and erroneous. These shortcomings are neither new nor surprising, and well documented. It is the purpose of this paper to provide further evidence that formal methods in general and process algebras in particular can overcome these problems. They provide powerful tools that help to analyse and evaluate protocols, already during the design phase. To illustrate this claim, I report how a combination of pen-and-paper analysis, model checking and interactive theorem proving has helped to perform a formal analysis of the Ad hoc On-Demand Vector (AODV) routing protocol.

Keywords Process algebra · (Routing) Protocol · Wireless mesh network · Formal specification · Verification · AODV

1 The Need for Better Protocols

> Despite the maturity of formal description languages and formal methods for analysing them, the description of real protocols is still overwhelmingly informal. The consequences of informal protocol description drag down industrial productivity and impede research progress. Pamela Zave [45]

In computing, protocols are omnipresent: examples reach from *internet communication protocols*, such as the Simple Mail Transfer Protocol (SMTP) [22, 33] and the Transmission Control Protocol (TCP) [34], via *cryptographic protocols*, such Kerberos [27] and the MD5 Message-Digest Algorithm [37], and *routing protocols*, such the Border Gateway Protocol (BGP) [36] and the Ad hoc On-Demand

P. Höfner (✉)
NICTA and UNSW, Locked Bag 6016, UNSW, Sydney, NSW 1466, Australia
e-mail: peter.hoefner@nicta.com.au

B. Anderssen et al. (eds.), *The Role and Importance of Mathematics in Innovation*, Mathematics for Industry 25,
DOI 10.1007/978-981-10-0962-4_8

Vector (AODV) protocol [31], to *multimedia protocols*, such as the Session Initiation Protocol (SIP) [38].

Due to this omnipresence, protocols should satisfy a few important properties: (a) Specifications should be given in a way that they are easy to understand and implement. (b) Specifications have to support cross-vendor interaction, meaning if the same protocol is implemented by different vendors, these implementations should be compatible—different implementations of the same standard should be able to cooperate. (c) Finally, protocols should, of course, be correct.

Many of the protocols standardised today, however, fail to satisfy at least one of these properties. This is mainly because of the state of the art in protocol design and development.

Protocols are Broken

There is a stunning number of protocols that have been standardised, but do not work as expected.

For example, Miskovic and Knightly showed that many routing protocols based on the IEEE 802.11s standard [19], proprietary protocols such as those developed by Motorola,[1] Cisco[2] and others, as well as research routing protocols such as AODV-ST [35] and HOVER [25] are likely to establish non-optimal routes. This leads not only to an overhead in network traffic, but also to significant delays in packet delivery [26].

In [14] van Glabbeek et al. analysed AODV and proved that this routing protocol is not a priori loop free, i.e. data packets could be send through the network without ever reaching the intended destination. They argue that loop freedom hinges on non-evident assumptions to be made when resolving ambiguities occurring in the standard.

Zave showed that some disruptions in the ring structure of the Chord protocol cannot be repaired by the Chord ring-maintenance protocol as specified in [41][3] and [42]; hence the protocol is provable incorrect. In fact she stated that no published version of Chord is correct; however, "by selecting the right pseudocode from several papers, incorporating the right hints from the text of another paper, and fixing small flaws revealed by analysis, it is possible to come up with a 'best' version that may be correct" [46].

The Border Gateway Protocol (BGP), which is designed to exchange routing and reachability information between internet service providers (ISPs), is the last protocol to be mentioned. Varadhan, Govindan and Estrin showed that this protocol does not necessarily converge, and could persistently oscillate [43]. That means that nodes can change persistently the information about routes, although the network is assumed to be static.

[1] http://www.wi-fiplanet.com/news/article.php/3600221.

[2] http://www.mikrotik.com/.

[3] This paper won the 2011 SIGCOMM Test-of-Time Award.

Designing Protocols: State of The Art

As shown, many protocols do not behave as expected. The question that arises is why does this happen. Is not it possible to correctly specify a protocol and test/prove fundamental properties before implementation and deployment? This paper illustrates that this is possible. It is, however, my belief that the state of the art of designing protocols—*rough consensus and running code*—is one of the problems, if the process is implemented as described below.

"[IETF's] working groups make decisions through a 'rough consensus' process. IETF consensus does not require that all participants agree although this is, of course, preferred. In general, the dominant view of the working group shall prevail. (However, 'dominance' is not to be determined on the basis of volume or persistence, but rather a more general sense of agreement). Consensus can be determined by a show of hands, humming, or any other means on which the WG agrees (by rough consensus, of course). Note that 51 % of the working group does not qualify as 'rough consensus' and 99 % is better than rough. It is up to the Chair to determine if rough consensus has been reached" [6].

In practice this usually means that somebody first creates a draft of a specification in natural language, such as English. This draft often contains an excellent idea and deep insights on how to tackle a specific problem, e.g. using sequence numbers to ensure loop freedom. This draft is then discussed by the working group and changes are applied to the textual draft. As soon as rough consensus on the (natural language) specification is reached and as soon as there are at least two 'running' implementations, the protocol is declared to be standardised. Using this approach the IETF had major successes, such as the development and the deployment of DHCP (Dynamic Host Configuration Protocol), DNS (Domain Name System) and BGP.[4]

These successes suggest that the use of natural languages for protocol descriptions without presenting a formal specification seems to be advantageous: everybody can easily read, understand and comment on the specification, and hence, the protocol is easy to implement. However, looking at contemporary protocol developments more closely, it turns out that natural languages are no proper specification languages at all. They may be easy to understand, but this comes at a price.

- *Specifications are (excessively) long*. The description of the Session Initiation Protocol (SIP) [38], for example, is 268 pages long (and is not even self-contained); the IEEE Std 802.11™-2012 [20] standard, which contains a set of media access control (MAC) and physical layer (PHY) specifications for wireless networks, is 2,793 pages long.
  The sheer length of these specifications makes it nearly impossible to read and understand the full specification.
- *Specifications are ambiguous and underspecified*. It is hard—maybe impossible—to write precise and unambiguous specifications using natural languages only. Ambiguities in the Ad hoc On-Demand Vector (AODV) protocol [31], for

[4] A list of IETF's successes and failures can be found at http://trac.tools.ietf.org/misc/outcomes/.

example, yielded five open-source implementations to behave in incompatible ways, although all following the standard closely [14].

- *Protocols are neither formally analysed nor verified.* The lack of an (unambiguous) formal specification makes a formal analysis impossible. Traditional approaches to analyse protocols are simulation and test-bed experiments. While these are important and valid methods for protocol evaluation, they have limitations with regards to guaranteeing basic protocol correctness properties. Experimental evaluation is resource intensive and time-consuming, and, even after a very long time of evaluation, only a finite set of scenarios can be considered—usually, no general guarantee can be given. This problem is illustrated by Miskovic's and Knightly's discovery of limitations in AODV-like protocols (see above) that have been under intense scrutiny over many years [26].

Better Protocols are Needed, Now!

These shortcomings are neither new nor surprising, and documented in several research papers, e.g. [45] or [39, Chap. 9]. I believe that many problems could be avoided if formal protocol descriptions would accompany the textual specification, already in the design phase, before rough consensus is reached. By this, different readings of the draft, or underspecification can easily be avoided. Another reason why formal methods should be used already during the design phase is that protocols are not deployed in a lab: as soon as protocols are shipped, deployed and in (regular) use, it is nearly impossible to replace them. A classic example is BGP, which is erroneous (see above), but runs at the backbone of the Internet since 1994.

It is the purpose of the remainder of this paper to provide further evidence that formal methods in general and process algebras in particular can overcome these problems. Formal methods are mathematical approaches used to formally reason about software and hardware systems. They are used from formalising systems' requirements, specifications and designs, through programming concepts and programming languages, to implementation. They are also used to relate different formalisations: for example, refinement can be used to show that an implementation 'follows' a formal specification. Formal methods are indispensable for software and protocol engineering, especially when safety, security or correctness is considered.

In the area of protocol development they provide powerful tools that help to analyse and evaluate protocols, already during the design phase. I will illustrate this by a formal analysis of AODV [31], a routing protocol currently standardised by the IETF MANET working group. I will report how a combination of pen-and-paper analysis, model checking and interactive theorem proving has helped to carry out the analysis. This case study shows (again) that formal methods are mature enough to support protocol design from the beginning. It is my belief that the use of formal methods could have found and prevented limitations in AODV-like protocols as reported in [26].

2 Formal Specification Languages

Formal specification languages and analysis techniques are now able to capture the full syntax and semantics of reasonably rich protocols. They are an indispensable augmentation to natural language, both for specification and analysis.

Even when formal analysis is not the final aim, the use of formal languages is useful: they are unambiguous, reduce significantly the number of misunderstandings, and clarify the overall structure. By this, they almost always avoid underspecification. Obviously, formal specification languages cannot prevent errors a priori, but they will make them less likely, and since they are unambiguous they do not allow different readings of a draft or a standard. If no formal analysis is required, it does not really matter which formalism is used. The choice of formal specification languages is numerous: it ranges from timed automata, which offer tool support by model checking (e.g. [8]), via the inductive approach, which offers interactive theorem proving support [30], to algebraic characterisations such as semirings (e.g. [15]) and process algebra (e.g. [10]). For our case study (see below), process algebra was chosen as specification language. It has the advantage that it is closely related to programming languages, and hence specifications are easy to understand by network researchers and software engineers as well, not only by theoreticians.

The Process Algebra AWN

The process algebra AWN (*Algebra for Wireless Networks*) [10] was initially developed for wireless networks such as AODV, and has therefore in-built support for node mobility, broadcast/multicast communication etc. However, AWN allows modelling any type of communicating concurrent processes, and can be used for a wide range of networks and protocols, e.g. [12].

The syntax of the AWN language, depicted in Table 1 and described below, is simple and reads much like a programming language, but it is implementation independent and has all the required ingredients to be able to formally reason about protocol and network properties, and to provide mathematically rigorous proofs.

AWN is a variant of standard process algebras [2, 3, 17, 24], extended with a *local broadcast* mechanism and a *conditional unicast* operator—allowing error handling in response to failed communications—and incorporating *data structures*.

In AWN, a protocol running in a (wireless) network is modelled as parallel composition of network nodes. On each node several processes may run in parallel. Network nodes communicate with their direct neighbours—those nodes that are currently in transmission range—using either broadcast, unicast, or an iterative unicast/multicast (called *groupcast*).

The basic components of *process expressions* are given in the first part of Table 1. A process name X comes with a *defining equation* $X(\mathtt{var}_1, \ldots, \mathtt{var}_n) \stackrel{def}{=} p$, where p is a process expression, and the $\mathtt{var}_i$ are data variables maintained by process X. A named process is similar to a procedure or a function: if it is called, data expressions exp_i are filled in for the variables $\mathtt{var}_i$. The process $p + q$ models choice: it may act either as p or as q, depending on which of the two is able to act at all. In a context

Table 1 Syntax of the process algebra AWN

Basic primitives of (node level) sequential process expressions	
$X(exp_1, \ldots, exp_n)$	Process name with arguments
$p + q$	Choice between processes p and q
$[\varphi]p$	Conditional process
$[\![\text{var} := exp]\!]p$	Assignment followed by process p
broadcast$(ms).p$	Broadcast ms followed by p
groupcast$(dests, ms).p$	Iterative unicast or multicast to all destinations $dests$
unicast$(dest, ms).p \blacktriangleright q$	Unicast ms to $dest$; if successful proceed with p; otherwise with q
send$(ms).p$	Synchronously transmit ms to parallel process on same node
receive(msg).p	Receive a message
deliver$(d).p$	Deliver d to application layer
Some advanced sequential process expressions	
$[\varphi]p + [\neg\varphi]q$	Deterministic choice with test
$X(n) \stackrel{def}{=} [\![n := n + 1]\!]\, X(n)$	Example of a loop
Parallel process expressions	
ξ, p	Process with valuation
$P \between Q$	Parallel processes on the same node

where both are able to act, a nondeterministic decision is made. The expression $[\varphi]p$ is a conditional process—an if-statement—if the Boolean expression φ evaluates to `true` then the process acts like p, it deadlocks otherwise.[5]

The process algebra also features (arbitrary) data structures. An update to a variable `var` is performed using the assignment $\texttt{var} := exp$, where exp is a data expression of the same type as `var`. The process $[\![\texttt{var} := exp]\!]p$ acts as p, but under the constraint that the value of the variable `var` is now exp.

AWN always provides data types for node identifiers, sets of node identifiers, and messages; variables of these types are used to model the transmission of messages, and are denoted by $dest$, $dests$ and ms, respectively. The process **broadcast**$(ms).p$ broadcasts (the data value bound to the expression) ms to all nodes in the network within transmission range of the sender, and subsequently acts as p; the process **groupcast**$(dests, ms).p$ transmits ms to all destinations within transmission range that are also listed in the set $dests$, and proceeds as p. Both expressions **broadcast** and **groupcast** continue as p, independently whether the transmission (to some nodes) was successful. In contrast to this, **unicast**$(dest, ms).p \blacktriangleright q$ tries to send the message ms to the sole destination dest; if successful it continues to act as p and otherwise

[5] In case φ contains free variables, values to these variables are chosen nondeterministically in a way that satisfies φ, if possible.

as q.[6] It models an abstraction of an acknowledgment-of-receipt mechanism that is typical for unicast communication but absent in broadcast communication, and implemented in wireless standards such as IEEE 802.11. All these mechanisms model internode message sending; for intranode communication the process **send**$(ms).p$ is used. This action can only take place if another process is able to receive the message.

The process **receive**(msg).p is able to receive any message m; it then proceeds as p under the constraint that the variable msg is updated to m. The message m can stem from another node (**broadcast/groupcast/unicast**), from the same node (**send**), or from an application layer process. The latter is modelled by the process **receive**(newpkt$(d, dip)).p$, where newpkt generates a message containing data d to be sent from the application layer, and the intended destination address dip. Data is delivered to the application layer by the process **deliver**$(d).p$.

It is straightforward to model a protocol (running on one node) using these basic process expressions. I will show a snippet of the AODV protocol in the next section. Other well-known programming constructs, such as if-then-else or loops can easily built from them; two examples are given in the second block of Table 1.

Processes running on the same node, can be combined as $P \between Q$. Here, P and Q are valuated processes, meaning that they are either a process expression p built from the syntax presented above and equipped with a valuation function ξ, which specifies values of variables maintained by p, or a parallel process itself.

In the full process algebra [10], parallel processes (processes describing the behaviour of a single network node) are combined into an expression modelling the entire network, including information about the transmission ranges of all nodes. Since we concentrate on modelling aspects, these details do not matter—the presented constructs are sufficient to describe protocols on the level of nodes.

The intuition of the syntax of AWN should be clear for anybody writing protocol specifications. However, to formally reason about protocols a formal semantics is required: AWN, as many other process algebras, is equipped with an operational semantics [10]. It describes a model's behaviour in terms of its execution. As a consequence, many desired properties, such as correctness and safety can often be verified by constructing proofs from these logical statements. An example that formally describes intranode communication is given by the rule

$$\frac{P \xrightarrow{\textbf{receive}(msg)} P' \quad Q \xrightarrow{\textbf{send}(msg)} Q'}{P \between Q \xrightarrow{\tau} P' \between Q'}.$$

This semantical rule states that if the process Q is able to **send** a message msg and, at the same time, process P is able to **receive** the same message, then both processes will execute their actions (**send/receive**); the resulting internal action is called τ.

[6] The unicast is unsuccessful if the destination *dest* is out of transmission range of the sender.

The main purpose of this paper is to illustrate that process algebras can be used to model and analyse reasonably rich protocols. Hence we abstain from a detailed presentation of the operational semantics.

3 Case Study: The AODV Routing Protocol

Together with my colleagues R. van Glabbeek, M. Portmann, W.L. Tan, A. McIver and A. Fehnker, I used the process algebra AWN to obtain the first rigorous formalisation of the specification of Ad hoc On-Demand Vector (AODV) routing protocol [31]. Based on the formalisation, a careful analysis of the protocol was performed, using pen-and-paper analysis in [10, 13], the proof assistant Isabelle/HOL [28] in [4, 5], and the model checker Uppaal [1, 23] in [9, 18].

The Protocol

AODV is a reactive protocol, meaning that routes are established on demand when needed. A route from a source node s to a destination node d is a sequence of nodes $[s, n_1, \ldots, n_k, d]$, where $n_1, \ldots, n_k$ are intermediate nodes located on a particular path from s to d. The intuition of AODV is best illustrated by a small example, depicted in Fig. 1. The network topology is given in Fig. 1a, where an edge between two nodes indicate that the nodes are within transmission range. In the example node s tries to send data packets to node d, but s does not yet have information about a route to d.

Node s initiates a route discovery mechanism by broadcasting a route request (RREQ) message, which is received by all neighbours in transmission range, nodes a and b in the example. Nodes receiving a RREQ message that do not know a route to the intended destination d rebroadcast the message. By this the RREQ message floods the network (cf. Fig. 1b). If one of the intermediate nodes has established a route to d before, it directly sends a route reply back towards the originator s. When a node forwards a RREQ message, it updates its routing table and adds a 'reverse route' entry to s, indicating via which next hop the node s can be reached, and other information about the route.

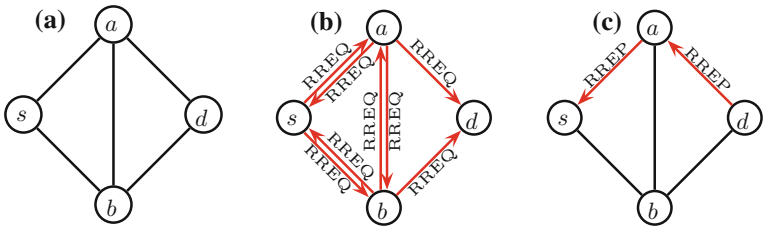


Fig. 1 AODV by example [13]. **a** Network topology. **b** Request floods the network. **c** Route reply is sent back

```
RREQ(hops, rreqid, dip, dsn, dsk, oip, osn, sip, ip, sn, rt, rreqs, store) ≝
  ...
7.  [ dip = ip ]    /* this node is the destination node */
      ...
9.    /* unicast a RREP towards oip of the RREQ */
10.   unicast(nhop(rt, oip), rrep(0, dip, sn, oip, ip)) . AODV(ip, sn, rt, rreqs, store)
11.   ▶ /* If the transmission is unsuccessful, a RERR message is generated */
        ...
18. + [ dip ≠ ip ]    /* this node is not the destination node */
19.   (
20.     [dip ∈ vD(rt) ∧ dsn ≤ sqn(rt,dip) ∧ sqnf(rt,dip) = kno ]    /* valid route to dip that is fresh enough */
          ...
24.       /* unicast a RREP towards the oip of the RREQ */
25.       unicast(nhop(rt, oip), rrep(dhops(rt, dip), dip, sqn(rt, dip), oip, ip)) .
26.         AODV(ip, sn, rt, rreqs, store)
27.       ▶ /* If the transmission is unsuccessful, a RERR message is generated */
            ...
34.     + [ dip ∉ vD(rt) ∨ sqn(rt,dip) < dsn ∨ sqnf(rt,dip) = unk ]    /* no valid route that is fresh enough */
35.       /* forward RREQ message as broadcast */
36.       broadcast(rreq(hops + 1, rreqid, dip, max(sqn(rt, dip), dsn), dsk, oip, osn, ip)) .
37.       AODV(ip, sn, rt, rreqs, store)
38.   )
```

Fig. 2 Excerpt of AWN specification for AODV: cases for handling a RREQ message [13]

Once the first RREQ message is received by the destination node d—we assume it stems from a—the destination node also adds a reverse route entry in its routing table, indicating that node s can be reached via node a. It then responds by sending a route reply (RREP) message addressed to node s to node a. In contrast to RREQ messages, RREP messages are unicast. Node a receives the RREP message; it creates a 'forward route' entry to d in its routing table and forwards the message to the next hop along the established reverse route. The RREP message will finally reach the originator of the RREQ message, and the route discovery process is completed and a route from s to d has been established—data packets can start to flow.

In the event of link failures, AODV uses route error (RERR) messages to inform affected nodes.

Formal Analysis

In contrast to the ambiguous de facto standard specification of AODV [31], which is written in English prose and about 35 pages long, the created AWN model is precise, yet very readable and consists only of roughly 200 lines. The model reflects precisely the intention of AODV and accurately captures all core aspects of the protocol specification, excluding all aspects of time. An excerpt, which shows the essential parts for handling a RREQ message, is given in Fig. 2. The full specification as well as a detailed explanation can be found in [13]. As the semantics of AWN is completely unambiguous, specifying a protocol in such a framework enforces total precision and obviously removes any ambiguity.

An analysis of this specification revealed that under a plausible interpretation of the original specification of AODV,[7] the protocol admits routing loops [14]; this is in direct contradiction with popular belief, the promises of the AODV standard, and

[7] As common, text placed between /* and */ are comments and not part of AWN.

the main paper on AODV [32] (with over 12,000 citations). However, we showed loop freedom of AODV under a subtly different interpretation of the original specification [13]. Our analysis, which I will report on in the remainder of this section, considered also route correctness, packet delivery, route optimality and other properties of the routing protocol. It has been carried out by pen-and-paper, with the proof assistant Isabelle/HOL [28], and with the model checker Uppaal [1, 23].

- Using the formal semantics of AWN, we verified properties of AODV that can be expressed as invariants by *pen-and-paper*. Invariants are statements that hold at all times when the protocol is executed. The most important invariants were route correctness and loop freedom.
  The term *route correctness* means that all routing table entries stored at a node are entirely based on information on routes to other nodes that is currently valid or was valid at some point in the past. In case of AODV, this property is not hard to prove, but already shows the power of formal methods, since a formal proof can be provided [13].
  Loop freedom is a critical property for any routing protocol, but it is particularly relevant and challenging for wireless networks, since the underlying network topology can change constantly. Garcia-Luna-Aceves describes a loop as follows: "A routing-table loop is a path specified in the nodes' routing tables at a particular point in time that visits the same node more than once before reaching the intended destination" [11]. Packets caught in a routing loop can quickly saturate the links and can decrease the overall network performance. To the best of our knowledge we are the first to give a complete and detailed proof of loop freedom [13]. The proof of loop freedom builds on another 30 invariants that needed to be proven before loop freedom could be verified.
- Providing a pen-and-paper proof of loop freedom was a major step in the understanding of AODV, but the proof itself is about 20 pages long. To add credibility and confidence we mechanised the proof in the theorem prover Isabelle/HOL [4, 5].
  Isabelle [29] is a generic *interactive theorem prover* based on a small logical core to ease logical correctness. The main application area is the formalisation of mathematical proofs and in particular formal verification. The most widespread instance of Isabelle nowadays is Isabelle/HOL [28], which provides a higher-order logic (HOL) theorem proving environment that is ready to use for big applications. Examples for such applications are the projects *L4.verified* and *Flyspec*. L4.verified used Isabelle/HOL to prove formal functional correctness of the seL4 microkernel, a small, 3rd generation high-performance microkernel with about 8,700 lines of C code [21]. Flyspec derived a formal proof of the Kepler conjecture on dense sphere packings using the Isabelle/HOL and HOL Light proof assistants [16].
  While the hand-written process-algebraic proof of loop freedom of AODV was already very formal, the implication that transfers statements about nodes to statements about networks involves coarser reasoning over execution sequences. The

mechanised proof clarifies this aspect by explicitly stating the assumptions made of other nodes. It consists of about 400 lemmas.

Besides the added confidence that comes with having even the smallest details fastidiously checked by a machine, the real advantage in encoding model, proof, and framework in an interactive theorem prover is that they can then be analysed and manipulated (semi-)automatically.

In [4] we showed how protocol variants, such as different readings of the textual standard or proposed improvements of the standard can quickly be analysed. Variants often only differ in minor details, most proofs stay the same or can be adapted automatically: an interactive theorem prover tries to 'replay' the original proof and, in case of a failure, it points at all proof steps that are no longer valid.[8] One only has to concentrate on these failures. This avoids the tedious, time-consuming, and error-prone manual chore of establishing which steps remain valid for each invariant, especially for long proofs.

- *Model checking* is in particular useful to discover protocol limitations and to develop improved variants; in our setting it can be seen as a diagnostic tool that complements the other verification techniques. Model checking is limited to networks of small size—due to state space explosion—and hence cannot verify properties for all networks, in contrast to the invariant proofs mentioned above that cover all topology.

  Based on our AWN specification we developed a model of AODV for the Uppaal model checker [9]. We checked important properties, such as route correctness and route optimality, against all topologies of up to 5 nodes, which also included dynamic topologies with one link going up or down. In the case a property does not hold, Uppaal produces evidence for the fault in the form of a 'counter-example' summarising the circumstances leading to it. Such diagnostic information provides insights into the cause and correction of these failures. For some problematic and undesired behaviour of AODV, automatically found by Uppaal, we provided fixes in form of improvements of AODV, which then were (semi-)automatically verified by Isabelle/HOL.

  Analysing small topologies often yields new insights, as does simulation, but the network sizes are far from realistic and quantitative information is not included. *Statistical model checking* [40, 44] can combine the systematic methodology of 'classical' model checking with the ability to analyse quantitative properties and realistic scenarios.[9] Using statistical model checking, we showed that quantitative reasoning is now feasible—for example we analysed the extent of establishing non-optimal routes—and illustrated that properties can be checked for networks of up to 100 nodes—of course an exhaustive search is not possible here.

[8] Reference [4] proves loop freedom for four variants of AODV, in average only one invariant needed major changes; and a few others needed systematic adaptions, such as changes of data types.

[9] SMC-Uppaal, the Statistical extension of Uppaal (release 4.1.11) [7] accepts the same input as standard Uppaal; the creation of a new model was not required.

4 Looking Ahead

In this paper, I have illustrated that formal specification languages and analysis techniques are now able to capture the full syntax and semantics of reasonably rich protocols. The use of formal methods in general and the use of process algebras in particular can be split into three layers (cf. Fig. 3): (a) the (syntax of the) formal description language, (b) its semantics, and (c) tools for analysing a protocol, based on the syntax and the semantics. Although it would be favourable if everybody would understand all three layers, this is wishful thinking: most likely only trained experts working in the area of formal methods will understand the full spectrum. But, for specifying a protocol in a precise and unambiguous manner, which also avoids underspecification, this is not necessary. To achieve this goal, only the syntax together with a good intuition about its semantics is required—neither a full understanding of the formal semantics nor of the formal analysis tools is needed.

I believe that state-of-the-art formal description languages are simple enough to be used by any network researcher and software engineer. These languages can be used to specify and analyse rather complicated protocols. To achieve more automation in the analysis, they often offer tool support, such as model checking.

So, the question remains why despite of the maturity of formal description languages and formal methods for analysing them, the description of real protocols is still overwhelmingly informal. As Zave pointed out, this drags down industrial productivity and impedes research progress [46]. It is my belief that three ingredients are still missing: (1) *Better (easy to use) tool support*: better tools and faster computers allow more and more automation. However, the use of tools often requires special knowledge (how to use the tool) or a special input format (e.g. timed automata). (2) *Code generation*: it is often believed that the combination of formal specification

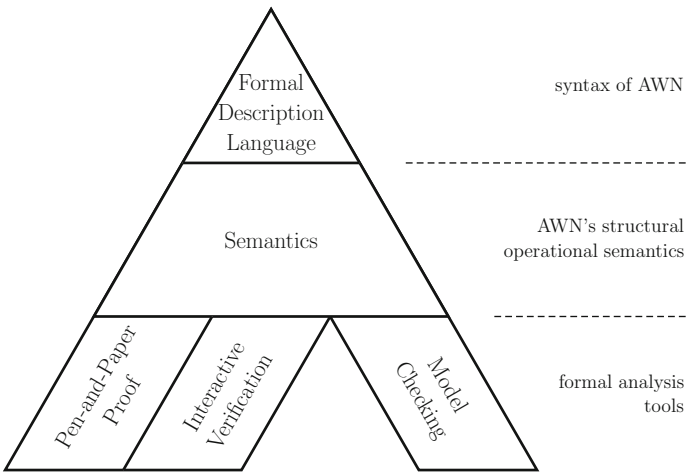


Fig. 3 Different layers of Formal Methods

followed by implementation requires more time (and hence more money) than just implementing the protocol straight away. If entire (or at least parts of) implementations could be generated out of formal specifications automatically, one could gain even more advantages from formal methods. (3) *Training*: to use formal methods, engineers working in industry must be aware of them; this can only be achieved by training. Current research tackles the first two items, the last one may be the hardest to achieve.

Acknowledgments Special thanks goes to all collaborators who contributed to the AODV case study; in particular Timothy Bourke, Ansgar Fehnker, Robert J. van Glabbeek, Annabelle McIver, Marius Portmann, and Wee Lum Tan. Further I would like to thank Robert J. van Glabbeek again for valuable comments on this paper. NICTA is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program.

References

1. Behrmann, G., David, A., Larsen, K.G.: A Tutorial on UPPAAL. In: Bernardo, M., Corradini, F. (eds.) Formal Methods for the Design of Real-Time Systems, *Lecture Notes in Computer Science*, vol. 3185, pp. 200–236. Springer, Berlin (2004)
2. Bergstra, J.A., Klop, J.W.: Algebra of communicating processes. In: de Bakker, J.W., Hazewinkel, M., Lenstra, J.K. (eds.) Mathematics and Computer Science, CWI Monograph 1, pp. 89–138. North-Holland (1986)
3. Bolognesi, T., Brinksma, E.: Introduction to the ISO specification language LOTOS. Comput. Netw. **14**, 25–59 (1987). doi:10.1016/0169-7552(87)90085-7
4. Bourke, T., van Glabbeek, R.J., Höfner, P.: A mechanized proof of loop freedom of the (untimed) AODV routing protocol. In: Cassez, F., Raskin, J.F. (eds.) Automated Technology for Verification and Analysis (ATVA'14), Lecture Notes in Computer Science, vol. 8837, pp. 47–63. Springer, Berlin (2014). doi:10.1007/978-3-319-11936-6_5
5. Bourke, T., van Glabbeek, R.J., Höfner, P.: Mechanizing a process algebra for network protocols. J. Autom. Reason. **56**(3), 309–341 (2016). doi:10.1007/s10817-015-9358-9. (in press)
6. Bradner, S. (ed.): IETF working group guidelines and procedures. RFC 2418 (Best Current Practice) (1998). https://tools.ietf.org/html/rfc2418
7. Bulychev, P., David, A., Larsen, K., Mikučionis, M., Bøgsted P., D., Legay, A., Wang, Z.: UPPAAL-SMC: Statistical model checking for priced timed automata. In: Wiklicky, H., Massink, M. (eds.) Quantitative Aspects of Programming Languages and Systems, *EPTCS*, vol. 85, pp. 1–16. Open Publishing Association (2012)
8. Chiyangwa, S., Kwiatkowska, M.: A timing analysis of AODV. In: Formal Methods for Open Object-based Distributed Systems (FMOODS'05), Lecture Notes in Computer Science, vol. 3535, pp. 306–322. Springer, Berlin (2005). doi:10.1007/11494881_20
9. Fehnker, A., van Glabbeek, R.J., Höfner, P., McIver, A.K., Portmann, M., Tan, W.L.: Automated analysis of AODV using UPPAAL. In: Flanagan, C., König, B. (eds.) Tools and Algorithms for the Construction and Analysis of Systems (TACAS '12), Lecture Notes in Computer Science, vol. 7214, pp. 173–187. Springer, Berlin (2012). doi:10.1007/978-3-642-28756-5_13
10. Fehnker, A., van Glabbeek, R.J., Höfner, P., McIver, A.K., Portmann, M., Tan, W.L.: A process algebra for wireless mesh networks. In: H. Seidl (ed.) European Symposium on Programming (ESOP '12), Lecture Notes in Computer Science, vol. 7211, pp. 295–315. Springer, Berlin (2012). doi:10.1007/978-3-642-28869-2_15

11. Garcia-Luna-Aceves, J.J.: A unified approach to loop-free routing using distance vectors or link states. In: Proceedings of the Symposium on Communications, Architectures & Protocols (SIGCOMM '89), ACM SIGCOMM Computer Communication Review, vol. 19(4), pp. 212–223. ACM (1989). doi:10.1145/75246.75268
12. van Glabbeek, R.J., Höfner, P.: SMACCM report: Formal specification of protocols for internal high-assurance network (2015)
13. van Glabbeek, R.J., Höfner, P., Portmann, M., Tan, W.L.: Modelling and verifying the aodv routing protocol. Distributed Computing (2016). (in press)
14. van Glabbeek, R.J., Höfner, P., Tan, W.L., Portmann, M.: Sequence numbers do not guarantee loop freedom—AODV can yield routing loops—. In: Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM '13), pp. 91–100. ACM, New York (2013). doi:10.1145/2507924.2507943
15. Griffin, T.G., Sobrinho, J.: Metarouting. SIGCOMM. Comput. Commun. Rev. **35**(4), 1–12 (2005). doi:10.1145/1090191.1080094
16. Hales, T.C., Adams, M., Bauer, G., Dang, D.T., Harrison, J., Le Hoang, T., Kaliszyk, C., Magron, V., McLaughlin, S., Nguyen, T.T., Nguyen, T.Q., Nipkow, T., Obua, S., Pleso, J., J., R., Solovyev, A., Ta, A.H.T., Tra, T.N., Trieu, D.T., Urban, J., Vu, K.K., Zumkeller, R.: A formal proof of the Kepler conjecture. CoRR (2015). http://arxiv.org/abs/1501.02155
17. Hoare, C.A.R.: Communicating Sequential Processes. Prentice Hall, Englewood Cliffs (1985)
18. Höfner, P., McIver, A.: Statistical model checking of wireless mesh routing protocols. In: Brat, G., Rungta, N., Venet, A. (eds.) NASA Formal Methods Symposium (NFM '13), Lecture Notes in Computer Science, vol. 7871, pp. 322–336. Springer, Berlin (2013). doi:10.1007/978-3-642-38088-4_22
19. IEEE: IEEE Standard for Information Technology—Telecommunications and information exchange between systems—Local and metropolitan area networks—Specific requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications Amendment 10: Mesh Networking (2011). http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6018236
20. IEEE: IEEE Standard for Information Technology—Telecommunications and information exchange between systems—Local and metropolitan area networks—Specific requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications (2011). (Revision of IEEE Std 802.11-2007)
21. Klein, G., Andronick, J., Elphinstone, K., Heiser, G., Cock, D., Derrin, P., Elkaduwe, D., Engelhardt, K., Kolanski, R., Norrish, M., Sewell, T., Tuch, H., Winwood, S.: seL4: Formal verification of an operating-system kernel. Commun. ACM **53**(6), 107–115 (2010). doi:10.1145/1743546.1743574
22. Klensin, J.: Simple mail transfer protocol. RFC 5321 (Draft Standard), Network Working Group (2008). https://tools.ietf.org/html/rfc5321
23. Larsen, K.G., Pettersson, P., Yi, W.: UPPAAL in a nutshell. Int. J. Softw. Tools Technol. Transf. **1**(1–2), 134–152 (1997)
24. Milner, R.: Communication and Concurrency. Prentice Hall, Upper Saddle River (1989)
25. Mir, S., Pirzada, A.A., Portmann, M.: HOVER: hybrid on-demand distance vector routing for wireless mesh networks. In: Proceedings of the Australasian Conference on Computer Science (ACSC'08), ACSC '08, pp. 63–71. Australian Computer Society, Inc. (2008)
26. Miskovic, S., Knightly, E.W.: Routing primitives for wireless mesh networks: Design, analysis and experiments. In: Proceedings of the Conference on Information Communications (INFOCOM '10), pp. 2793–2801. IEEE (2010). doi:10.1109/INFCOM.2010.5462111
27. Neuman, C., Yu, T., Hartman, S., Raeburn, K.: The Kerberos network authentication service (v5). RFC 4120 (Standards Track) (2005). http://tools.ietf.org/html/rfc4120
28. Nipkow, T., Paulson, L.C., Wenzel, M.: Isabelle/HOL: A Proof Assistant for Higher-Order Logic, Lecture Notes in Computer Science, vol. 2283. Springer, Berlin (2002)
29. Paulson, L.C.: The foundation of a generic theorem prover. J. Autom. Reason. **5**(3), 363–397 (1989). doi:10.1007/BF00248324

30. Paulson, L.C.: The inductive approach to verifying cryptographic protocols. Comput. Secur. **6**(1–2), 85–128 (1998)
31. Perkins, C.E., Belding-Royer, E.M., Das, S.: Ad hoc on-demand distance vector (AODV) routing. RFC 3561 (Experimental), Network Working Group (2003). https://tools.ietf.org/html/rfc3561
32. Perkins, C.E., Royer, E.M.: Ad-hoc On-Demand Distance Vector Routing. In: Mobile Computing Systems and Applications (WMCSA '99), pp. 90–100. IEEE (1999). doi:10.1109/MCSA.1999.749281
33. Postel, J.B.: Simple mail transfer protocol. RFC 821 (Internet Standard) (1982). https://tools.ietf.org/html/rfc821
34. Postel, J.B. (ed.): Transmission control protocol. RFC 793 (Internet Standard) (1981). https://tools.ietf.org/html/rfc793
35. Ramachandran, K., Buddhikot, M., Chandranmenon, G., Miller, S., Belding-Royer, E.M., Almeroth, K.: On the design and implementation of infrastructure mesh networks. In: Proceedings of the IEEE Workshop on Wireless Mesh Networks (WiMesh'05)). IEEE Press (2005)
36. Rekhter, Y., Li, T., Hares, S.: A border gateway protocol 4 (BGP-4). RFC 4271 (Draft Standard), Network Working Group (Errata Exist) (2006). https://tools.ietf.org/html/rfc4271
37. Rivest, R.: The MD5 Message-Digest Algorithm. RFC 1321 (Informational, Errata Exist) (1992). http://tools.ietf.org/html/rfc1321
38. Rosenberg, J., Schulzrinne, H., Camarillo, G., Johnston, A., Peterson, J., Sparks, R., Handley, M., Schooler, E.: SIP: Session initiation protocol. RFC 4728 (Proposed Standard), Network Working Group (Errata Exist) (2002). https://tools.ietf.org/html/rfc3261
39. Ryan, P., Schneider, S., Goldsmith, M., Lowe, G., Roscoe, A.: The Modelling and Analysis of Security Protocols: The CSP Approach, (first published 2000) edn. Pearson Education (2010)
40. Sen, K., Viswanathan, M., Agha, G.A.: Vesta: A statistical model-checker and analyzer for probabilistic systems. In: Quantitative Evaluaiton of Systems (QEST'05), pp. 251–252. IEEE (2005)
41. Stoica, I., Morris, R., Karger, D., Kaashoek, M.F., Balakrishnan, H.: Chord: A scalable peer-to-peer lookup service for internet applications. SIGCOMM Comput. Commun. Rev. **31**(4), 149–160 (2001). doi:10.1145/964723.383071
42. Stoica, I., Morris, R., Liben-Nowell, D., Karger, D.R., Kaashoek, M.F., Dabek, F., Balakrishnan, H.: Chord: A scalable peer-to-peer lookup protocol for internet applications. IEEE/ACM Trans. Netw. **11**(1), 17–32 (2003). doi:10.1109/TNET.2002.808407
43. Varadhan, K., Govindan, R., Estrin, D.: Persistent route oscillations in inter-domain routing. Comput. Netw. **32**(1), 1–16 (2000). doi:10.1016/S1389-1286(99)00108-5
44. Younes, H.: Verification and planning for stochastic processes with asynchronous events. Ph.D. thesis, Carnegie Mellon University (2004)
45. Zave, P.: Experiences with protocol description. In: Rigorous Protocol Engineering (WRiPE' 11) (2011)
46. Zave, P.: Using lightweight modeling to understand Chord. SIGCOMM Comput. Commun. Rev. **42**(2), 49–57 (2012). doi:10.1145/2185376.2185383

Relational Hash

Avradip Mandal and Arnab Roy

Abstract Traditional cryptographic hash functions allow one to easily check whether the original plaintexts are equal or not, given a pair of hash values. Probabilistic hash functions extend this concept where given a probabilistic hash of a value and the value itself, one can efficiently check whether the hash corresponds to the given value. However, given distinct probabilistic hashes of the same value it is not possible to check whether they correspond to the same value. In this work we introduce a new cryptographic primitive called *Relational Hash* using which, given a pair of (relational) hash values, one can determine whether the original plaintexts were related or not. We formalize various natural security notions for the Relational Hash primitive—one-wayness, twin one-wayness, unforgeability and oracle simulatability. We develop a Relational Hash scheme for discovering linear relations among bit-vectors (elements of $\mathbb{F}_2^n$) and $\mathbb{F}_p$-vectors. Using the linear Relational Hash schemes we develop Relational Hashes for detecting proximity in terms of hamming distance. The proximity Relational Hashing schemes can be adapted to a privacy preserving biometric identification scheme, as well as a privacy preserving biometric authentication scheme secure against passive adversaries.

Keywords Probabilistic hash functions · Functional encryption · Biometric authentication

1 Introduction

Consider a scenario where there is a database of fingerprints of known criminals. The database should not reveal the actual fingerprints, even internally. An investigative officer might want to check, whether a candidate fingerprint digest matches with

A. Mandal · A. Roy (✉)
Fujitsu Laboratories of America, Sunnyvale, CA, USA
e-mail: aroy@us.fujitsu.com

A. Mandal
e-mail: amandal@us.fujitsu.com

B. Anderssen et al. (eds.), *The Role and Importance of Mathematics in Innovation*, Mathematics for Industry 25,
DOI 10.1007/978-981-10-0962-4_9

the database. How can we build a biometric identification scheme which guarantees complete template privacy to both the server, as well as to the investigating officer?

We observe that Homomorphic Encryption [3] does not entirely solve this problem. In Homomorphic Encryption, one of the parties has to have a decryption key, in order to decrypt the final result. However, this decryption key also enables the party to decrypt the corresponding input from the other party, thus eluding true bipartite privacy. In particular, if the decryption key is leaked, the whole database could be compromised.

We propose a cryptographic primitive called *Relational Hash* [4] which attempts to address the question above. One of the key ideas is to have distinct, but related, hashing systems for the individual co-ordinates, i.e., have two hash functions h_1 and h_2 and enable checking of $x_1 \stackrel{?}{=} x_2$, given $h_1(x_1)$ and $h_2(x_2)$. Extending equality, we define *Relational Hash* with respect to a relation R, such that given two hashes $h_1(x_1)$ and $h_2(x_2)$, we can efficiently determine whether $(x_1, x_2) \in R$ holds. It may also be desirable to compute ternary relations R' on x_1, x_2 and a third plaintext parameter z, so that given $h_1(x_1)$, $h_2(x_2)$ and z, we can efficiently determine whether $(x_1, x_2, z) \in R'$ holds. For any Relational Hash primitive, we formalize a few natural and desirable security properties, namely one-wayness, unforgeability, twin one-wayness and oracle simulatability. We emphasize here that there are no secret keys in the system to decrypt the input data, thus mitigating the drawback of Homomorphic Encryption wherein leakage of a secret key can compromise the database.

In this talk I will describe a relational hash construction for checking linear relations and then show how we use the linear relational hash to construct a relational hash system for verifying proximity, which addresses the biometric privacy scenario that we set out with. The biometric application is depicted in Fig. 1.

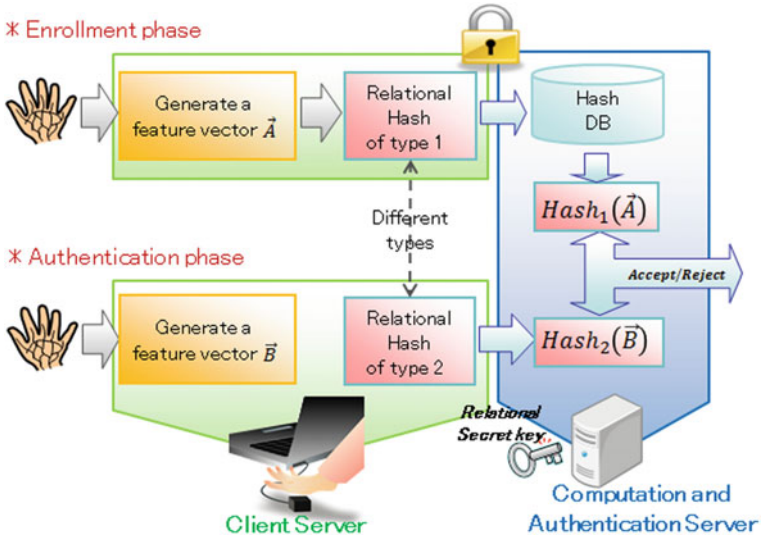


Fig. 1 Relational hash for biometric authentication

1.1 Why Traditional Hash Functions Are Not Sufficient

Traditional cryptographic hash functions, like MD-5 and SHA-3, enable checking for equality while hiding the plaintexts. Since these are deterministic functions, this just involves checking if the hashes are identical. However, hash functions are not useful for biometric matching as biometric data are noisy.

The notion of probabilistic hash functions was developed in [1, 2]. In this setting, the computation of hashes is randomized and thus no two independently generated hashes of the same plaintext look same. However, given the plaintext and a hash, it can be checked efficiently if the hash corresponds to the plaintext. However, probabilistic hashes suffer from the drawback that for verification of equality the plaintext has to be provided in the clear, which deterministic hashes do not require. Probabilistic hashes do not allow checking whether the plaintexts are equal, given two distinct hash values. This drawback can preclude use of probabilistic hashes in certain scenarios where it is desirable to hide the plaintext from the verifier as well. For example, consider a scenario where password equality is to be checked by a server. If the server uses deterministic hashes, then only the hash of the password could be transmitted to the server. However, with probabilistic hashes, the actual password has to be sent to the server for verification. With relational hashes, we allow verification given two distinct hashes of the passwords, as shown in Fig. 1.

References

1. Canetti, R.: Towards realizing random oracles: hash functions that hide all partial information. In: Burton, S., Kaliski, Jr., (ed.) CRYPTO'97, LNCS, vol. 1294, pp. 455–469. Springer, Heidelberg (1997)
2. Canetti, R., Micciancio, D., Reingold, O.: Perfectly one-way probabilistic hash functions (preliminary version). In: 30th ACM STOC, pp. 131–140. ACM Press (1998)
3. Gentry, C.: Fully homomorphic encryption using ideal lattices. In: Mitzenmacher, M (ed.) 41st Annual ACM Symposium on Theory of Computing, pp. 169–178, Bethesda, Maryland, USA, May 31 June 2, ACM Press (2009)
4. Mandal, A., Roy, A.: Relational hash: probabilistic hash for verifying relations, secure against forgery and more. In: Gennaro, R., Robshaw, M. (eds.) CRYPTO 2015 LNCS, Part I, vol. 9215, pp. 518–537. Springer, Heidelberg (2015)

Cryptography and Financial Industry

Takenobu Seito

Abstract Cryptographic algorithms are widely used for the purpose of ensuring security of data used in financial transactions (e.g., ATM transactions, the online banking). RSA has been one of the most widely used cryptographic algorithms. Currently, the migration from RSA to Elliptic Curve Cryptography is an important agenda in many sectors including the financial industry. Comparing with RSA, Elliptic Curve Cryptography has the following two advantages: (1) It can provide the same security level by smaller key sizes, (2) The probability of the occurrence of security flaws by operational issues is much lower. In this paper, we will introduce a brief overview of the recent situation of RSA and explain the usage of Elliptic Curve Cryptography in the financial industry. Also, we will show the recent study on the security evaluation of Elliptic Curve Cryptography.

Keywords Financial industry · Public-key cryptography · RSA · Elliptic curve cryptography

1 Introduction

Background. In the financial industry, cryptographic algorithms are widely used as fundamental techniques in order to ensure security of financial transaction data. For example, the algorithms are used to ensure confidentiality as well as integrity of important data (e.g., a personal identification number, or a password) and authenticity of smart cards (e.g., bank cards, debit cards or credit cards) in ATM transactions. In the retail financial service through the Internet such as the online banking,

The views expressed in this paper are those of the author and do not necessarily the official views of Bank of Japan.

T. Seito (✉)
Center for Information Technology Studies, Institute for Monetary and Economic Studies, Bank of Japan, 2-1-1 Nihonbashi-Hongokucho, Tokyo, Chuo-ku 103-8660, Japan
e-mail: takenobu.seitou@boj.or.jp

B. Anderssen et al. (eds.), *The Role and Importance of Mathematics in Innovation*, Mathematics for Industry 25,
DOI 10.1007/978-981-10-0962-4_10

many financial institutions adopt the cryptographic protocol called Secure Socket Layer/Transport Layer Security[1] (SSL/TLS for short) [4] in order to ensure confidentiality and integrity of their transmission data between the financial institutions and end users. There are two major classes of the cryptographic algorithms, Symmetric-Key Cryptography and Public-Key Cryptography. In Symmetric-Key Cryptography, a cryptographic key for the encryption is identical to that for the decryption. The key is shared securely among users. On the other hand, in Public-Key Cryptography, a cryptographic key for the encryption is different from that for the decryption. The encryption key is called a public-key because it can be widely distributed for unspecified users. The decryption key is kept secret by its owner as secret-key.

In order to design these algorithms, mathematics is fundamentally used. Especially, Public-Key Cryptography makes use of it more fundamentally than Secret-Key Cryptography does. For instance, RSA [19], one of the most famous algorithms of Public-Key Cryptography, is constructed based on Euler's theorem. Moreover, the security of RSA depends on the hardness of a large integer factoring problem which is evaluated using the recent study of the number theory. On the other hand, the security of Symmetric-Key Cryptography is not based on the mathematically hard problems. From the above fact, it can be considered that Public-Key Cryptography relates more closely with mathematics than Symmetric-Key Cryptography. Thus, we will focus on Public-Key Cryptography in this paper.

Current Issues of RSA. RSA (named after Rivest, Shamir and Adleman) is the most popular algorithms for Public-Key Cryptography according to major international standards and guidelines in the financial industry. RSA, however, has the following two issues. The first is that its key size is relatively large. Therefore, many researchers point out that it will be difficult to implement RSA in cryptographic hardware modules with a limited computational power. The second is that RSA has the vulnerability of generating weak keys which cause the security flaws.

Elliptic Curve Cryptography: Alternative to RSA. Elliptic Curve Cryptography (ECC for short), which is another type of Public-Key Cryptography, has attracted much attention as an alternative Public-Key Cryptography to RSA. Comparing with RSA, ECC has the following two advantages: (1) It can provide the same security level by smaller key sizes (about $1/10$), (2) The probability of generating weak keys is much lower. The migration from RSA to ECC has been an important topic in many sectors including the financial industry. In fact, ECC is specified in several international standards (e.g., ISO and IEC) and guidelines (e.g., EMV specification) regarding security techniques in the financial industry. ECC is also specified in SSL/TLS as one of utilizable cryptographic algorithms. Therefore, ECC will become a mainstream of Public-Key Cryptography in the future.

In this paper, we will introduce a brief overview of the recent situation around RSA and explain the usage of ECC in the financial industry. After that, we will show the recent study on the security evaluation of ECC.

[1] It is specified by International Engineering Task Force (IETF) which develops and maintains international standards regarding information techniques used on the Internet.

Table 1 Role of keys for achieving confidentiality and integrity

Kind of security functions	Role of keys	
	Public-key	Secret-key
Confidentiality	Encrypt a message	Decrypt a ciphertext
Integrity (Authenticity)	Verify the validity of a signature	Generate a signature of a message

2 Recent Situation Around RSA

2.1 Public-Key Cryptography

Public-Key Cryptography can realize both confidentiality and integrity (authenticity) by using two different keys: the public-key and the secret-key (see Table 1). To guarantee confidentiality, the public-key is used for encrypting a message (plaintext), and the secret-key is used for decrypting a ciphertext to obtain the message. On the other hand, to guarantee integrity, the secret-key is used for generating a digital signature of a message, and the public-key is used for verifying the validity of the signature.

The most basic security requirement of Public-Key Cryptography is that it is difficult to compute the secret-key from the corresponding public-key. This difficulty is generally based on the assumption of mathematically hard problems that are infeasible to solve such as the integer factoring problem, the discrete logarithm problem, and the elliptic curve discrete logarithm problem.

2.2 RSA and Its Issues

We describe the RSA encryption algorithm which consists of key generation, encryption, and decryption phases.

1. **Key Generation**. Each user generates two large primes P and Q. Then, the user computes a composite integer $N = P \times Q$. Also, the user selects two natural numbers e and d which satisfy the following two equations:

$$\gcd(e, (P-1) \times (Q-1)) = 1,$$
$$e \times d \equiv 1 \mod (\varphi(N)),$$

   where $\varphi(N)$ is the Carmichael function. The user sets (P, Q, d) and (N, e) as the secret-key and the public-key, respectively.
2. **Encryption**. For the message M, a message sender computes the ciphertext $C = M^e \mod N$ by using the public-key and sends it to a message receiver.

3. **Decryption**. After receiving the ciphertext C, the receiver decrypts the message $M = C^d \mod N$ by using the secret-key.

The security of RSA is based on the difficulty of the integer factoring problem (IFP for short) which is defined as follow.

Definition 1 (*IFP*) Given a large composite integer N, find prime factors of N.

Roughly speaking, the secret-key of RSA is a pair of two primes P and Q, and the public-key is the composite integer $N = P \times Q$. It is easy to compute the public-key N from the secret-key P and Q. Conversely, it is hard to compute the secret-key P and Q from the public-key N based on the assumption of IFP. Furthermore, the hardness of IFP is proportional to the size of the composite integer N. So far, no algorithms has been proposed to solve RSA efficiently.

Currently, it is pointed out that RSA has two practical issues: a large key size and a weak keys.

Large key size. The first issue is the restriction of the hardware implementation by increasing the key size of RSA in the future. In general, the security level declines gradually due to the development of new attack algorithms and the improvement of the cost performance of computers. To prevent the security level from decreasing, it is necessary to lengthen the key size. So far, the key size of RSA has been expanded from 768-bits to 2,048-bits periodically. A current mainstream of its key size is 2,048-bits. National Institute for Standards and Technology (NIST) recommends the migration from 2,048-bit key size to 3,072-bit key size up to 2030 [17].

It will be more difficult to use RSA with longer keys in hardware devices such as smart cards and embedded devices in the future. Especially, it is pointed out that RSA with 4,096-bits is infeasible to be implemented in currently used smart cards. Since the smart cards are widely used in the financial industry, it will become a very important issue. Therefore, there is a need for cryptographic algorithms which can provide a same security level as RSA by smaller key sizes.

Weak keys. It is well known that the vulnerability of weak keys exists due to the inappropriate implementation. For instance, if two different users have the same prime (at least one) as the secret-key, it is easy to compute both users' secret-keys from the corresponding public-keys by the following attack algorithm (Table 2).

Obviously, the secret-key can be computed in polynomial time. Thus, the impact on the security of RSA is significant. Since the number of primes which can be

Table 2 Attack algorithm by using weak keys

Input: Two public-keys $N_1 = P_1 \times Q$ and $N_2 = P_2 \times Q$ where P_1, P_2, and Q are primes
Output: Two secret-keys (P_1, Q) and (P_2, Q)
Step 1. Compute $Q = \gcd(N_1, N_2)$
Step 2. Compute $P_1 = N_1/Q$ and $P_2 = N_2/Q$

Table 3 Typical concrete algorithms of ECC

Security functionality	Typical example
Confidentiality	Elliptic Curve LeGamal Encryption (ECElGamal)
Integrity (authenticity)	Elliptic Curve Digital Signature Algorithm (ECDSA)
Key Agreement	Elliptic Curve Diffie–Hellman Key Agreement (ECDH) Elliptic Curve Menezes–Qu–Vanstone Key Agreement (ECMQV)

used as the secret-key is usually large,[2] the probability which the same prime is selected by the different two users is negligible. However, it has been pointed out that the distribution of secret-keys is biased if the implementation of the pseudo-random number generator is inappropriate. In fact, some researchers pointed out that there exists vulnerable keys which are used over the Internet [7, 13]. Even if two uses select secret-key as (P_1, Q) and (P_2, Q) where P_1, P_2 and Q are large primes sufficing $P_1 \neq P_2$, the corresponding public-keys $N_1 = P_1 \times Q$ and $N_2 = P_2 \times Q$ are different integers. Therefore, such weak keys are difficult to be detected by system administrators. In order to detect the weak keys, it is needed to check all pairs of public-keys by using the Euclidean algorithm.

3 ECC: Alternative Public-Key Cryptography to RSA

3.1 Overview

ECC was introduced by Koblitz and Miller independently of one another [11, 15]. The security of ECC is based on the hardness of the elliptic curve discrete logarithm problem (ECDLP for short). An elliptic curve is a special type of a cubic equation over finite fields. As finite fields, prime and binary fields are often adopted to construct algorithms. It is well known that the point addition can be defined on the elliptic curve, and this property is applied for designs and security evaluations of ECC. There are some concrete algorithms in ECC. Here we summarize typical concrete algorithms of ECC categorized by the security functionality in Table 3.

We describe the algorithm of ECElGamal as a typical example of ECC. Before generating each user's key, it is necessary to select a common parameter which is shared by all users. This parameter determines the type of an elliptic curve and it is closely related to the security of ECC. It is a parameter unique to ECC and such a parameter is not required in RSA. The common parameter is usually generated by a trusted authority. The algorithm of ECElGamal consists of the following phases.

[2] For example, in RSA with 2,048-bits key, the number of such primes is about $2^{1,015}$ from the prime number theorem.

1. **Common Parameter Generation**. A trusted authority selects the following parameters: (i) the type of an elliptic curve (i.e., the type of a finite field and coefficients of the curve), and (ii) the point $\mathbf{G}$ on the elliptic curve. Then, the trusted authority publishes a set of these parameters as the common parameter.
2. **Key Generation**. Each user selects a natural number s. After that, the user computes a point $\mathbf{T} = s \times \mathbf{G}$ by using the common parameter. The secret-key is s and the public-key is $\mathbf{T}$.
3. **Encryption**. For the message $\mathbf{M}$, the sender selects a random number r and computes $\mathbf{C}_1 = r \times \mathbf{G}$, and $\mathbf{C}_2 = \mathbf{M} + (r \times \mathbf{T})$ by using the public-key and the common parameter. After that, the sender constructs a ciphertext $C = (\mathbf{C}_1, \mathbf{C}_2)$ and sends it to the receiver.
4. **Decryption**. After receiving the ciphertext $C = (\mathbf{C}_1, \mathbf{C}_2)$, the receiver decrypts the message $\mathbf{M} = \mathbf{C}_2 - (s \times \mathbf{C}_1)$ by using the secret-key.

The security of ECC is based on the difficulty of ECDLP which is defined as follow.

Definition 2 (*ECDLP*) Given an elliptic curve over a finite field (the prime field or the binary field) and two points $\mathbf{G}$ and $\mathbf{T}$, find a natural number s with $\mathbf{T} = s \times \mathbf{G}$.

In ECC, it is easy to compute the public-key $\mathbf{T} = s \times \mathbf{G}$ from the secret-key s and the common parameter $\mathbf{G}$. Conversely, it is hard to compute the secret-key s from the public-key $\mathbf{T}$ and the common parameter $\mathbf{G}$. The hardness of ECDLP is proportional to the size of the secret-key s.

3.2 Comparison Between ECC and RSA

Comparing with RSA, ECC has the following practical advantages.

1. **Smaller key sizes**. ECC can provide the same security level by smaller key sizes (about 1/10). The reason for this advantage is that ECDLP is harder to solve than IFP. Currently, the most efficient algorithm solves ECDLP in exponential time. On the other hand, the efficient algorithm for IFP solves in sub-exponential time. Here, we show a comparison of key sizes which provide the same security levels in Table 4. This was estimated by comparing the current hardness of ECDLP and IFP on the basis of the recent computational power of computers and the improvement of attack algorithms. This advantage makes ECC more suitable for the implementation to smart cards and embedded devices.
2. **Advantage of operational aspect**. The mechanism of the implementation to the key generation is different from that of RSA. In ECC, each user randomly selects the natural number s as the secret-key. It is well known that the number of natural numbers is much larger than that of primes. Therefore, the probability of selecting the same secret-key among different users is much lower than that of RSA. In addition, if different users select the same natural number as secret-keys, the

Table 4 Comparison of key sizes between ECC and RSA (NIST [17])

Key sizes of ECC (bits)	Key sizes of RSA (bits)
160–223	1,024
224–255	2,048
256–383	3,072
384–511	7,680
512–521	15,360

corresponding public-keys are the same value. Thus, it is easy to notice such an event by comparing the existing public-keys.

These are why ECC has been paid much attention to as an alternative Public-Key Cryptography to RSA.

3.3 Usage of ECC in the Financial Industry

SSL/TLS Standard. In the online banking, SSL/TLS is widely used as the cryptographic protocol. ECC is standardized in the latest version of SSL/TLS (TLS1.2) [4] as the utilizable cryptographic algorithm. More precisely, SSL/TLS consists of the following phases.

- Phase 1. Authentication of the server using Public-Key Cryptography.
- Phase 2. Key Agreement for Symmetric-Key Cryptography.
- Phase 3. Secure communication using Symmetric-Key Cryptography.

ECDSA and ECDH can be used in the phases 1 and 2, respectively.

EMV Specification. EMV specification[3] is an international standard for debit/credit cards payment systems using smart cards. This specification is managed by EMVCo which is a consortium of some international payment service corporations (American Express, Discover, JCB, MasterCard, UnionPay, and VISA). In EMV specification, RSA is now used for authenticity of smart cards (e.g., bank cards, debit cards, and credit card) in the transaction. In 2009, EMVCo indicated the roadmap for the migration of RSA to ECC in this specification to enhance the security and the practicality of the transaction environment which is compliant with EMV specification [5].

ISO Standards. ECC is now specified in many ISO standards regarding security techniques (Table 5). Referring to these, the financial institutions seem to adopt ECC in their services.

[3]Japanese Bankers Association (JBA) IC Cash Card Specifications are compliant with the EMV specification.

Table 5 ISO standards specifying ECC (ISO [9], ISO/IEC [8, 10])

International standard	Specified ECC algorithms
ISO 11568 Key Management (retail)	Authenticity: RSA, ECDSA Key Agreement: ECDH, ECMQV
ISO/IEC 11770 Key Management	Key Agreement: ECDH
ISO/IEC 14888 Digital Signature with Appendix	Authenticity: RSA, ECDSA

In addition, U.S financial standards also adopt ECC as secure Public-Key Cryptography algorithms. For instance, ANSI X9.62 specifies ECC as the algorithm for Public-Key Cryptography for the financial services [2].

3.4 Security Evaluation of ECC

The security of ECC is based on the hardness of ECDLP. However, if the selected common parameter satisfies specific conditions, then ECDLP can be efficiently solved by known attack algorithms. For example, MOV reduction [14], FR reduction [6], SSSA algorithm [20, 21, 23], and Index calculus [1, 12] are well known as major attack algorithms. Therefore, it is important to select the common parameter in order to defend against these algorithms. The trusted third party such as NIST and SECG (The Standards for Efficient Cryptography Group) publish recommended common parameters for using ECC securely [3, 17].

The key size is also an important factor for the security of ECC, because the brute force can be applied. So far, Pollard ρ algorithm [18] and Baby-Step Giant-Step algorithm [22] have been proposed. Thus it is needed to select the key size in such way to prevent from these algorithms. At present, NIST recommends 224-255-bit or higher key size for use up to 2030 [17].

4 Conclusion

Cryptographic algorithms are crucial techniques to ensure the security of various financial services such as the online banking and smart card transactions. So far, RSA has been widely used as Public-Key Cryptography in the financial sector. Recently, ECC has been paid much attention as an alternative Public-Key Cryptography to RSA. Comparing with RSA, ECC can provide the same security level by smaller key sizes and the probability of causing the vulnerability of weak keys is much lower. Thus, the migration from RSA to ECC is the important agenda in the financial industry. Major international standards regarding the cryptographic algorithms and protocols specify ECC as a new algorithm. It is expected that the security of ECC would be discussed more carefully in the future.

References

1. Adleman, L.: A Subexponential algorithms for the discrete logarithm problem with applications to cryptography. In: Proceedings of Foundations of Computer Science (FOCS), pp.55–60 (1979)
2. American National Standards Institute (ANSI).: X9.62: Public-key cryptography for the financial service industry: The elliptic curve digital signature algorithm (ECDSA). ANSI (2005)
3. Certicom.: SEC 2: recommended elliptic curve domain parameters version 2.0. Standards for efficient cryptography group (2010)
4. Dierks, T., Rescorla, E.: The transport layer security (TLS) protocol version 1.2. Request for Comments (RFC), vol. 5246 (2008)
5. EMVCo.: EMVCo common contactress terminal roadmap. General Bulletin, vol. 43 (2009)
6. Frey, G., Ruëk, H-G.: A remark concerning m-divisibility and the discrete logarithm in the divisor class group of curves. Math. Comput. **62**(206), 865–874 (1994)
7. Heninger, N., Durumeric, Z., Wustrow, E., Halderman, A.: Mining your Ps and Qs: detection of widespread weak keys in network devices. In: Proceedings of USENIX Security Symposium (2012)
8. International Organization for Standarization (ISO) and International Electrotechnical Commission (IEC).: ISO/IEC 14888-3: Information technology – security techniques – digital signatures with appendix – Part 3: discrete logarithm based mechanism. ISO and IEC (2006)
9. International Organization for Standarization (ISO).: ISO 11568-4: Banking – Key management (retail) – Part 4: asymmetric cryptosystems – key management and life cycle. ISO (2007)
10. International Organization for Standarization (ISO) and International Electrotechnical Commission (IEC).: ISO/IEC 11770-3: information technology – security techniques – key management – Part 3: mechanisms using asymmetric techniques. ISO and IEC (2015)
11. Koblitz, N.: Elliptic curve cryptosystems. math comput **48**, 203–209 (1987)
12. Kraitchik, M.: Thòrie des nombres. Gauthier-Villars (1922)
13. Lenstra, A.K., Hughes, J., Augier, M., Bos, J., Kleinjung, T., Wachter, C.: Ron was wrong, whit is right. In: International Association for Cryptographic Research (IACR) Cryptography ePrint Archive, vol. 64 (2012)
14. Menezes, A., Vanstone, S., Okamoto, T.: Reducing elliptic curve logarithms to logarithms in a finite field. In: Proceedings of Symposium on Theory of Computing (STOC), pp.80–89 (1991)
15. Miller, V.: Uses of elliptic curves in cryptography. In: Williams, H.C. (ed.) CRYPTO 1985, vol. 218, pp. 417–426. LNCS, Springer, Heidelberg (1986)
16. National Institute of Standard and Technology (NIST).: Advanced encryption standard. In: Federal Information Processing Standardization (FIPS) 197 (2001)
17. National Institute of Standard and Technology (NIST).: Recommendation on key management. Special Publication (SP), pp. 800–57 (2012)
18. Pollard, John M.: Monte Carlo method for index computation (mod p). Math Comput **32**(143), 918–924 (1978)
19. Rivest, R., Shamir, A., Adleman, L.: A Method of obtain digital signatures and public key cryptosystems. Commun. ACM **21**(2), 361–396 (1978)
20. Satoh, T., Araki, K.: Fermat quotieents and the polynomial time discrete log algorithm for anomalous elliptic curves. Commentarii Mathematici, Universiatis Sancti Pauli, pp. 81–92 (1998)
21. Semaev, I.: Evaluation of discrete logarithms in a group of p-torsion points of an elliptic curve in characteristic p. Math. Comput. **67**(221), 353–356 (1998)
22. Shanks, D.: Class number, a theory of factorization, and genera. In: Proceedings of Symposia in Pure Mathematics, 20, pp. 415–440 (1971)
23. Smart, N.: The discrete logarithm on elliptic curves of trace one. J. Crypto. **12**(3), 193–196 (1999)

Relaxation of an Energy Model for the Triangle-to-Centred Rectangle Transformation

Pierluigi Cesana

Abstract We model and analyze the two-dimensional triangle-to-centred rectangle transformation of elastic crystals. By considering a Ginzburg–Landau type model, we compute the relaxation of the total energy both in the case of compressible and incompressible materials and construct some possible explicit microstructures as the approximate solutions of a non-convex minimization problem.

Keywords Phase-transformations · Microstructure · Variational calculus

1 Introduction

The austenite-to-martensite phase transformation is a first-order solid-to-solid transition characterized by an abrupt change of shape of the crystalline lattice driven by temperature or applied stresses. First observed in steel, it has then been discovered in ceramics, biological systems and, most important for technological applications, shape-memory alloys (SMAs) [4]. The transition from the high-temperature phase (austenite) to the low-temperature state (martensite) is usually activated by a decrease of the temperature below a critical threshold. The low-symmetry and disordered phase usually appears in the form of a mixture of symmetry-related crystal variants, called martensitic microstructure. This system is characterized by the presence of interfaces separating plates composed of various variants, possibly coexisting at different scales, which may result in complicated patterns rich in misalignments as well as vacancies of the crystal lattice usually modelled as topological defects [6, 14].

The mathematical modelling of the austenite-to-martensite phase-transition via an energy-minimization approach traces back to the work of Ericksen, Ball and James.

P. Cesana (✉)
Institute of Mathematics for Industry, Kyushu University, Fukuoka, Japan
e-mail: p.cesana@latrobe.edu.au

P. Cesana
Australia Branch, Department of Mathematics and Statistics, La Trobe University, Bundoora, VIC 3086, Australia

B. Anderssen et al. (eds.), *The Role and Importance of Mathematics in Innovation*, Mathematics for Industry 25,
DOI 10.1007/978-981-10-0962-4_11

In their paper *Fine phase mixtures as minimizers of energy*, Arch. Rat. Mech. Anal., 100 (1987), Ball and James explain the mechanism of the formation of microstructure in the general framework of minimization problems for non-convex free energies in non-linear elasticity, and of differential inclusion problems for non-linear partial differential equations with algebraic constraints associated to the crystallographic properties of the transformation. This approach has been employed by a number of authors giving rise to an extensive platform of both analytical and numerical work related to the austenite-to-martensite transformation (see [4, 13] and references quoted therein).

In this contribution, we analyze an energy model of the triangle-to-centred rectangle (briefly, TR) transformation. This is the two-dimensional version of the three-dimensional hexagonal-to-orthorhombic transformation observed in materials such as the MgCd ordered alloys [6, 17]. The TR transformation has been the subject of investigation in relation to the modelling of crystal defects (disclinations) and of their interaction with microstructure [6, 10, 12, 14, 17]. Focusing on purely elastic deformations, in what follows we characterize the low-energy states and construct possible microstructures associated to the full relaxation of a continuum model of the TR transformation.

2 Microstructure Modelling

We follow the approach of [6, 17] (see also [10, 12]) and model the triangle-to-centred rectangle (TR) transformation in the framework of the Ginzburg–Landau theory of phase transitions by considering the mechanical strain as the order parameter of the system. Formation of microstructure results as the approximate minimizer of the non-convex model.

2.1 Continuum Model

We consider Ω an open, bounded subset of R^2 with Lipschitz boundary which we address as the reference configuration occupied by the elastic crystal. Based on the assumption that the microstructure is essentially homogeneous perpendicular to the plane of the paper, we assume $u : \Omega \to R^2$ to be the two-dimensional displacement vector. In the framework of linearized elasticity, the mechanical energy density of the system depends only on the symmetric part of the 2×2 displacement gradient $F = \nabla u$ which we denote with

$$E := E(F) = \frac{F + F^T}{2}.$$

We introduce the multiwell energy density

$$\Psi(F) := \Psi_{lin}(E) = 2\mu \min_{i=1,2,3}\left[|E - E_i|^2\right] + \kappa(\mathrm{tr}\, E)^2 \tag{1}$$

by means of the auxiliary function Ψ_{lin} defined on $R^{2\times 2}_{sym}$, the space of symmetric matrices. Here μ and κ are the classical (positive) Lamé constants of linearized elasticity. Slightly different expressions for Ψ_{lin} are also possible and available in the literature (see [6, 17]). The following two properties of Ψ_{lin} are crucial:

$$\Psi_{lin}(E) \begin{cases} = 0 & \text{if } E \in \mathscr{T} \\ > 0 & \text{otherwise} \end{cases} \tag{2}$$

where $\mathscr{T} = \{E_1, E_2, E_3\}$. Conditions (2) imply that the free energy density has three minimum points, corresponding to the three martensitic variants, with strain matrices

$$E_1 = \gamma \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad E_2 = \gamma \begin{pmatrix} -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & \frac{1}{2} \end{pmatrix}, \quad E_3 = \gamma \begin{pmatrix} -\frac{1}{2} & -\frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} & \frac{1}{2} \end{pmatrix}. \tag{3}$$

$\gamma > 0$ is a constant material parameter.

By integrating Ψ over $\Omega \subset R^2$, we obtain the total elastic energy of the crystal

$$I(u) := \int_\Omega \Psi\big(\nabla u\big)dx \equiv \int_\Omega \Psi_{lin}\Big(\frac{\nabla u + \nabla^T u}{2}\Big)dx. \tag{4}$$

Correctly, the total energy depends only on the symmetrized displacement gradient. According to a variational principle, we consider as the equilibrium configurations of our elastic crystal the absolute minimizer of I in the presence of applied boundary conditions. To be more precise, we consider minimization problems for functionals defined in Sobolev spaces by applying the direct method of the calculus of variations. Here and in what follows the notation is standard [7, 8].

2.2 Relaxation

The direct method of the calculus of variations is a technique that provides a sufficient condition for the existence of solutions to minimum problems. Fix a function $\overline{u} \in H^1(\Omega, R^2)$. We investigate whether the boundary value problem for the energy I introduced in (4)

$$\min_{u-\overline{u}\in H^1_o(\Omega,R^2)} I(u) \tag{5}$$

admits a solution. The two main ingredients for the direct method to apply are boundedness of the minimizing sequences (so that one can extract a sub-sequence

converging in the weak topology) and lower semi-continuity of the functional (with respect to the weak convergence). Let u_j be a minimizing sequence for I, that is,

$$\lim_{j\to\infty} I(u_j) = \inf_{u-\overline{u}\in H^1_o(\Omega,R^2)} I(u). \tag{6}$$

From a combination of Poincaré and Korn's inequalities [7] there follows that the minimizing sequence $\{u_j\}$ is bounded in $H^1(\Omega, R^2)$. However, the direct method fails here because of the non-convexity and, more essentially, non-lower semi-continuity of I. Although this is only a sufficient condition, it is actually easy to construct boundary conditions for which (5) does not have a solution.

Furthermore, if we take $\overline{u} \equiv 0$, the following holds. Let u_j be a minimizing sequence for I. Then ∇u_j develops finer and finer oscillations as $j \to \infty$. This is what we call a microstructure. The asymptotic behaviour of the minimizing sequence is captured by the lower semi-continuous envelope of I also called the relaxation of I, defined as

$$\overline{I}(u) := \inf\left\{\liminf_{k\to+\infty} I(u_k),\ u_k \rightharpoonup u \ in\ H^1(\Omega, R^2), u_k - \overline{u} \in H^1_o(\Omega, R^2)\right\}.$$

The relaxation is characterized by the fundamental property

$$\inf_{u-\overline{u}\in H^1_o(\Omega,R^2)} I(u) = \min_{u-\overline{u}\in H^1_o(\Omega,R^2)} \overline{I}(u) \tag{7}$$

and that the minimizing sequence defined in (6) converges (possibly up to a subsequence) to a minimizer of $\overline{I}$. Explicit knowledge of the relaxation $\overline{I}$ turns out to be a powerful tool to achieve a qualitative as well as quantitative insight on the microstructure and its influence on the macroscale properties of the system, what is often called a mesoscale or effective modelling. In the current situation, by taking advantage of the geometric structure of the function Ψ it turns out to be possible to compute the relaxation of I exactly. This result follows essentially from ideas and constructions already contained in [3, 5, 9, 11, 16].

To begin, we introduce some terminology.

Definition 1 (*Quasiconvexity*) We recall the fundamental definition of quasiconvexity [1]. A continuous function $f : R^{2\times2} \to R$ is said to be quasiconvex if and only if $\forall Z \in R^{2\times2}$, ω open bounded subset of R^2, $w \in C^1_o(\omega, R^2)$ we have

$$f(Z) \le |\omega|^{-1} \int_\omega f(Z + \nabla w(y))dy.$$

Definition 2 (*Rank-1 convexity*) A function $f : R^{2\times2} \mapsto R$ is rank-1 convex [8] if $f(sZ_1 + (1-s)Z_2) \le sf(Z_1) + (1-s)f(Z_2)$ for every $s \in [0, 1]$, $Z_1, Z_2 \in R^{2\times2}$ with

$$\operatorname{rank}(Z_1 - Z_2) \leq 1. \tag{8}$$

By removing the constraint (8) we encounter the usual definition of convexity.

Definition 3 (*Envelopes*) Let us now consider a generic function $0 \leq g : R^{2\times 2} \to R$. We define the convex envelope of g as $g^{co}(Z) := \sup\{h(Z) : h \leq g, h \text{ convex}\}$. In the same way we define the quasi- and rank-1 convex envelopes, by requiring that the function h satisfies the corresponding requirement of partial convexity.

If we restrict our attention to the case of real-valued functions, the following chain of inequalities follows by definition (see [8], p. 265)

$$g^{co} \leq g^{qc} \leq g^{rc}. \tag{9}$$

The following characterization of the rank-1 convex envelope (see [8, Theorem 6.10] and see [8, Sect. 6.4]) has a relevant role in what follows

$$g^{rc}(Z) = \inf\Big\{\sum_i^K \lambda_i g(Z_i) : 0 \leq \lambda_i \leq 1, \sum_i^K \lambda_i = 1, \sum_i^K \lambda_i Z_i = Z, \ \{\lambda_i, Z_i\} \text{ satisfy } (H_K)\Big\}. \tag{10}$$

Here $Z_i \in R^{2\times 2}$ for all i. Condition (H_K) (for brevity here not reported, see [8, Sect. 5.2.5]) consists in a series of algebraic constraints ensuring that the matrix Z in (10) is obtained by a kinematically compatible combination of matrices Z_i. As an example, in the case $K = 2$, the pair (λ_1, Z_1) and (λ_2, Z_2) satisfy (H_2) if $\lambda_1, \lambda_2 \in [0, 1]$ with $\lambda_1 + \lambda_2 = 1$ and Z_1, Z_2 verify

$$\operatorname{rank}(Z_1 - Z_2) = 1,$$

that is, Hadamard's jump condition [2]. As a relevant case for the TR transformation, for $K = 3$ we have that (λ_i, Z_i) satisfy condition (H_3) if $\lambda_i \in [0, 1]$, $\sum_i^3 \lambda_i = 1$ and if, up to a permutation,

$$\operatorname{rank}(Z_2 - Z_3) \leq 1 \tag{11}$$

$$\operatorname{rank}\Big(Z_1 - \frac{\lambda_2 Z_2 + \lambda_3 Z_3}{\lambda_2 + \lambda_3}\Big) \leq 1. \tag{12}$$

Our relaxation result of Theorem 1 is based on the computation of a compatible combination of matrices that is optimal in the sense that it realizes the minimum in Eq. (10). This procedure is frequently referred to as lamination construction (see [15] for the definition of laminates).

Remark Note that for the integral functional $J : H_o^1(\Omega, R^2) \to R$ defined as $J := \int_\Omega f(\nabla u)dx$ with $f(F)$ quasiconvex and $c|F|^2 - \tilde{c} \leq f(F) \leq C|F|^2$ (with $0 < c < C, \tilde{c} > 0$) there follows the lower semi-continuity of J with respect to the weak topology of $H^1(\Omega, R^2)$. The existence of a solution to minimum problem for J over $H_o^1(\Omega, R^2)$ then follows by the direct method.

We are now in a position to state our main relaxation results for the model of compressible (Theorem 1) and incompressible materials (Theorem 2).

Theorem 1 (Compressible materials) *Let $I : H^1_o(\Omega, R^2) \to R$ as defined in (1) and (4). Then*

$$\overline{I}(u) = \int_{\Omega} \overline{\Psi}(\nabla u) dx$$

where $\overline{\Psi} \equiv \Psi^{co}$ the convex envelope of Ψ.

Sketch of the proof. From the theory of Acerbi-Fusco [1] there follows that $\overline{\Psi} \equiv \Psi^{qc}$. The proof of Theorem 1 then consists in showing that $\Psi^{qc} \equiv \Psi^{co}$. We do so by matching an upper bound with a lower bound to Ψ^{qc}.

Lower bound. Note that $\Psi^{co} \leq \Psi^{qc}$ because (real-valued) convex functions are quasiconvex (9).

Upper bound. Consider the particular geometric structure of Ψ_{lin}. Introducing the distance function, we have

$$\Psi(F) = \Psi_{lin}(E) = 2\mu \min_{i=1,2,3} |E - E_i|^2 + \kappa(\operatorname{tr} E)^2 = 2\mu \operatorname{dist}^2\big(E, \mathscr{T}\big) + \kappa(\operatorname{tr} E)^2 \quad (13)$$

where $E = \frac{F+F^T}{2}$. By recalling that $\Psi^{qc} \leq \Psi^{rc}$ (9), it is then enough to show that

$$\Psi^{rc}(F) \leq 2\mu \operatorname{dist}^2\big(E, co\mathscr{T}\big) + \kappa(\operatorname{tr} E)^2 \leq \Psi^{co}(F) \quad (14)$$

where $co\mathscr{T}$ denotes the convex envelope of the set $\mathscr{T}$. By definition of convex envelope, the properties of the distance function and since $\mathscr{T} \subseteq co\mathscr{T}$, the last inequality in (14) follows immediately. We are left to show the first inequality in (14). We begin with considering the case $F \in co\mathscr{T}$. This implies that the both the trace and the skew part of F are zero, in other words $F = E$ and $\operatorname{tr} E = 0$. We perform a lamination construction thus obtaining $co\mathscr{T}$ as the average of kinematically compatible matrices in the sense of (H_K) whose symmetric parts belong to the set $\mathscr{T}$. Let us write

$$E = \gamma \begin{pmatrix} e_2 & e_3 \\ e_3 & -e_2 \end{pmatrix} \quad (15)$$

with $e_2, e_3 \in R$. The assumption $E \in co\mathscr{T}$ can be implemented by the parameterization

$$e_2 = \frac{3\lambda_1 - 1}{2}, \quad e_3 = \alpha(1 - \lambda_1)$$

with $0 \leq \lambda_1 \leq 1$ and $-\frac{\sqrt{3}}{2} \leq \alpha \leq \frac{\sqrt{3}}{2}$ so that both λ_1 and α can be easily expressed in terms of e_2, e_3. We can then take

$$\begin{array}{c} \bullet \quad \lambda_2 = \lambda(1-\lambda_1), \lambda_3 = (1-\lambda)(1-\lambda_1) \\ \bullet \quad Z_1 = E_1 + W(\phi_1), Z_2 = E_2 + W(\phi_2) + W(\theta_2), Z_3 = E_3 + W(\phi_2) + W(\theta_3) \end{array} \tag{16}$$

where

$$\begin{array}{c} \bullet \quad \phi_1 = \pm(1-\lambda_1)\sqrt{\frac{9}{4}+\alpha^2}, \phi_2 = \mp\lambda_1\sqrt{\frac{9}{4}+\alpha^2} \\ \bullet \quad \theta_2 = \pm\sqrt{3}(1-\lambda), \theta_3 = \mp\sqrt{3}\lambda \\ \bullet \quad \lambda = \frac{\alpha}{\sqrt{3}} + \frac{1}{2} \end{array} \tag{17}$$

and

$$W(\phi) = \gamma \begin{pmatrix} 0 & +\phi \\ -\phi & 0 \end{pmatrix}. \tag{18}$$

Choice of parameters (16–17) is determined by a series of algebraic constraints represented by condition (H_K). More precisely, the role of the off-diagonal matrix $W(\phi)$ is to guarantee that the lamination construction is kinematically compatible while leaving the symmetric part of the matrix Z_i equal to E_i for $i = 1, 2, 3$. Thus, it is immediate to see that $(\lambda_i, Z_i)_i^3$ satisfy (H_K) with $K = 3$ (see Eqs. 11–12) and that the family $(\lambda_i, Z_i)_i^3$ is *optimal* in the sense that

$$\Psi^{rc}(F) \leq \sum_i^3 \lambda_i \Psi(Z_i) = \sum_i^3 \lambda_i \Big[2\mu \min_{i=1,2,3} \Big| \frac{Z_i + Z_i^T}{2} - E_i \Big|^2 \Big] = 0 = 2\mu \, \mathrm{dist}^2(F, co\mathscr{T}). \tag{19}$$

The first inequality in (19) follows from the characterization of Ψ^{rc} (see Eq. (10)) while the chain of equalities holds because the symmetric part of Z_i with $i = 1, 2, 3$ belongs to $\mathscr{T}$. The general proof with any $F \in R^{2\times 2}$ (possibly with non-zero trace or non-zero skew part) follows easily. □

The classical way to model incompressible materials in linearized elasticity is to consider the limit ratio $\kappa/\mu = +\infty$. This is equivalent to restricting the admissible deformation gradients to the class of traceless matrices, and hence to define an energy functional in the presence of a linear constraint on the gradient of the displacement. This yields the following results.

Theorem 2 (Incompressible materials) *Let $I : H_o^1(\Omega, R^2) \to R$ as in Theorem 1. Define the functional $I_\infty : H_o^1(\Omega, R^2) \to R \cup \{+\infty\}$ as*

$$I_\infty(u) = \begin{cases} I(u) & if \ \ div\, u = 0 \\ +\infty & otherwise. \end{cases} \tag{20}$$

Then

$$\overline{I_\infty}(u) = \begin{cases} \int_\Omega \overline{\Psi}(\nabla u) dx & if \ \ div\, u = 0 \\ +\infty & otherwise \end{cases} \tag{21}$$

where $\overline{\Psi} \equiv co\Psi$ *is the convex envelope of* Ψ.

Remark Proof of Theorem 2 can be easily obtained by modifying the Proof of [5, Theorem 2] accordingly.

Remark Although in Theorems 1 and 2 we are focussing on functionals defined over the subspace of zero boundary displacement, the relaxation result still holds in the presence of more general boundary conditions as in the situation of [5].

3 Discussion and Summary

The main outcome of Theorem 1 (and, similarly, Theorem 2) consists in providing an explicit formula for the relaxed energy $\overline{I}$. Although it is known that the relaxed energy density $\overline{\Psi}$ coincides with the quasiconvex envelope of Ψ [1, 8], this information is in general of little use to fully characterize quantitatively the relaxed functional. Trying to compute the relaxed density by applying the definition of quasiconvex envelope turns out to be a prohibitive task. In practical cases, as the scenario analyzed in this paper, the possibility of computing quasiconvex envelopes reduces in matching an upper bound (based on the notion of rank-1 convexity) with a lower bound (based on convexity). In the current situation, it is found that the relaxed energy density $\overline{\Psi}$ is a convex function and coincides with the measure of the distance from the set of zero-energy mechanical strains, represented by the convex envelope of the set $\mathscr{T}$. The computation of $co\mathscr{T}$ is a trivial algebraic computation and in Fig. 1 we represent a parameterization of $co\mathscr{T}$ in the space (e_2, e_3).

As a first result, we have all the low-energy states of the system, and, therefore all the possible microstructures, occur as a mixture of kinematically compatible combinations of the matrices E_1, E_2 and E_3 at the level of simple laminates or, at most, laminates within laminates (see our construction of the upper bound in

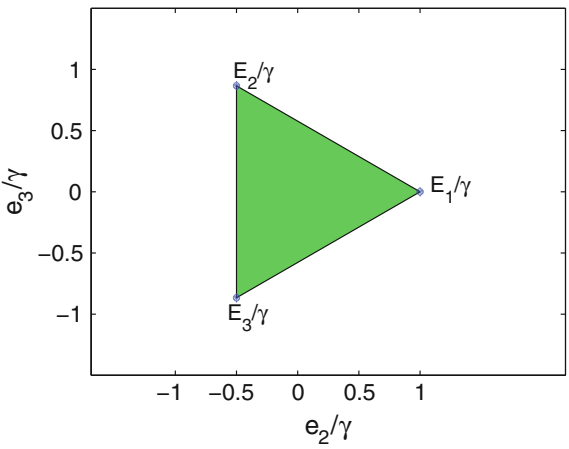


Fig. 1 Representation of the set $\mathscr{T}$ in the space of coordinates e_2, e_3. With some abuse of notation we represent the three matrices E_1, E_2 and E_3 in the same image. The convex envelope of $\mathscr{T}$ is represented by the closed *triangle* (*green* region in the online version of this proceeding)

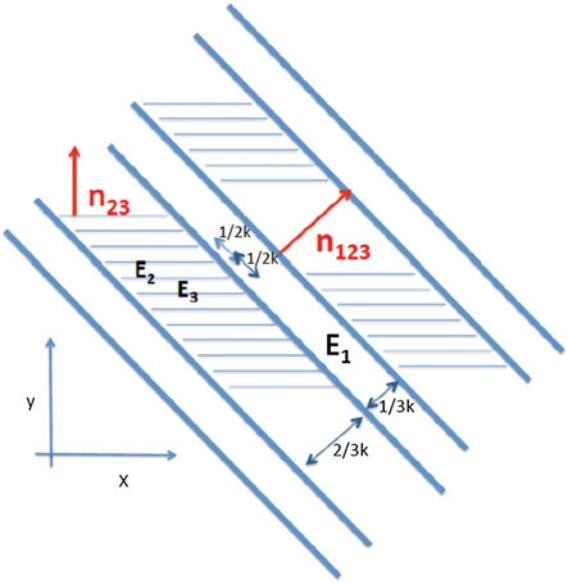


Fig. 2 Right: schematic representation of a possible microstructure associated to the TR transformation

the proof Theorem 1). As by-product of our relaxation result we gain some insight on the nature of the minimizing sequences for I. In Fig. 2 we sketch one element for the minimizing sequence $\{u_k\} \subset H^1_o(\Omega, R^2)$ for the problem $\inf_{H^1_o(\Omega,R^2)} I(u)$ corresponding to a given k. As the gradient of u_k oscillates, the symmetric part $(\nabla u_k + \nabla^T u_k)/2$ takes value in $\mathscr{T}$. Note that the volume fraction of each of the three martensitic variants involved in this construction corresponds to $\frac{1}{3}$. Therefore, we have

$$\inf_{H^1_o(\Omega,R^2)} I(u) = \min_{H^1_o(\Omega,R^2)} \overline{I}(u) = \overline{I}(0) = 0$$

and

$$u_k \rightharpoonup 0 \text{ in } H^1(\Omega, R^2).$$

The construction for u_k involves a laminate-within-laminate construction where bands occupied by the variants E_1 are matched with a further mixture composed of the two remaining variants E_2 and E_3 occurring at a smaller scale. The boundary layer occurring between at the interface separating the variant 1 with the mixture E_2/E_3 is schematically represented in Fig. 2 by a thicker line.

Acknowledgments The research of PC was partially supported by the ERC under the EU's Seventh Framework Programme (FP7/2007–2013)/ERC grant agreement no. 291053.

References

1. Acerbi, E., Fusco, N.: Semicontinuity problems in the calculus of variations. Arch. Rat. Mech. Anal. **86**, 125–145 (1984)
2. Ball, J.M., James, R.D.: Fine phase mixtures as minimizers of energy. Arch. Rat. Mech. Anal. **100**, 13–52 (1987)
3. Bhattacharya, K.: Comparison of the geometrically nonlinear and linear theories of martensitic transformation. Continuum Mech. Thermodyn. **5**, 205–242 (1993)
4. Bhattacharya, K.: Microstructure of Martensite. Oxford University Press, Oxford (2003)
5. Cesana, P.: Relaxation of multiwell energies in linearized elasticity and applications to nematic elastomers. Arch. Rat. Mech. Anal. **197**, 903–923 (2010)
6. Cesana, P., Porta, M., Lookman, T.: J. Mech. Phys. Sol. **72** (2014)
7. Ciarlet, P.G.: Math. Elast., vol. 1. Elsevier, Amsterdam (1988)
8. Dacorogna, B.: Direct Methods in the Calculus of Variations, 2nd edn. Springer, Heidelberg (2008)
9. Desimone, A., Dolzmann, G.: Arch. Rat. Mech. Anal. **161** (2002)
10. Jacobs, A.E., Curnoe, S.H., Desai, R.C.: Mater. Trans. **45** (2004)
11. Kohn, R.: Cont. Mech. Thermodyn. **3** (1991)
12. Lookman, T., Shenoy, S.R., Rasmussen, K.O., Saxena, A., Bishop, A.R.: Phys. Rev. B **67** (2003)
13. Müller, S.: Variational methods for microstructure and phase transitions. In: Bethuel, F., Huisken, G., Müller, S., Steffen, K., Hildebrandt, S., Strüwe, M. (Eds.) Proceedings of C.I.M.E. Summer School Calculus of Variation and Geometric Evolution Problems, vol. 1713, Cetraro 1996, Springer, LNM (1999)
14. Patching, S., Cesana, P., Rueland, A.: In preparation
15. Pedregal, P.: Europ. J. Appl. Math. **4** (1993)
16. Pipkin, A.C.: Quart. J. Mech. Appl. **44** (1991)
17. Porta, M., Lookman, T.: Acta Mater. **61** (2013)

Capillary Surfaces Modeling Liquid Drops on Wetting Phenomena

Rafael López

Abstract The aim of the present work is to relate the shape of a liquid drop in some contexts on capillarity and wetting with the surfaces that are mathematical models of these droplets. When a liquid drop is deposited on a support substrate, we are interested whether the geometry of the support imposes restrictions to the possible configurations of the droplet. Recently there is a progress in experiments done for liquid drops deposited on (or between) spherical rigid bodies, an assembly of cylinders and on a cone that allows to consider new theoretical problems in the field of capillary surfaces. We exploit the symmetries of these supports to apply the maximum principle of elliptic equations concluding that in some cases the drop inherits part of the symmetries of the support.

Keywords Capillarity · Wetting · Mean curvature · Delaunay surfaces · Free boundary problem · Tangency principle

1 Introduction

1.1 A Brief Approach to Capillarity and Wetting

Following [8], capillarity studies the interfaces between two immiscible phases and wetting refers how a liquid deposited on a solid (or liquid) substrate spreads out. Capillarity and wetting appear in a variety of industrial and engineering processes (e.g., automobile manufacturing, textile production, ink-jet printing, or colloid-polymer mixtures) where it is of interest to understand the physical and chemical behavior of a fluid. Many experiments consist of modifying the characteristics of the liquid and the solid until to attain the desirable wetting/spreading properties [5, 8]. A simple, but illustrative example, is when a given amount of an incompressible liquid

R. López (✉)
Departamento de Geometría y Topología Instituto de Matemáticas (IEMath-GR), Universidad de Granada, 18071 Granada, Spain
e-mail: rcamino@ugr.es

B. Anderssen et al. (eds.), *The Role and Importance of Mathematics in Innovation*, Mathematics for Industry 25,
DOI 10.1007/978-981-10-0962-4_12

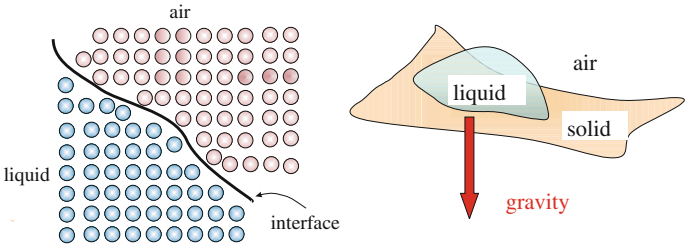


Fig. 1 *Left* an interface is the boundary of two homogenous systems with different physical and chemical properties. *Right* a liquid droplet deposited on a substrate under ideal physical conditions

is deposited on a solid substrate. Under idealized conditions (non-roughness, constant pressure and temperature, purity or low viscosity), the only forces acting on the liquid molecules are of order of a few nanometers and are determined by the van der Waals and electrostatic interactions. These forces are balanced except for the molecules on the liquid–air interface S of the drop which, to be in contact with the air and solid phases, are mainly attracted inward and to the sides so that the attraction energy at the interface is less than in the interior. See Fig. 1, left. Under the above physical assumptions, the total energy E of the system is $E = E_S + E_A + E_G$ where E_S is the surface tension, E_A is a wetting energy, and E_G the gravitational energy (Fig. 1, right). The energy E_S is the surface energy to create the interface S and is proportional to the number of interfacial molecules, that is, the surface area of S. The energy per area of S is called the *surface tension* σ. Similarly, E_A is the energy by the adhesion of the droplet on the solid phase which is also proportional to the number of molecules of the droplet in contact with the solid. Finally, E_G represents the weight of the drop and can be written as an integral $\int_V gz$, where V is the volume of the drop, g is the gravitational constant and z is the height at a point of S with respect to a reference system. In this physical system, there are three different phases present, namely, liquid–air, solid–liquid, and solid–air phase, and the three corresponding surface tensions σ, σ_{SL} and σ_{SA}, respectively: see Fig. 2, left.

In thermodynamic equilibrium, the interface S is free to change the shape in order to minimize its total free energy E. Assuming that the volume V of the drop remains fix (no evaporation), or in other words, if V is a Lagrange multiplier, and according

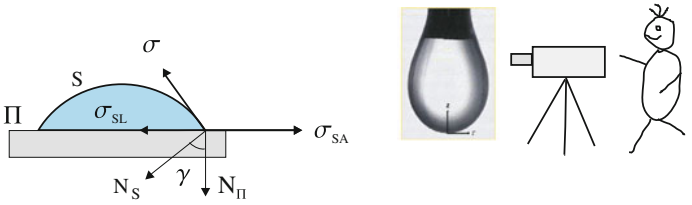


Fig. 2 *Left* the contact angle γ between S and Π and the equilibrium between the three surface tensions. *Right* the pendant drop method to measure the surface tension for an axisymmetric droplet

to the principle of virtual work, the system will be in equilibrium if the energy E attains a critical point in the position of S. Then S satisfies the well-known Laplace equation

$$(P_L - P_A) + (\Delta d)\, g\, z = \left(\frac{1}{R_1} + \frac{1}{R_2}\right)\sigma = 2H\sigma. \tag{1}$$

Here $P_L - P_A$ is the difference between the liquid pressure P_L under S and the air pressure P_A just above S, Δd is the difference of densities between the liquid and air phases and H is the mean curvature of S. The mean curvature H at each point of S is defined by $2H = 1/R_1 + 1/R_2$, where R_i are the curvature radii. Because we are assuming ideal conditions, $P_L - P_A$ is constant, as well as, Δd, g and σ. In particular, the mean curvature H is a linear function of z, that is, for each point $(x, y, z) \in S$, we have $H(x, y, z) = \lambda z + \mu$, where $\lambda = g\Delta d/(2\sigma)$, $\mu = (P_L - P_A)/(2\sigma)$. Usually there are two extra boundary conditions. The first one supposes that the liquid–solid phase is prescribed, that is, the part that the drop wets the solid is confined in a fixed region so the boundary ∂S of S is a prescribed curve. A second and more natural scenario is assuming that the droplet can move freely on the substrate Π (free boundary condition). In this situation, S satisfies the so-called Young equation

$$\cos\gamma = \frac{\sigma_{SA} - \sigma_{SL}}{\sigma}, \tag{2}$$

where γ is the angle that makes S with the liquid–solid–air contact line ∂S. Because the three surfaces tensions are constant, the Young equation (2) establishes that *the contact angle γ between the drop and the substrate is constant along ∂S* [11, 16]. Here γ is the angle between the unit normal vectors N_S and N_Π of S and Π, respectively: $\cos\gamma = \langle N_S, N_\Pi\rangle$, where N_S points to the liquid drop and N_Π points outward the drop.

In a specific problem it is necessary for the prediction of the magnitude of the capillary forces for eliminating or minimizing undesirable events, for example, an uncontrolled growth of agglomeration of particles or an abrupt change of the flow behavior of a fluid. According to this, the wetting state of the fluid is determined once the three surface tensions are known. In general, it is difficult to compute all of them, although the difference $\sigma_{SA} - \sigma_{SL}$ in (2) is a property of the solid and independent off the liquid used. Thus the interest focuses to compute the surface tension σ which is obtained from the Laplace equation (1) once calculated H or from the Young equation (2) if the contact angle γ is known.

1.2 The Measurement of the Surface Tension

Among the numerous measurement techniques of the surface tension σ, we describe the sessile and pendant drop method [1]. A drop is sitting (or hanging) on a horizontal plane which takes aside-view photographs of the profile and we use a snapshot to

determine the shape of S (or the angle γ) by comparing the actual shape of the drop with theoretical simulations based on the parameter σ; see Fig. 2, right. However in order to use the Young equation (2), it is actually difficult to compute explicitly the contact angle γ because the liquid is easily contaminated. The other (and more common) procedure consists to determine the mean curvature H adequating the profile shape of the drop to a well-controlled geometry and extracting σ from the Laplace equation (1). The mean curvature H of a surface $z = u(x, y)$ in Euclidean space $\mathbb{R}^3$ satisfies

$$(1 + u_y^2)u_{xx} - 2u_x u_y u_{xy} + (1 + u_x^2)u_{yy} = 2H(1 + u_x^2 + u_y^2)^{3/2}. \tag{3}$$

We observe that Eq. (3) is a PDE of order two, which cannot be integrated, even if H is constant, and only be numerically approximated by analytic methods. Assuming a small scale (wetting) or that the typical size of the meniscus is much smaller than the capillary length (capillarity), the surface tension dominates the gravitational force, so the gravity can be neglected. Thus $g = 0$ in the Laplace equation (1) and we deduce that the mean curvature H is constant. As a consequence we can affirm that *the liquid–air interface S of a liquid droplet is modeled by a surface in Euclidean space where the mean curvature is the same at every point and makes a constant contact angle with the support substrate*. Surfaces with zero mean curvature everywhere ($H = 0$) are called minimal surfaces and they appear when the pressures coincide in both sides of S. Constant mean curvature surfaces are easily obtained when we dip in and out a closed wire in a container with soapy water. The soap film spanning by the wire is a minimal surface because there is not pressure difference across it. However if the wire traps air inside it, or if we blow air on it making a bubble, then there is an enclosed volume, the pressure difference is nonzero (but constant), and the surface has nonzero constant mean curvature.

Therefore experimentalists need to simplify Eq. (3) and the usual idea is assuming symmetric shapes so the discrete computational procedures developed to simulate the mathematical behavior of these processes can be fast and manageable. In this sense, it would be useful to reduce this equation into an ODE if, owing to symmetries, the equation depends only on one coordinate. The most common situation is assuming axisymmetric solutions of (3), that is, S is a surface of revolution. If $u = u(r)$ is the distance to the rotation axis, a first integration of (3) is

$$Hu^2 - \frac{u}{\sqrt{1 + u'^2}} = c \tag{4}$$

for some $c \in \mathbb{R}$. From this equation, we can solve some cases: if $c = 0$, the solution of (4) is the circle $u(r) = \sqrt{1 - H^2 r^2}/H$ and S is a sphere of radius $1/|H|$; if u is a constant function, then the solution is $u(r) = 1/(2H)$ (for $1 + 4Hc = 0$) and S is a cylinder of radius $1/(2|H|)$; if $H = 0$, then $c = m^2 > 0$, $u(r) = m\cosh(r/m)$ and S is a catenoid. However for arbitrary c, the solutions of (4) cannot integrate completely and they can only be represented by elliptic integrals. The profile curves of the solutions of (4) are mathematically characterized to be the roulettes of the focus

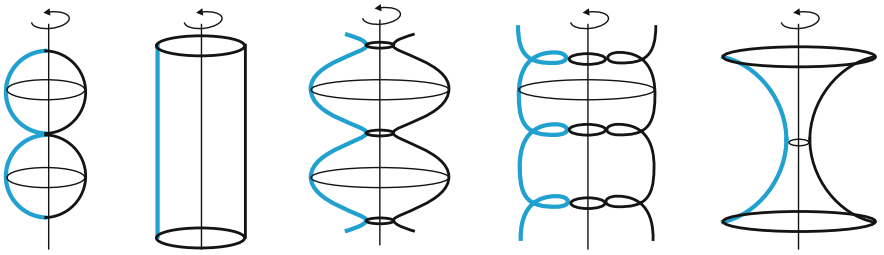


Fig. 3 Delaunay surfaces. From *left* to *right* sphere, cylinder, unduloid, nodoid, and catenoid

of a conic and the surfaces are called Delaunay surfaces [9, 10]. Besides spheres and cylinders, they are unduloids, nodoids, and catenoids: see Fig. 3. Usually, experiments utilize symmetric devices to sit or hang a droplet from a circular opening where *the observed interface is assumed to be a surface revolution*. In general, pendant drops are more utilised because they are easily controllable. Once we know that the interface is rotational, determining the geometry of the drop consists to capture and digitalize its image, extracting its contour, smoothing the profile, and comparing the shape with the theoretical Delaunay surfaces (Fig. 4). Finally, a software (for example, a Runge–Kutta method, a technique based on finite elements or the Surface Evolver) works to compute the mean curvature H. This measurement method is simple and it does not require a sophisticated machinery or any special cleanliness of the solid substrate.

In contrast to the assumption that a droplet hanging for a circular opening is axisymmetric (independently with or without gravity), and from the theoretical viewpoint, the shape of a surface with constant mean curvature (cmc surface in short) in Euclidean space spanning a circle $\mathbb{S}^1$ is not well known up today and only some partial results ensure that a compact cmc surface in $\mathbb{R}^3$ spanning $\mathbb{S}^1$ is a planar disk or a spherical cap. For the state of the art in this topic, see [18]. In the free boundary problem, it is unknown whether the geometry of the substrate affects to the geome-

Fig. 4 *Left* description of the typical apparatus of the pendant drop method. *Right* a pendant drop is modeled by an axisymmetric surface by adjusting its contour

try of a cmc surface supported on it, for example, if it inherits its symmetries. First mathematical results were obtained by Wente in [34] assuming embeddedness of the surface.

2 Capillary Surfaces Supported on Spheres, Cylinders, Cones, and Wedges

Recently there is a great interest in the study of liquid drops deposited on (or between) configurations formed by spherical rigid bodies, an assembly of cylinders, cones or planes because this variety of systems may be found like a crystallization, agglomeration, phase sintering, liquid foams, and emulsions [15, 17, 23, 27, 33]. Moreover, the improvement of the numerical analysis methods as well as the modeling software allows to consider new theoretical problems in capillarity and wetting. When the size of the liquid drop is very small, the effect of gravity is negligible and no other force is considered. In such a case, the interface S has the same mean curvature H everywhere. We need to again model the liquid bridges as Delaunay surfaces where the geometry associated is relatively simple or at least giving conditions that ensure that S is rotational. In this section, by a *capillary surface* we mean a cmc surface S with free boundary on a substrate Π and S makes a constant contact angle with Π along its boundary ∂S. Since we are considering bounded droplets, we also suppose that S is compact. The symmetry of the mentioned supports in this section allows to get (at least theoretically) explicit examples of pieces of Delaunay surfaces that are capillary surfaces. Some examples appear in Fig. 5, where the support Π is a sphere, a circular cylinder and a circular cone and the rotation axis of S coincides with the one of Π.

In what follows, we show some results on the symmetry of a capillary surface when the support substrate is a sphere, a right cylinder, a cone, and a wedge. See [19–22].

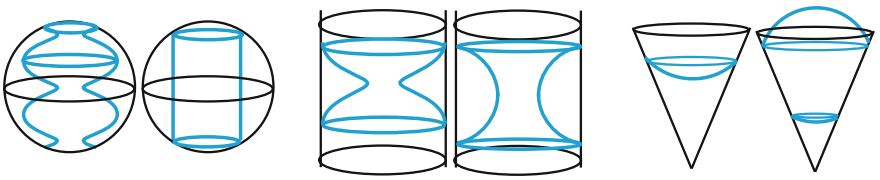


Fig. 5 Pieces of Delaunay surfaces that make a constant contact angle with a *sphere*, a *circular cylinder*, and a *cone*

2.1 Droplets on a Sphere

Consider a cmc surface S whose boundary lies on a sphere, which we suppose to be the unit sphere $\mathbb{S}^2$ and denote by $\mathbb{B}^3$ the unit ball enclosed by $\mathbb{S}^2$. Previous results on capillary surfaces on $\mathbb{S}^2$ included in $\mathbb{B}^3$ were obtained assuming that the contact angle is constant [12, 29, 30].

Theorem 1 *Let S be an embedded cmc surface on $\mathbb{S}^2$ whose boundary ∂S is included in a hemisphere $\mathbb{S}^2_+$. Suppose S is a capillary surface. Let W be the 3-domain bounded by $S \cup \Omega$, where $\Omega \subset \mathbb{S}^2_+$ is the domain bounded by ∂S. If $W \subset \overline{\mathbb{R}^3 - \mathbb{B}^3}$ or $S \subset \mathbb{B}^3$, then S is part of a sphere.*

This result extends if we replace the capillary condition by assuming that the boundary ∂S is a circle. In such a case and when $W \subset \overline{\mathbb{R}^3 - \mathbb{B}^3}$, we add the hypothesis that the mean curvature H satisfies $|H| \geq 1$.

2.2 Droplets on a Right Cylinder

By a right cylinder we mean $\Sigma = C \times \mathbb{R}$, where $C \subset \mathbb{R}^2$ is a simple planar closed curve. The cylinder is said to be circular if C is a circle. The cylinder Σ determines two domains in $\mathbb{R}^3$, namely, the inside and the outside, that is, $\Omega \times \mathbb{R}$ and $\mathbb{R}^3 \backslash \overline{\Omega \times \mathbb{R}}$, respectively, where $\Omega \subset \mathbb{R}^2$ is the bounded domain by C. Consider a capillary surface S on Σ that lies in one side of Σ. A first question to elucidate is if the boundary ∂S is a curve nullhomotopic in Σ or if ∂S is homotopic to C. For example, the first setting could occur if the volume of S is very small, and the second one when a cylindrical tube is introduced in a container of liquid and the liquid rises up by capillarity. In the latter one, we ask if S is a graph $z = u(x, y)$ on Ω.

Theorem 2 *Let Σ be a right cylinder and let S be an embedded capillary surface on Σ such that $S \subset inside(\Sigma)$.*

1. *If ∂S is homotopic to C, then S is a graph on Ω. If Σ is a circular cylinder, then S is a planar disk or a spherical cap.*
2. *If $\partial S = C_1 \cup C_2$ and each C_i, $(i = 1, 2)$ is homotopic to C, then S has a symmetry with respect to a plane orthogonal to the axis.*
3. *If Σ is a circular cylinder and ∂S is contained in a half cylinder of Σ, then S has two mutually planes of symmetry and S is a topological disk.*

In the item 3, by a half cylinder of Σ we mean one of the two components remains when we intersect Σ by a plane containing the rotation axis.

In case that S has zero mean curvature, we have a strong result under the hypothesis that the surface is immersed.

Theorem 3 *Let S be capillary minimal surface on Σ such that $S \subset inside(\Sigma)$. If ∂S is a graph on C, then S is a horizontal planar domain.*

2.3 Droplets on a Cone

Consider $\Omega \subset \mathbb{S}^2$ a simply connected domain of $\mathbb{S}^2$ and included in a hemisphere of $\mathbb{S}^2$. If $\Gamma = \partial\Omega$, the cone determined by Γ is defined as $\Sigma = \{\lambda p : \lambda > 0,\, p \in \Gamma\}$, that is, the set of all rays starting from the origin O through all points of Γ. If Γ is a circle, we say that Σ is a circular cone. The inside of the cone Σ is the corresponding 3-domain $\{\lambda p : \lambda > 0,\, p \in \Omega\}$.

We consider a capillary surface S whose boundary lies on Σ and contained in the inside of Σ. As in the case of a right cylinder, we do not know whether ∂S is nullhomologous in $\Sigma - \{O\}$ or if ∂S is homotopic to Γ in $\Sigma - \{O\}$ and S has a one-to-one central projection on Ω (a radial graph), that is, each ray starting from the vertex intersects S one point at most. See Fig. 6, left. We obtain

Theorem 4 *Let S be an embedded capillary surface supported on a cone Σ and let us fix N the unit normal vector field of S pointing towards the liquid drop. If $H \leq 0$, then S is a radial graph and the boundary ∂S has only one connected component which is homologous to Γ in $\Sigma - \{O\}$. In the particular case that the cone is circular, then S is a planar disk or a spherical cap.*

In other words, Theorem 4 says that the non-positivity of H implies that S is a topological disk and that there are no capillary bridges between the walls of Σ. As a consequence, and dropping the assumption on the sign of H, we have (Fig. 6, right)

Corollary 1 *If S is a capillary surface on a circular cone Σ such that the contact angle γ satisfies $\gamma \leq (\pi - \varphi)/2$, being φ the amplitude of Σ, then S is a planar disk or a spherical cap.*

In this case, the hypothesis on γ implies $H \leq 0$: this is a consequence of comparing S with spherical caps or planar disks having the same mean curvature and the same contact angle with Σ.

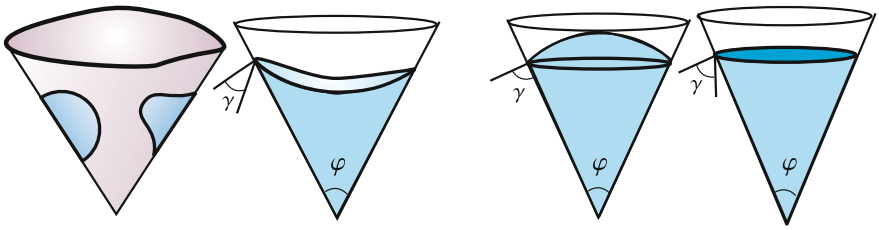


Fig. 6 *Left* possible configurations of a liquid drop deposited on a cone. *Right* spherical caps and planar disk are examples of capillary surfaces on a circular cone

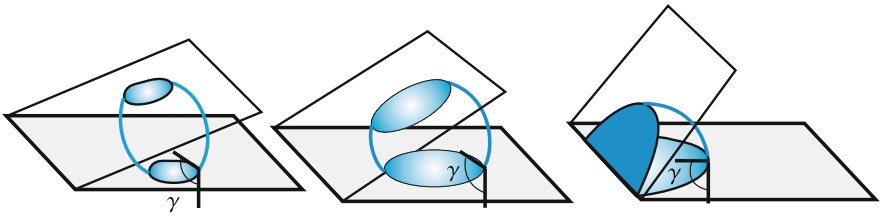


Fig. 7 *Left* capillary surfaces on a wedge. *Right* a spherical cap meeting orthogonally the walls of a wedge

2.4 Droplets on a Wedge

If we intersect appropriately, a Delaunay surface by two orthogonal planes $\Pi_1 \cup \Pi_2$ to the rotation axis, we obtain a capillary surface contacting $\Pi_1 \cup \Pi_2$ with the same contact angle. It is known by experiments that only some pieces of Delaunay surfaces are physically realized, that is, only some surfaces are stable in the sense that the second variation of the energy E is non-negative. Early results of Vogel and Athanassenas prove that the only stable capillary surfaces connecting two parallel planes are rotational surfaces and that if the contact angle γ is $\pi/2$, then the half-sphere and the cylinder are the only possibilities [4, 32]. However, the problem is far to be completely known for a general contact angle or other assumptions replacing stability [2, 6, 14, 24, 35].

A similar situation occurs when Π_1 and Π_2 are not parallel planes. In this case, the 3-domain determined by $\Pi_1 \cup \Pi_2$ is called a wedge. For this support, there are explicit examples of capillary surfaces when we place a sphere centered in the plane bisecting the wedge ($\gamma > \pi/2$), or if the center lies in the axes of the wedge ($\gamma = \pi/2$). See Fig. 7. A first question posed is on the existence of capillary surfaces with cylindrical topology connecting Π_1 and Π_2: see [7, 25, 26]. Under this context and $\gamma = \pi/2$ (Fig. 7, right), we prove

Theorem 5 *Consider a cmc surface S on a wedge with contact angle $\gamma = \pi/2$. If S is stable or S is embedded, then S is part of a sphere centered at the vertex.*

3 The Proof Methods

Motivated by experiments on wetting and capillarity, we assume that the interface of a droplet is an embedded surface. In our context, and since our surfaces are compact, embeddedness is equivalent to say that the surface has not self-intersections. In the theory of embedded cmc surfaces, one of the main ingredients in the proofs is the Alexandrov reflection principle. Alexandrov proved that the sphere is the only embedded closed to cmc surface [3]. Although this result was expected, the novelty

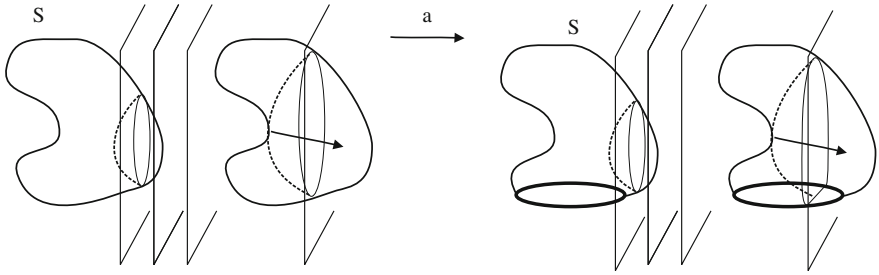


Fig. 8 The Alexandrov reflection method. *Left* the closed embedded case. *Right* the circular boundary case

came from the proof, where the very surface is utilized as a barrier with itself to obtain the desired result. This idea has been extensively utilized not only in geometry but also in PDE theory, starting with the breaking paper of Serrin [31]. We briefly explain the Alexandrov method. The mean curvature equation (3) is elliptic but not linear. However if u_1 and u_2 are two solutions of (3), the difference function $u = u_1 - u_2$ satisfies a *linear* elliptic equation $Lu = 0$ and we can apply the maximum principle [13]. In the context of cmc surfaces, this result is known as the tangency principle which asserts that if S_1 and S_2 are two surfaces with the same constant mean curvature, which are tangent at a point $p \in S_1 \cap S_2$ and S_1 lies in one side of S_2 around p, then S_1 and S_2 coincide in a neighborhood of p, and by extension of the argument, S_1 and S_2 coincide in a common open and closed set [18]. For the proof, let S be an embedded closed cmc surface and let us fix a direction $\mathbf{a} \in \mathbb{R}^3$. Consider a plane coming from infinity and orthogonal to $\mathbf{a}$ until arriving the first contact point with S: Fig. 8, left. Next, we follow moving the plane and reflecting the surface that lies behind the plane until the first time that the reflected surface (with respect to a plane $P_{\mathbf{a}}$) reaches the initial surface. In the touching point between both surfaces, the tangency principle implies that the reflected surface and the part of the surface in that side of $P_{\mathbf{a}}$ must coincide, proving that $P_{\mathbf{a}}$ is a plane of symmetry of S. Doing the same argument for all spatial directions $\mathbf{a}$, we conclude that S must be a round sphere.

In case the boundary of S is a circle, we need to assume that S lies in one side of the plane containing ∂S, see Fig. 8, right. This prevents that the first contact point may occur between an interior point with the boundary ∂S because in such a case, the reflected surface and S are not tangent at the first touching point and we cannot utilize the tangency principle.

In each one of the support substrates considered in Sect. 2, we have different possibilities of choices of planes to start with the reflection method. We explain in each case [19–22].

1. Suppose that the support is a sphere $\mathbb{S}^2$ and that ∂S is a circle in $\mathbb{S}^2_+$ (Theorem 1). Recall that in this case, we are assuming $|H| \geq 1$. Let Π be the horizontal plane containing the center O of $\mathbb{S}^2$. A first step is proving that if W lies outside $\mathbb{B}^3$, then $O \notin W$. On the contrary, consider the uniparametric family of spherical caps

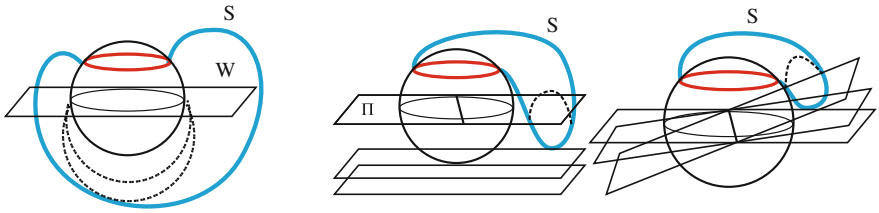


Fig. 9 Reflection method for a capillary surface droplet supported on a sphere

of radius bigger than 1 below Π and with the common boundary to be the circle $\Pi \cap \mathbb{S}^2$. Starting from the radius $r = 1$, we increase the radius of these caps until the first contact with S: Fig. 9, left. At the contact point, the mean curvature of the cap, namely $1/r$, must be bigger than $|H|$, which it is not possible because $|H| \geq 1$. As a conclusion, if W lies outside $\mathbb{B}^3$, then $O \notin W$. The reflection process starts with horizontal planes coming from below until we reach S (Fig. 9, right). Next, we follow moving up the plane and reflecting. Since ∂S lies in the upper hemisphere, there is not a touching point before arriving to the plane Π since, on the contrary, there would be a horizontal plane of symmetry: a contradiction because $\partial S \subset \mathbb{S}^2_+$. Once arrived to the origin, we fix a horizontal straight line $L \subset \Pi$ passing through O. Let us replace the above planes by a family of planes all containing L (Fig. 9, right). Then we go rotating the plane and we follow the reflection method until the first touching point p. If p is an interior point, a standard argument implies that the plane is a plane of symmetry, so of ∂S. If p is a boundary point, then the plane is a plane of symmetry of ∂S. Repeating this argument for any horizontal straight line L through the center of $\mathbb{S}^2$, we conclude that S is a spherical cap.

In case that S is a capillary surface, the only difference in the above argument is that if the first touching point p is a boundary point (necessarily with respect to a plane containing L), the condition on the constancy of the contact angle implies that the reflected surface and the initial one are tangent at p. Thus we apply the (boundary version) tangency principle [13] concluding that the plane is a plane of symmetry of the surface.

2. Suppose that the support is a right cylinder $\Sigma = C \times \mathbb{R}$. In the item 1 of Theorem 2, the reflection method uses a family of orthogonal planes to a vertical line and coming from infinity (Fig. 10, left). In case of existence of a horizontal plane where the reflected surface touches the first time with the initial surface at some interior point, then this plane is a plane of symmetry. This is a contradiction with that ∂S is a curve homotopic to C. If the first contact point occurs at a boundary point, the condition on the constancy of the contact angle implies that the initial and the reflected surface are tangent at that point, and the proof works. For the item 2, the argument is similar.

   For the item 3, and because ∂S lies in a half cylinder, then ∂S is nullhomotopic in Σ. Thus S together a domain of Σ bounds a 3-domain $W \subset \mathbb{R}^3$. A first step

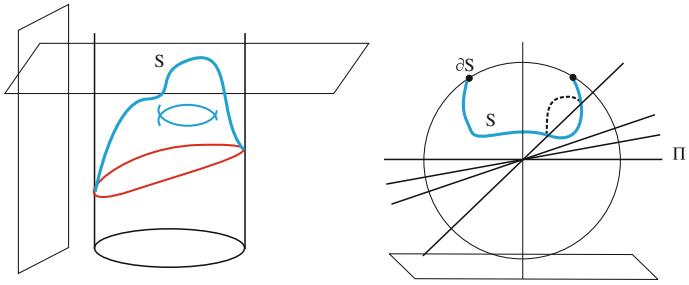


Fig. 10 Reflection method for a capillary surface supported on a right cylinder. *Right* a *top view* of Σ

consists to apply the reflection method with a uniparametric family of planes orthogonal to the rotation axis. Hence we obtain a first plane of symmetry P_1 of S. Let Π be the plane containing the rotation axis that leaves in one side ∂S. We now use a uniparametric family of planes parallel to Π and all of them lie in the other side of Π (not containing ∂S). The reflection method works until we arrive to the very plane Π (Fig. 10, right). The hypothesis on ∂S to be contained in a half cylinder prevents the existence of a first contact point. At this position, we replace the planes by a family of planes containing the axis. We follow the reflection method by rotating these planes until the first (interior or boundary) contact point, obtaining a new plane of symmetry P_2 of S. The plane P_2 contains the axis so P_2 is orthogonal to P_1. Because S is symmetric by these orthogonal planes P_1 and P_2, then S is a topological disk.

3. In the case of a cone, the reflection process with respect to planes is substituted by a spherical reflection method, which appeared first in [25] replacing inversions about a one parameter family of spheres all centered at the center O of $\mathbb{S}^2$. Although an inversion does not preserve H, there is a certain control of the mean curvature of the inverted surface in order to use the tangency principle. Exactly if $\mathbb{S}^2_r \subset \mathbb{R}^3$ is the sphere of radius r centered at O, the spherical reflection about $\mathbb{S}^2_r$ is the inversion mapping defined by

$$\phi_r : \mathbb{R}^3\backslash\{O\} \to \mathbb{R}^3\backslash\{O\}, \quad \hat{p} := \phi_r(p) = \frac{r^2}{|p|^2} p.$$

Let H be the mean curvature of S with respect to a unit normal vector field N. Denote by $\hat{S}_r$ the spherical reflection of S about ϕ_r and consider on $\hat{S}_r$ the orientation

$$\hat{N}(\hat{p}) = N(p) - \frac{2\langle N(p), p\rangle}{|p|^2} p.$$

Then the mean curvature of $\hat{S}_r$ is

$$\hat{H}(\hat{p}) = \frac{H|p|^2 + 2\langle N(p), p\rangle}{r^2}. \tag{5}$$

We start the spherical reflection method from spheres $\mathbb{S}^2_r$ with r sufficiently big until the first contact point p_0 with S. Because N points are inside the liquid, then $\langle N(p_0), p_0\rangle < 0$. We have from (5) that $\hat{H}(p_0) \leq H|p_0|^2/r^2 < H$, where we use $H \leq 0$. Following the reflection across inversions and using the assumption on the non-positivity of H, we conclude that there is not a contact point between the inverted surface with the part of S inside $\mathbb{S}^2_r$, which proves that S is a radial graph on Ω. For the second part of Theorem 4, we use that the only cmc surface in $\mathbb{R}^3$ that is invariant by an inversion about a sphere is an open set of a sphere or a plane.

4. We only prove Theorem 5 when S is an embedded surface. First we extend the Ros formula [28] proving that if a compact embedded surface with no necessarily constant mean curvature H meets a wedge orthogonally, then

$$\int_S \frac{1}{H}\, dM \geq 3V, \tag{6}$$

where V is the volume of S and the equality holds if and only if S is part of a sphere. The proof of (6) involves the Reilly formula for a solution of PDE with Dirichlet and Neumann boundary conditions and the classical Minkowski formula

$$\int_S (1 + H\langle N, x\rangle)\, dS = -\frac{1}{2}\int_{\partial S} \langle \nu, x\rangle\, ds,$$

where ν is the inward unit conormal along ∂S. After a rigid motion, the orthogonality intersection condition means that $\langle \nu, x\rangle = 0$ and as H is constant, we get $A - 3HV = 0$, where A is the area of S. This implies equality in (6) and the result follows.

4 Conclusions

In the present paper we have discussed under what conditions some geometric configurations of a liquid droplet in thermodynamic equilibrium is a surface of revolution. Our motivation comes from the fact that experiments devoted to compute the surface tension σ of a liquid (e.g., the pendant drop method) assume previously that if the boundary of the air–liquid interface is symmetric, or if the drop is supported on a highly symmetric substrate, the liquid drop receives the same symmetries. In recent years there is a great progress in the creation of new materials and experimentation at nanometer and microscopic scales of fluids deposited between configurations of

spheres, cylinders, and planes. In some industrial experiments, there exist processes of crystallization and agglomeration that require the knowledge of the effects of the capillary forces of the liquids bridges connecting these solids and avoiding an abrupt change in the liquid shape, or preventing undesirable overflowing events. To quantify and estimate these forces, the mathematical models for droplets and liquid bridges are cmc surfaces which are assumed to be surfaces of revolution because the analytic expression of the mean curvature equation (3) in the axisymmetric case (4) is easier. Recent progress in experiments with a wider variety of morphologies on the substrate has given a new boost in the theoretical study that it was not previously considered.

Our results show that if these drops are modeled by a surface with constant mean curvature and under assumptions of embeddedness, then the droplet inherits some symmetries of the support substrate Π when Π is a sphere, a right cylinder, a cone, or a wedge. This allows to provide a mathematical understanding of why the shapes of these drops are axisymmetric. These results provide us new directions of investigation, for example, assuming that the droplet has self-intersections which means that in the fluid there may appear empty chambers of liquid or that the droplet does not lie completely in one side of the substrate.

Acknowledgments The author has been partially supported by the MINECO/FEDER grant MTM2014-52368-P.

References

1. Adamson, A.W., Gast, A.P.: Physical Chemistry of Surfaces. Wiley, New York (1982)
2. Ainouz, A., Souam, R.: Stable capillary hypersurfaces in a half-space or a slab (2014). arXiv:1411.4241
3. Alexandrov, A.D.: Uniqueness theorems for surfaces in the large V. Vestnik Leningrad Univ. Math. **13**, 5–8 (1958). (English translation: AMS Transl. **21**, 412–416 (1962))
4. Athanassenas, M.: A variational problem for constant mean curvature surfaces with free boundary. J. Reine Angew. Math. **377**, 97–107 (1987)
5. Bonn, D., et al.: Wetting and spreading. Rev. Mod. Phys. **81**, 739–805 (2009)
6. Choe, J., Koiso, M.: Stable capillary hypersurfaces in a wedge. Pacific J. Math. **280**, 1–15 (2016)
7. Concus, P., Finn, R.: Discontinuous behavior of liquids between parallel and tilted plates. Phys. Fluids **10**, 39–43 (1998)
8. de Gennes, P., Brochard-Wyart, F., Quéré, D.: Capillarity and Wetting Phenomena. Springer, New York (2004)
9. Delaunay, C.: Sur la surface de révolution dont la courbure moyenne est constante. J. Math. Pures Appl. **6**, 309–315 (1841)
10. Eells, J.: The surfaces of Delaunay. Math. Intell. **9**, 53–57 (1987)
11. Finn, R.: Equilibrium Capillary Surfaces. Springer, Berlin (1986)
12. Fraser, A., Schoen, R.: Sharp eigenvalue bounds and minimal surfaces in the ball. Invent. Math. **203**, 823–890 (2016)
13. Gilbarg, D., Trudinger, N.S.: Elliptic Partial Differential Equations of Second Order. Reprint of the 1998 edition. Springer, Berlin (2001)
14. Koiso, M., Palmer, B.: A uniqueness theorem for stable anisotropic capillary surfaces. SIAM J. Math. Anal. **39**, 721–741 (2007)

15. Kubalski, G.P., Napiorkowski, M.: A liquid drop in a cone-line tension effects. J. Phys.: Condens. Matter **12**, 9221–9229 (2000)
16. Langbein, D.: Capillary Surfaces. STMP, vol. 178. Springer, Berlin (2002)
17. Lian, G., Thornton, C., Adams, M.J.: A theoretical study on the liquid bridge forces between two rigid spherical bodies. J. Colloids. Interfaces Sci. **161**, 138–147 (1993)
18. López, R.: Constant Mean Curvature Surfaces with Boundary. Springer Monographs in Mathematics. Springer, Berlin (2013)
19. López, R.: Capillary surfaces with free boundary in a wedge. Adv. Math. **262**, 476–483 (2014)
20. López, R., Pyo, J.: Constant mean curvature surfaces with boundary on a sphere. Appl. Math. Comput. **220**, 316–323 (2013)
21. López, R., Pyo, J.: Capillary surfaces of constant mean curvature in a right solid cylinder. Math. Nachr. **287**, 1312–1319 (2014)
22. López, R., Pyo, J.: Capillary surfaces in a cone. J. Geom. Phys. **76**, 256–262 (2014)
23. Lukas, D., et al.: Morphological transitions of capillary rise in a bundle of two and three solid parallel cylinders. Phys. A **371**, 226–248 (2006)
24. Marinov, P.: Stability of capillary surfaces with planar boundary in the absence of gravity. Pac. J. Math. **255**, 177–190 (2012)
25. McCuan, J.: Symmetry via spherical reflection and spanning drops in a wedge. Pac. J. Math. **180**, 291–323 (1997)
26. Park, S.H.: Every ring type spanner in a wedge is spherical. Math. Ann. **332**, 475–482 (2005)
27. Rabinovich, Y.U., Esayanur, M.S., Moudgil, B.M.: Capillary forces between two spheres with a fixed volume liquid bridge: theory and experiment. Langmuir **21**, 10992–10997 (2005)
28. Ros, A.: Compact hypersurfaces with constant higher order mean curvatures. Rev. Mat. Iberoamericana **3**, 447–453 (1987)
29. Ros, A., Souam, R.: On stability of capillary surfaces in a ball. Pac. J. Math. **178**, 345–361 (1997)
30. Ros, A., Vergasta, E.: Stability for hypersurfaces of constant mean curvature with free boundary. Geom. Dedicata **56**, 19–33 (1995)
31. Serrin, J.: A symmetry problem in potential theory. Arch. Ration. Mech. Anal. **43**, 304–318 (1971)
32. Vogel, T.I.: Stability of a liquid drop trapped between two parallel planes. SIAM J. Appl. Math. **47**, 516–525 (1987)
33. van Honschoten, J.W., Tas, N.R., Elwenspoek, M.: The profile of a capillary liquid bridge between solid surfaces. Am. J. Phys. **78**, 277–286 (2009)
34. Wente, H.C.: The symmetry of sessile and pendent drops. Pac. J. Math. **88**, 387–397 (1980)
35. Wente, H.C.: The capillary problem for an infinite trough. Calc. Var. Partial Differ. Equ. **3**, 155–192 (1995)

Deformable Human Body Modeling from 3D Medical Image Scans

Taehyun Rhee, Patrick Lui and J.P. Lewis

Abstract Creating an accurate virtual human body model is challenging but required in many fields. This study presents a method to create 3D human body models from medical image scans. Visible light scanning of articulated 3D objects such as the human hand has limitation due to the self-occlusion of surfaces in certain poses. We present a complete system to create a deformable articulated human body volume from multiple 3D MRI scans of a living person, which can produce accurate volume deformation containing inner anatomical layers. The method combines technologies involving medical imaging, and computer vision, as well as computer graphics, to address the practical issues involved in producing detailed volume models from living human scans. The results provide an occlusion free person-specific 3D human body model, asymptotically accurate inner tissue deformations, and realistic volume animation of articulated movements driven by standard joint control estimated from the actual skeleton.

Keywords Registration · Deformation · Human modeling · Volume animation

1 Introduction

An anatomically accurate deformable 3D human body model including the bones, muscles, tendons, and other anatomical layers is challenging. Articulated body regions such as the human knee or hand are capable of a wide range of skeletal

T. Rhee (✉) · P. Lui · J.P. Lewis
School of Engineering and Computer Science,
Victoria University of Wellington,
Gate 6, KelburnParade, Wellington, New Zealand
e-mail: taehyun.rhee@ecs.vuw.ac.nz

P. Lui
e-mail: patrick.lui@ecs.vuw.ac.nz

J.P. Lewis
e-mail: john.lewis@ecs.vuw.ac.nz

B. Anderssen et al. (eds.), *The Role and Importance of Mathematics in Innovation*, Mathematics for Industry 25,
DOI 10.1007/978-981-10-0962-4_13

movements, resulting in complex deformation of the surrounding soft tissues. The difficulty in manually or algorithmically defining complex articulated body structures of an individual subject can be avoided by adopting a data-driven approach. Since scans of a living subject at multiple poses can be used as the training samples, accurate deformable models can be built from actual data. Also, a model constructed from living human scans reflects characteristics of the subject and provides personalized information, which is often essential to create a virtual clone for medical and other applications.

Volume data obtained from 3D medical image scans (e.g. MRI or CT) represents 3D interior anatomy. Translucent volume rendering can successively visualize anatomical layers without losing the overall context of the subject. Previous scan based approaches have focused on surface scans [1, 2] and deformation [3, 4]. This study describes a data-driven approach in the volume domain using appropriate deformation algorithms [6], resulting in accurate volume deformation informed by multiple scans of articulated body regions from a living person [7].

2 Articulated Volume Registration

One of the challenging issues in scan-based deformation is to obtain geometric correspondences across the samples. In case of medical image volumes, the geometrical information is represented by voxel properties without explicit geometric parameterizations, and creating iso-surfaces of each layer from in vivo MRI volumes is difficult due to poor delineation of different tissue layers. Working with volumetric data brings computational scaling issues, since the volume of data is considerably larger than the surface data. Volume registration methods for producing correspondence across multiple scans of different poses must deal not only with the volume of data, but also with the many degrees of freedom (DOFs) arising from both non-rigid tissue deformation and rotations of the underlying skeletal joints. In addition, the optimization must handle the strong local minima inherent in complex articulated subjects. There are also issues involving the use of in vivo MRIs that do not arise with cadavers or non-articulated subjects.

In our study [7], the issues described above are successively addressed, resulting in sophisticated solutions for quantifying and visualizing complex volume deformation arising from a wide range of skeletal movements of the human body. Differently posed volumes of real human body regions (e.g. the hand and knee) are obtained by MRI scans. Skeleton models of each pose sample are created by semi-automatic hierarchical bone volume registration. Then, the kinematic joint structures of each volume sample are estimated. In order to solve the correspondence problem across scans, a template volume is registered to each volume sample. The wide range of pose variations is first approximated by the volume blend deformation algorithm [6], providing proper initialization for subsequent non-rigid volume registration. The initialized volume is then automatically registered to the target volume while mini-

Fig. 1 From the *left*, the original image, images warped by 2D biharmonic clamped plate spline (CPS) and thin plate spline (TPS). Control point correspondences (*small squares*) are indicated as *yellow* (source) and *white* (target). The right most image is comparison of the CPS (*top*) with TPS (*bottom*) interpolating the same points. The CPS is both smooth and localized, decaying to zero at the boundary of the unit disc

mizing the mutual information or sum of squared intensity difference without relying on fiducial markers.

Non-rigid volume registration requires high DOFs warping functions for accurate registration. To address this, a locally adaptive registration algorithm that efficiently reduces the search domain and DOFs of the warping function is developed. The volume is hierarchically and spatially decomposed and dissimilar regions are locally and adaptively registered using deformation based on the clamped plate spline (CPS) [5, 7]. The CPS minimizes the standard spline curvature functional, but subject to having zero value and derivative on the boundary of the unit disc in $\mathbf{R}^n$ as in the Fig. 1. The derivation of the CPS resembles that of radial basis functions (RBFs), with the solution being a weighted sum of the Green's function of the solution's differential operator, but the Green's function in this case is not a radially symmetric function but instead depends implicitly on the location relative to the origin. The function for the biharmonic case in three dimensions is:

$$\begin{cases} G(x, y) & = \|x - y\|(A + 1/A - 2) \\ A(x, y) & = \frac{\sqrt{\|x\|^2\|y\|^2 - 2x^T y + 1}}{\|x - y\|} \end{cases} \tag{1}$$

The x component of the resulting interpolated deformation at a point $\mathbf{p}$ is:

$$d_x = \sum w_k G(\mathbf{p}, \mathbf{c}_k) \tag{2}$$

(and similarly for the y, z components) where $\mathbf{c}_k$ are the locations of the feature points to be interpolated.

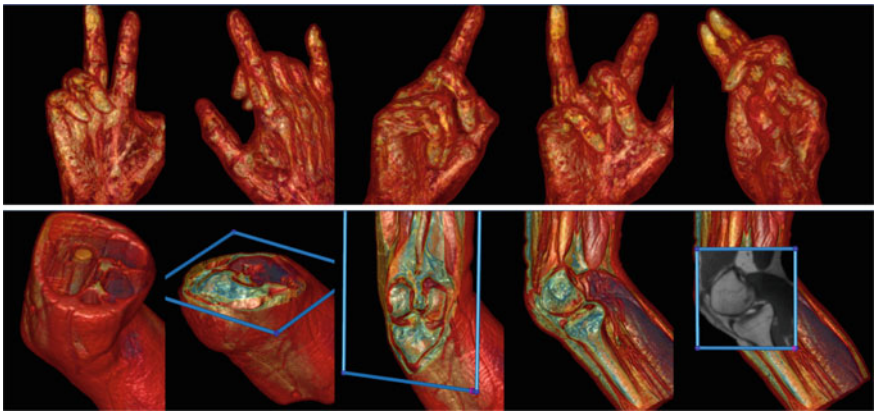


Fig. 2 Volume deformation: MRI volumes of the *human hand* and *knee* are deformed to arbitrary poses; translucent volume rendering visualizes the anatomical layers without losing the context of the subject

3 Example Based Volume Deformation

Accurately registered volume samples are now used to compute volumetric displacements between the samples as is required for example based deformation algorithms such as pose space deformation [3, 4, 7] that generate arbitrary poses of an articulated subject. The result (Fig. 2) is a rapidly deformable volumetric model of an articulated body containing accurate data-driven deformation of all anatomical layers. The model can be visualized by any volume rendering algorithm. Although the example based volume deformation (EVD) requires complex precomputations to handle raw medical volume samples, note that the EVD deformation time itself is small: the time is around 3.5 s to deform the human hand volume ($255 \times 255 \times 90$ voxel grid points) and 1.7 s to deform the knee volume ($255 \times 255 \times 123$ voxel grid points) to any arbitrary pose. The end result has potential uses in many applications of medical image analysis, biomechanics, and computer graphics.

4 Conclusion

This study presents a complete pipeline to produce a person specific volume deformation of articulated body regions, while managing the practical problems arising in multiple volume scans of a living human. The approach is demonstrated on MRI volumes of articulated body regions such as the human knee and hand. In particular, the human hand is one of the most complex articulated human body regions. It requires a novel and powerful registration approach to avoid the strong local minima inherent in registering highly articulated body region. Given the results obtained with this

complex subject, we feel there is a good argument that the method will work well for many simpler cases. Since we focused on several challenging problems, some related issues such as high quality volume visualisation and accurate joint modeling and animation were omitted or simplified and left for future work.

Acknowledgments We appreciate organizer and committee of FMfI 2015 and deliver special thanks to Prof. Hiroyuki Ochiai, Yoshihiro Mizoguchi at Kyushu University, Prof. Shizuo Kaji at Yamaguchi University, and Dr. Ken Anjyo at OLM digital for sincere discussion regarding mathematical issues of the topic.

References

1. Allen, B., Curless, B., Popovic, Z.: Articulated body deformation from range scan data. In: SIGGRAPH '02: Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques, pp. 612–619. ACM Press (2002)
2. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: Scape: shape completion and animation of people. ACM Trans Graph **24**(3), 408–416 (2005)
3. Kurihara, T., Miyata, N.: Modeling deformable human hands from medical images. In: Proceedings of the 2004 ACM SIGGRAPH Symposium on Computer Animation (SCA-04), pp. 357–366 (2004)
4. Lewis, J.P., Cordner, M., Fong, N.: Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In: SIGGRAPH '00: Proceedings of the 27th Annual Conference on Computer Graphics And Interactive Techniques, pp. 165–172. ACM Press/Addison-Wesley Publishing Co (2000)
5. Marsland, S., Twining, C.J.: Constructing diffeomorphic representations for the groupwise analysis of nonrigid registrations of medical images. IEEE Trans Med Imag **23**(8), 1006–1020 (2004)
6. Rhee, T., Lewis, J., Neumann, U., Nayak, K.: Soft-tissue deformation for in-vivo volume animaton. In: Proceedings of Pacific Graphics, pp 435–438 (2007)
7. Rhee, T., Lewis, J.P., Neumann, U., Nayak, K.: Scan-based volume animation driven by locally adaptive articulated registrations. IEEE Trans Vis Comput Graph **17**(3), 368–379 (2011)

The Mathematics Describing Two-Phase Geothermal Fluid Flows: Quantifying Industrial Applications and Innovations

Graham Weir

Abstract Geothermal energy generates about 10 % of New Zealand's electricity. At shallow depths, due to low pressure, geothermal fluid begins to boil, and forms a two-phase flow system. The corresponding equations are of mixed type, containing a parabolic equation for pressure, but hyperbolic equations for the liquid fraction and for dissolved chemicals. The steady flow equations are highly constrained, and are useful in the design of heat exchangers, and to chromatography. The transient flow equations are essential to the validation of reservoir models. However, the strong heterogeneity of the earth produces fractal-like behaviour in tracer transport, which raises many open questions. We present a dimensional argument, showing that a previously derived fractal Green's function can be derived by assuming a one-sided Gaussian distribution of permeability, and noting that an inversion of this distribution produces the corresponding tracer profile. Such tracer profiles are characterised by asymptotic inverse-square time behaviour, and consequently, all nonzero moments are unbounded.

Keywords Geothermal energy · Boiling · Tracer profiles · One-sided Gaussian distributions · Scale-dependent dispersivity

1 Introduction

Geothermal energy is an important cultural and energy source, especially for countries around the "Ring of Fire" [3]. Bathing in geothermal springs is popular, and believed by many to have therapeutic properties. Geysers, hot pools, boiling mud and volcanoes are attractive to many tourists. Most of the earth's gold, silver, and copper deposits have resulted from mineral transport by geothermal convection cells, with deposition occurring in the two-phase zone. Base-load electric power generation from geothermal fields is an important energy source in many countries [10].

G. Weir (✉)
Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand
e-mail: grahamweir@xtra.co.nz

B. Anderssen et al. (eds.), *The Role and Importance of Mathematics in Innovation*, Mathematics for Industry 25,
DOI 10.1007/978-981-10-0962-4_14

This paper introduces the equations used in numerical modelling of geothermal reservoirs. Large length scales are often appropriate to geothermal models, and consequently diffusive, conductive, and capillary effects can be of minor importance. We make such an assumption, and derive simplified idealised flow regimes associated with steady vertical flows, as well as, considering shock transport processes, and indicate some of the associated innovations.

Heterogeneity characterises geological media. Tracer tests are an important method for identifying preferred flow paths in a geothermal field. Tracer profiles are often analysed by assuming a fracture-block system, or by using a scale-dependent dispersivity. It has been found recently that many tracer profiles are well approximated by a probability function for which all nonzero moments are infinite. We show how such probability functions for tracer profiles result from an inversion performed on a one-sided Gaussian probability distribution of permeability.

2 The Mathematical Equations

We assume that the transport of mass, energy, and chemicals in a porous medium, in which the pore space is occupied by a single phase fluid (either a liquid or a gas), follows from Darcy's Law

$$\mathbf{V} = -\frac{k}{\mu}(\nabla P - \rho \mathbf{g}) \tag{1}$$

where $\mathbf{V}$, k, μ, P, ρ, $\mathbf{g}$ denote volumetric flux, permeability, dynamic viscosity, fluid pressure, density, and gravitational acceleration, respectively. Volumetric flux has the dimensions of volume of fluid per area of rock per unit time, and so V has the dimensions of a velocity (the Darcy velocity). Since fluid flow occurs within the pores of the rock matrix, typical fluid velocities are significantly greater than the Darcy velocity.

Darcy's law was derived observationally from geophysical measurements. It also follows from assuming low-velocity, non-turbulent flows subject to the Navier–Stokes equations. About feed points for wells, where high-velocity fluid flow occurs, Darcy's equation needs modification, for example, to the Brinkman equation [7].

In a two-phase (boiling) region, where both liquid and vapour phases exist within the pore space, Darcy's law in (1) is extended by assuming that each phase obeys separate Darcy equations, with fluid flow fluxes being the sum of the corresponding liquid and vapour phase fluxes, $\mathbf{V}_l$ and $\mathbf{V}_v$, respectively,

$$\mathbf{V}_l = -\frac{kk_l}{\mu_l}(\nabla P_l - \rho_l \mathbf{g}); \quad \mathbf{V}_v = -\frac{kk_v}{\mu_v}(\nabla P_v - \rho_v \mathbf{g}) \tag{2}$$

where k_l, k_v, μ_l, μ_v, P_l, P_v, ρ_l, ρ_v are the relative permeabilities, dynamic viscosities, pressures, and densities for the liquid and vapour phases, respectively. Different relative permeabilities arise because each phase obstructs the motion of the other phase. Non-flowing residual phases are possible, if this obstruction to flow from the

other phase becomes too great. Similarly, surface tension effects produce different pressures in the liquid and vapour phases, and an empirical relationship needs to be introduced to relate the two-phase pressures.

The conservation equations for mass, energy and chemicals are of the form

$$\partial_t \rho_i + \nabla \cdot \mathbf{J}_i = 0; \quad i = M, E, c \tag{3}$$

where ρ_i, $\mathbf{J}_i$ are the total density and flux of conserved quantity i. For example,

$$\rho_M = \phi\rho_l S + \phi\rho_v(1 - S); \quad \mathbf{J}_M = \rho_l \mathbf{V}_l + \rho_v \mathbf{V}_v \tag{4}$$

where ϕ, S are porosity and liquid saturation. The energy density and fluxes are

$$\rho_E = (1 - \phi)\rho_R U_R + \phi\rho_l U_l S + \phi\rho_v U_v(1 - S); \quad \mathbf{J}_E = \rho_l h_l \mathbf{V}_l + \rho_v h_v \mathbf{V}_v - K\nabla T \tag{5}$$

where $\rho_R, U_R, U_l, U_v, h_l, h_v, K, T$ are rock density, internal rock energy, internal liquid energy, internal vapour density, liquid enthalpy, vapour enthalpy, thermal conductivity and rock temperature, respectively.

The chemical density and chemical fluxes (for one conserved chemical) are

$$\rho_c = \phi\rho_l X_l S + \phi\rho_v X_v(1 - S); \quad \mathbf{J}_c = \rho_l X_l \mathbf{V}_l + \rho_v X_v \mathbf{V}_v - D_{cl}\nabla X_l - \rho_v \tau\phi S D_{cv}\nabla X_v \tag{6}$$

where X_l, X_v, D_{cl}, τ, D_{cv} are chemical liquid mass density, chemical vapour mass density, chemical liquid diffusivity, tortuosity, chemical vapour diffusivity, respectively [5, 13]. For multiple chemicals in the fluid, there are correspondingly multiple chemical conservation equations, as in (3) and (6) [14].

In a two-phase region, the mass fraction of the chemical in the vapour phase will be related to that in the liquid phase, usually by a Henry law,

$$X_v = X_v(X_l) \tag{7}$$

which allows X_v to be eliminated from the equations.

Finally, the difference between liquid and vapour pressures is typically defined as an empirical function of liquid saturation S,

$$P_l - P_v = P_{lv}(S) \tag{8}$$

to model capillary effects.

2.1 Steady Vertical Flow Equations

The aim of this subsection is to assume steady, vertical flows and constant permeabilities, ignoring diffusive, conductive, and capillary effects, in order to gain insight into the equations above. Then the three conservation equations (4)–(6) allow three

constant vertical fluxes, J_M, J_E, J_c, which depend essentially only on the local derivative of pressure. Because three free variables are essentially determined by only one variable ($\partial_z P$), there are two implicit constraints.

One of these constraints typically [1] implies that the liquid saturation S can take on one of two values. The low (high) value relates to vapour- (liquid-) saturated conditions, such as hold at the Larderello geothermal field in Italy (the Taupo Volcanic Zone in New Zealand). Utilisation of these two saturation values has led to the innovation of geothermal heat pipes [4].

The second constraint determines how the boiling point of the fluid depends on both fluid properties and the local flow [9].

2.2 *Infinitesimal Discontinuities*

Many numerical experiments with geothermal simulators show that rapid changes in pressure, for example from opening a well in a field, can result in shock-like propagation of saturation and chemical concentrations through the geothermal field. Diffusive, conductive, and capillary effects will ensure that pure shocks do not develop, but nevertheless, very rapid spatial variations in some variables can occur.

In the idealised case when diffusive, conductive, and capillary effects are ignored, pure shocks can occur. The corresponding Rankine–Hugoniot equations relate changes across the shock surface, in flux and density through $[J_i] = c[\rho_i]$, where c is the shock speed. In the case of only one chemical, shocks can occur only in saturation and chemical mass fractions, $[S]$, $[X_l]$, since (7) can be used to eliminate $[X_v]$. We obtain

$$c = \frac{\rho_l[V_l] + \rho_v[V_v]}{\phi(\rho_l - \rho_v)[S]} = \frac{\rho_l h_l[V_l] + \rho_v h_v[V_v]}{\phi(\rho_l h_l - \rho_v h_v)[S]} = \frac{\rho_l[X_l V_l] + \rho_v[X_v V_v]}{\phi(\rho_l[X_l S] + \rho_v[X_v(1-S)])} \quad (9)$$

where we have used thermodynamic identity $\rho_l U_l - \rho_v U_v = \rho_l h_l - \rho_v h_v$. Rearranging the first two expressions in (9) gives $[V_l + V_v] = 0$, or that volumetric flux is continuous. This follows since the pore space is always completely filled by fluid, and flow of a volume of fluid through a surface through the porous medium must be matched by a corresponding flow of an equal volume away from that surface.

The last two equations in (9) describe how shocks in chemical mass fraction transmit, but this depends on the expression in (7). An important innovation following from this difference in wave speed of different chemicals in a porous media is chromatography, but in most industrial applications, capillary effects should be considered, because the corresponding length scales are much smaller than for those in a geothermal application. A theory of multiple reacting chemicals in a porous medium is given in [14]. A general theory of infinitesimal shocks in a porous medium is given in [11].

2.3 Classification of Equations

The presence of approximate shock solutions to (2)–(6) arise because these equations are almost singular. The large length scales in geothermal reservoirs, and the frequent dominance of convection over diffusion, means that diffusive, conductive, and capillary effects are often small. Then the Laplacian of pressure can be eliminated from the energy and chemical conservation equations above, yielding one elliptic equation for pressure, and wave equations for saturation and chemical mass fraction, analogous to that occurring in Buckley–Leverett theory. A general separation of diffusive and wave equations in a porous medium is given in [11].

3 Analysing Tracer Returns

Geological media are characterised by immense variations in permeability. Consequently, tracer tests in geological media can be characterised by fast returns along preferential flow paths, combined with slow flows from low permeability paths. This has led to the concept of scale-dependent dispersion, in which the variance of tracer returns increases linearly with the scale of the experiment [6].

The aim of this section, and the main result of this paper, is to derive ideal tracer profiles, obtained from a specific distribution of permeability, which yields results consistent with a linear scale-dependent dispersion, as observed in field experiments. This will yield permeabilities to be considered in construction of a numerical model of a geothermal field.

The properties of the probability distribution of permeability we seek are

1. A peak in permeability, consistent with a set of preferred geological flow paths.
2. A rapid decrease in probability, for permeabilities greater than the peak value.
3. A nonzero value of probability for low values of permeability.

The last of these conditions reflects the great many dead-ends in a porous media, which will contribute to the low permeability paths. Note that it is not unusual to work with the logarithm of permeability, in order to capture the large range of naturally occurring permeabilities. However, this will typically impose a zero probability for zero permeability paths, unlike the requirement in (3).

The one-sided Gaussian distribution $p(k)$,

$$p(k) = \frac{\sqrt{2}}{\sqrt{\pi}s\left[1 + \operatorname{erf}\left(\frac{k_0}{\sqrt{2}s}\right)\right]} \exp\left[-\frac{(k-k_0)^2}{2s^2}\right]; \quad \int_0^\infty p dk = 1, \tag{10}$$

where k_0 is the peak permeability, and s is a variance, satisfies the three points above.

Consider a well discharging liquid water at a constant mass rate q (kg s^{-1}), with water density ρ, and tracer mass density in the water X (kg m^{-3}). Then the rate of

mass discharge of tracer $\dot{M}$ is qX/ρ (kg s^{-1}). In an infinitesimal amount of time, dt, the increment dM in tracer mass discharged is

$$dM = \dot{M}dt = \frac{qXdt}{\rho} = Mp(k)dk; \quad \int_0^\infty \dot{M}dt = M, \tag{11}$$

where M is the total mass of tracer recovered, and we associate an infinitesimal increment dk of permeability producing the discharged tracer, with probability $p(k)$ given in (10).

If the mean speed of fluid flowing in a pore is u, then

$$u = \frac{L}{t} = \frac{k\triangle P}{\phi\mu L}; \quad \rightarrow k = \frac{\mu\phi L^2}{t\triangle P}; \quad u_0 = \frac{k_0\triangle P}{\phi\mu L}, \tag{12}$$

where L is the distance between the injection and production well, $\triangle P$ is the pressure difference between the injection and production wells, t is the time since the tracer was injected into the injection well, and u_0 is a mean speed of tracer between injection and production wells.

If (12) is used to relate permeability k to time t, then from (11) and (10)

$$X = \frac{\rho\mu\phi L^2 M}{q\triangle Pt^2} p\left(\frac{\mu\phi L^2}{t\triangle P}\right) = \frac{\sqrt{2}\rho\mu\phi L^2 M \exp\left[-\frac{(L-u_0t)^2}{2\sigma t^2}\right]}{q\triangle Pt^2\sqrt{\pi}s\left[1+\text{erf}\left(\frac{1}{\sqrt{\theta}}\right)\right]}, \tag{13}$$

where

$$\sigma = \left(\frac{s\triangle P}{\mu\phi L}\right)^2; \quad \theta = \frac{2s^2}{k_0^2} \tag{14}$$

These expressions are identical to those derived earlier [12], by assuming a dispersion varying linearly with time, $D = \sigma t$, but where σ is now given explicitly in (14), rather than empirically. It was shown in [12] that many tracer profiles are well approximated by (13). It is clear from (13), that for large time, X decreases as t^{-2}, and so all nonzero moments of X with respect to time, are unbounded, since asymptotically $\int t^n t^{-2} dt$ is unbounded at the upper limit, for each integer n greater than zero.

If t_m is the time that the peak occurs in measured tracer mass density X, then

$$t_m = \frac{L(\sqrt{1+4\theta}-1)}{2u_0\theta}; \quad u_0 = \frac{L(\sqrt{1+4\theta}-1)}{2\theta t_m} \tag{15}$$

We can now find explicit expressions for k_0 and s,

$$k_0 = \frac{\phi\mu L^2(\sqrt{1+4\theta}-1)}{2\theta t_m\triangle P}; \quad s = \sqrt{\frac{\theta}{2}}k_0 \tag{16}$$

which fixes the the one-sided Gaussian probability function in (10).

When tracer returns are well approximated by (13), then the time t_m of peak returns can be read off the experimental record, and the parameter θ obtained by fitting (13) to the tracer return. Field measurements will also provide field temperatures and hence provide μ, as well as estimates of the pressure difference $\triangle P$ between the injection and production wells, and the distance L. An estimate of the porosity remains undetermined, with a possible default value being perhaps 0.1, which is a typical low value for a consolidated sandstone [2]. Then (16) gives k_0 and s.

Writing the probability function in (13) in nondimensional units

$$k(x) = \frac{2}{\sqrt{\pi}\left[1 + \text{erf}\left(\frac{1}{\sqrt{\theta}}\right)\right]} \exp\left[-\left(x - \frac{1}{\sqrt{\theta}}\right)^2\right]; \quad x = \frac{k}{k_0\sqrt{\theta}} \tag{17}$$

shows that for small θ, the permeability distribution is essentially a delta function centred on k_0, whereas for large θ, the permeability is essentially a one-sided Gaussian, with zero mean. Figure 1 plots four one-sided Gaussians. For large values of permeability, the probability is essentially zero, corresponding to zero tracer returns for early times. The nonzero value of probability for zero k produces the t^{-2} decay of tracer profiles, for long times.

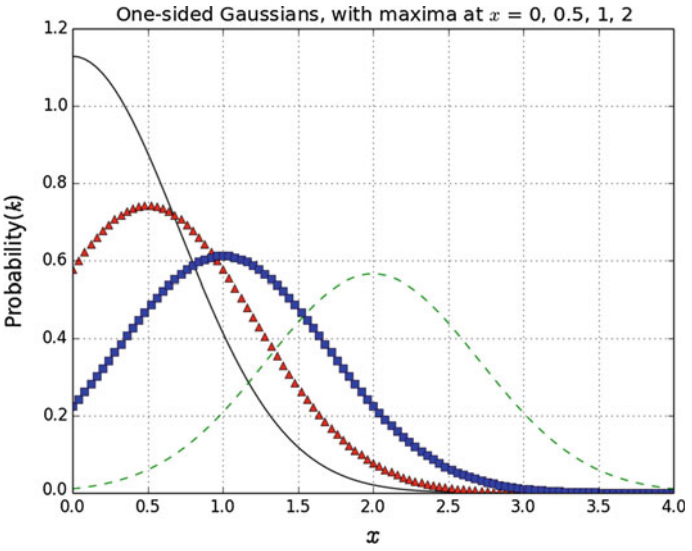


Fig. 1 One-sided Gaussian distributions, $k(x)$ versus nondimensional x $(=k/(k_0\sqrt{\theta}))$ with maxima $(k = k_0)$ at $x = 0$ (*black line*), 0.5 (*red triangles*), 1 (*blue squares*), 2 (*green dashes*)

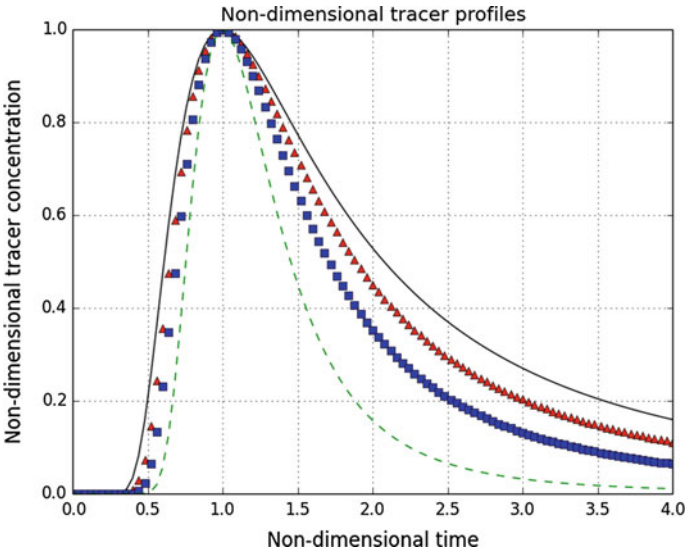


Fig. 2 Nondimensional tracer profiles, $Y(\tau)$ versus τ, for $\theta = \infty$ (*black line*), 4 (*red triangles*), 1 (*blue squares*), 0.25 (*green dashes*), corresponding to Fig. 1

The tracer profile X in (13) can be nondimensionalised to Y, through

$$X_0 = \frac{\sqrt{2}\rho\mu\phi L^2 M}{q\triangle P t_m^2 \sqrt{\pi} s\left[1 + \text{erf}\left(\frac{1}{\sqrt{\theta}}\right)\right]} \exp - \left(\frac{\sqrt{1+4\theta}-1}{2\sqrt{\theta}}\right) \tag{18}$$

$$Y = \frac{\exp\left(\frac{\sqrt{1+4\theta}-1}{2\sqrt{\theta}}\right)}{\tau^2} \exp - \left[\frac{\left(\sqrt{1+4\theta}+1\right)}{2\sqrt{\theta}\tau} - \frac{1}{\sqrt{\theta}}\right]^2 \tag{19}$$

$$X = X_0 Y; \tau = \frac{t}{t_m}; \quad 2\sigma = u_0^2; \quad Y(\tau = 1) = 1 \tag{20}$$

Figure 2 plots the dependence of $Y(\tau)$ vs τ, for $\theta = \infty$, 4, 1, 0.25. All four plots have been scaled for their maxima to occur at $\tau = 1$, and their maximum value there to be unity. In the limiting case of $\theta = \infty$, we have

$$Y = \frac{e}{\tau^2} \exp\left(-\frac{1}{\tau^2}\right); \quad k = \frac{2}{\sqrt{\pi}} \exp(-x^2) \quad \text{when} \quad \theta = \infty \tag{21}$$

4 Constructing the Numerical Model

A widely used numerical simulator for geothermal modelling is TOUGH2 [8], an acronym for Transport Of Unsaturated Groundwater and Heat. Often the size of numerical blocks used in TOUGH2 simulations are as large as 100m, whereas the interior of pores in a geological setting can be as small as 10^{-7} m. Consequently, there is considerable craft in choosing spatial permeability distributions in geothermal modelling, and specific models will often depend on the viewpoint of the modeler.

5 Conclusions

This paper has briefly reviewed the two-phase flow equations used in geothermal modelling, and outlined some of the flow regimes which can follow from these equations. Some of the innovations associated with these idealised flow regimes include the technologies of chromatography, geothermal heat pipes and geothermal energy utilisation.

The main new result of this paper was given in (13), which showed that a previously derived set of empirical tracer profiles, incorporating the concept of a scale-dependent dispersivity, follow identically from an underlying permeability structure satisfying a one-sided Gaussian permeability distribution. We outlined how a tracer profile can be used to derive the mean and variance of this Gaussian distribution. The corresponding tracer profile results from an inversion of this distribution, resulting in a new probability distribution in which all nonzero moments are unbounded.

Acknowledgments The author gratefully acknowledges funding from the Institute of Mathematics for Industry, Kyushu University.

References

1. Bai, W., Xu, W., Lowell, R.P.: The dynamics of submarine geothermal heat pipes. Geophys. Res. Lett. **30**(3), 1108–1111 (2003)
2. Corey, A.T.: Mechanics of heterogeneous fluids in porous media. Water Resources Publications, Fort Collins, Colorado **80522**, 4 (1977)
3. https://en.wikipedia.org/wiki/Ring_of_Fire
4. Kusaba, S., Suzuki, H., Hirowatari, K., Mochizuki, M., Mashiko, K., Nguyen, T., Akbarzadeh, A.: Extraction of geothermal energy and electric power generation using a large scale heat pipe. In: Proceedings of World Geothermal Congress, 2000, Kyushu-Tohuku, Japan, May 28–June 10, 2000, 3489–3494 (2000)
5. McKibbin, R., Pruess, K.: On non-condensible gas concentrations and relative permeabilities in geothermal reservoirs with gas-liquid co- or counterflows. In: 10th New Zealand Geothermal Workshop, Auckland, New Zealand (1988)
6. Neuman, S.P.: Universal scaling of hydraulic conductivities and dispersivities in geologic media. Water Resour. Res. **26**, 1949–1958 (1990)

7. Nield, D.A., Bejan, A.: Convection in Porous Media, 2nd edn. Springer, New York (1999)
8. Pruess, K.: SHAFT, MULKOM, TOUGH: A set of numerical simulators for multiphase fluid and heat flow, Geotermia, Revistica Mexicana de Geoenergia, pp. 185–202 (1988)
9. Sutton, F.M., McNabb, A.: Boiling curves at Broadlands geothermal field, New Zealand. N. Z. J. Sci. **20**, 333–337 (1977)
10. Weir, G.J. (ed.): Special Issue on geothermal energy. Transport in Porous Media **33**(1–2) (1998)
11. Weir, G.J.: Nonreacting chemical transport in a two-phase reservoir - factoring diffusive and wave properties. Transp. Porous Media **17**, 201–220 (1994)
12. Weir, G., Burnell, J.: Analysing tracer returns from geothermal reservoirs. In: Proceedings of 36th NZ Geothermal Workshop, Auckland, New Zealand, pp. 24–26 (2014)
13. Weir, G.J., Kissling, W.: Enhanced conduction in steady, vertical flows of water, steam and carbon dioxide in a porous medium. Transp. Porous Media **24**, 297–313 (1996)
14. White, S.P.: Multiphase nonisothermal transport of systems of reacting chemicals. Water Resour. Res. **31**(7), 1761–1772 (1995)

Leveraging Progress in Analytical Groundwater Infiltration for New Solutions in Industrial Metal Solidification

Dimetre Triadis

Abstract Previous analytical solutions for Stefan solidification problems with nonlinear heat conduction relied on a boundary flux proportional to $1/\sqrt{t}$. These can now be generalised to analytical series solutions for a large family of boundary fluxes with the same leading-order form, representing significant progress towards analytical treatment of simple casting systems for industrial metal manufacture. Mathematically, these nonlinear Stefan problems are intimately related to integrable solutions of Richards' equation, governing unsaturated one-phase groundwater flow though soil. For the most general known integrable soil model, it has taken 20 years to move from analytical treatment of a free-surface boundary condition implying constant rainfall to one implying surface saturation. Here the corresponding advances both in theory and efficient algorithms for symbolic computation are utilised directly to generalise the class of boundary fluxes for nonlinear Stefan solidification problems.

Keywords Stefan problems · Casting · Metal solidification · Heat conduction · Integrable PDEs · Symbolic computation · Phase-change

1 Introduction

Specialised continuous and billet casting processes utilised in industrial metal manufacture are often the subject of large-scale numerical simulations that provide predictions relevant to the particular process being studied. Analytical solution methods can only be applied to simplified casting geometries, however, they yield important physical insights into the general behaviour of casting systems, and are needed as benchmarks to evaluate the accuracy of more versatile numerical techniques. In the study of groundwater flow, relatively complex analytical series solutions have

D. Triadis (✉)
Institute of Mathematics for Industry, Kyushu University–Australia Branch,
La Trobe University, Melbourne, Australia
e-mail: D.Triadis@latrobe.edu.au

B. Anderssen et al. (eds.), *The Role and Importance of Mathematics in Innovation*, Mathematics for Industry 25,
DOI 10.1007/978-981-10-0962-4_15

yielded concrete insights into the very popular but often misunderstood Green–Ampt [4] infiltration model [13].

Understanding the simplified one-dimensional system illustrated in Fig. 1 is important to a wide variety of industrial metal casting processes. Here a layer of molten metal is initially placed upon a solid, cooler metal at time $t = 0$, with a known time-dependent quantity of heat extracted from the lower surface. The illustration shows a solidification front and a single secondary phase-change front, both of which originate at the metal-metal interface $x = 0$, and migrate away as time passes.

To the author's knowledge, there is no analytical solution method for the illustrated system that addresses practical nonlinear thermal properties. Exact solutions do exist for base and production metal layers that are assumed to have infinite thickness [3, 14]. Of course these also apply to finite-thickness metal layers at times small enough for boundary effects to be neglected. For the above solutions the natural heat flux at the metal-metal interface is found to be proportional to $1/\sqrt{t}$. Hence solutions are also known for systems that neglect explicit consideration of the base metal layer altogether, and consider only a production metal of infinite extent with a known heat extraction rate at its lower surface [7, 8]. We expect that explicitly accounting for a finite base layer will alter the leading-order interface flux through the action of lower-order terms in a small-time expansion. Hence the present study is aimed at generalising known results by considering a production metal of infinite extent with

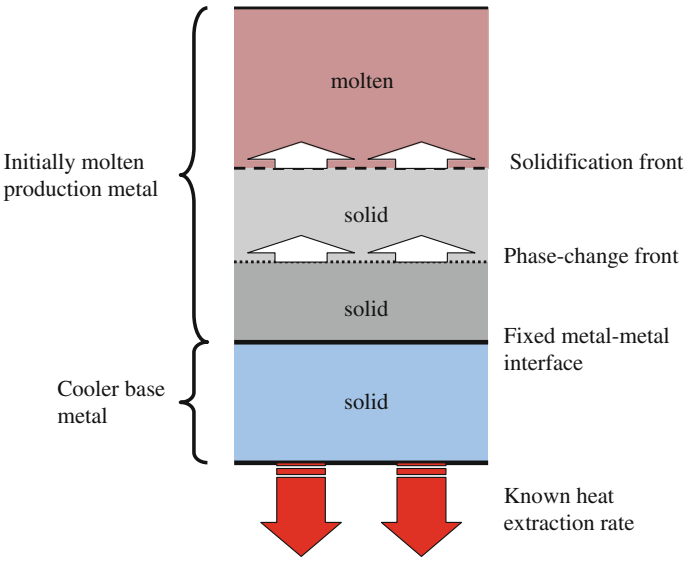


Fig. 1 Simplified casting system with solidification and phase-change fronts. Heat flux at the bottom surface of the cast wall metal (shown in *blue*), is known. Two production metal phase-change fronts are shown migrating away from the metal-metal interface $x = 0$, where they are assumed to originate at time $t = 0$

a known heat flux at the bottom surface that may be expressed as a power series in $\sqrt{t}$ with leading-order term proportional to $1/\sqrt{t}$.

The canonical $1/\sqrt{t}$ heat flux results in a constant metal-metal interface temperature. Our generalised results predict a time-dependent interface temperature, that has the potential to be compared to experimental studies that are capable of recording this cast interface temperature [6].

While there is often a larger number of phase changes relevant to a particular casting process, our model below will only consider two phases of the production metal; a molten and a solid phase. Thus the phase-change front of Fig. 1 and similar phase-change fronts are neglected to demonstrate the solution method as clearly as possible. As our modelling assumptions do not include fluid motion, there is no mathematical difference between the treatment of a solid-liquid phase-change front and a solid-solid phase-change front. From earlier studies [14] and the development below it should be clear that there is no significant difficulty involved in extending the model to a larger number of phases.

2 Linearising the Governing Equations

We assume that heat flows in both the solid phase ($i = 1$) and the molten phase ($i = 2$) are governed by the nonlinear diffusion equations

$$c_i(\theta_i)\frac{\partial \theta_i}{\partial t} = \frac{\partial}{\partial x}\left[k_i(\theta_i)\frac{\partial \theta_i}{\partial x}\right], \tag{1}$$

for the temperature $\theta_i(x, t)$ as a function of the distance from the interface boundary x and time t. The thermal properties of the production metal manifest through the volumetric heat capacity $c_i(\theta_i)$ and thermal conductivity $k_i(\theta_i)$ of each phase, all of which are assumed to be known functions of the relevant temperature range.

The phase change from molten to solid metal takes place at position $x = X(t)$ and fixed temperature θ_c

$$\theta_i\big(X(t), t\big) = \theta_c \qquad \text{for} \quad t > 0, \tag{2}$$

and we impose no compatibility criterion for the thermal properties on either side of this boundary. We account for a known latent heat of solidification λ at the phase-change front through a Stefan boundary condition for the flux:

$$k_1(\theta_1)\frac{\partial \theta_1}{\partial x} - k_2(\theta_2)\frac{\partial \theta_1}{\partial x} = \lambda\, \dot{X}(t) \quad \text{at} \quad x = X(t). \tag{3}$$

At $t = 0$, $X(0) = 0$; and the molten metal is assumed to be at a temperature $\theta_0 > \theta_c$, so that

$$\theta_2(x, 0) = \theta_0. \tag{4}$$

Finally, as discussed in the introduction, we assume that the outer-surface flux $U_0(t)$ is known, and can be expressed in power series form through the set of coefficients $\{\zeta_n\}$:

$$U_0(t) = k_1(\theta_1)\frac{\partial \theta_1}{\partial x} = \sum_{n=0}^{\infty} t^{\frac{n-1}{2}}\zeta_n \quad \text{at} \quad x = 0. \tag{5}$$

We first form dimensionless variables θ_*, t_*, x_*, c_{*i}, k_{*i}, etc…with some characteristic temperature θ_s, conductivity k_s heat capacity c_s and time scale t_s. Note that the leading-order problem in time with only $\zeta_0 \neq 0$ has a scaling symmetry related to the fact that there is no evident time scale t_s or length scale l_s in the system (1)–(5); only the ratio l_s^2/t_s has a natural scale, k_s/c_s.

Introducing new 'heat density' dependent variables $\Theta_i(x, t)$ with arbitrary constants Θ_{ci}

$$\Theta_i \equiv \Theta_{ci} + \int_{\theta_{*c}}^{\theta_*} c_{*i}(\bar{\theta})\, d\bar{\theta}, \tag{6}$$

standardises the form of our governing equations

$$\frac{\partial \Theta_i}{\partial t_*} = \frac{\partial}{\partial x_*}\left[\frac{k_{*i}(\theta_{*i})}{c_{*i}(\theta_{*i})}\frac{\partial \Theta_i}{\partial x_*}\right]. \tag{7}$$

These can be rendered integrable by assuming a particular functional form for the heat diffusivity in both phases

$$\frac{k_{*i}(\theta_{*i})}{c_{*i}(\theta_{*i})} = \frac{\alpha_i}{\Theta_i^2}, \tag{8}$$

where the constants α_i may be chosen with the Θ_{ci} to fit the properties of a particular metal. For some metals, functions of the above form provide a reasonable fit to their thermal properties [9]. The heat diffusivity of other metals may be approximated accurately by versatile segments of the above type, equivalent to introducing additional fictitious phase-change boundaries with zero latent heat [14]. Models with less sophisticated piecewise-constant approximation of material properties have also been presented [11].

We proceed to linearise the governing equations via the 'reciprocal Bäcklund transformation' formalism of [5]. That is, we adopt the new independent variables,

$$y_1 = \int_0^{x_*} \Theta_1(\bar{x}, t)\, d\bar{x} + I_0(t_*) \tag{9}$$
$$= -\int_{x_*}^{X_*(t_*)} \Theta_1(\bar{x}, t)\, d\bar{x} + \Theta_{c1} X_*(t_*) + I_{c1}(t_*)$$
$$y_2 = -\int_{X_*(t_*)}^{x_*} \Theta_2(\bar{x}, t)\, d\bar{x} + \Theta_{c2} X_*(t_*) + I_{c2}(t_*) \tag{10}$$

while retaining time t_*, and introduce dependent variables

$$u_i(y_i, t_*) = \frac{1}{\Theta_i}, \tag{11}$$

where $I_0(t_*)$ and the $I_{ci}(t_*)$ represent the total heat passing through various boundaries

$$I_0(t_*) = \int_0^{t_*} U_{*0}(\tau)\, d\tau, \qquad I_{ci}(t_*) = \int_0^{t_*} \frac{\alpha_i}{\Theta_i^2} \frac{\partial \Theta_i}{\partial x_*}\bigg|_{x_*=X_*(\tau)} d\tau. \tag{12}$$

The alternative forms of (9) can be derived from the reciprocal Bäcklund transformation by integrating over different regions, as in (3.13) and (3.15) of [8]. The above method of linearisation is essentially equivalent to successive transformations attributed to Kirchhoff, and Knight or Storm, as detailed in [14].

A final transformation follows from the scaling symmetry of the leading-order problem. We change independent variables from y_i and t_* to

$$\omega_i = \frac{y_i}{\sqrt{\alpha_i t_*}} \quad \text{and} \quad t_*. \tag{13}$$

Let $\Omega_0(t_*)$ denote the value of ω_1 corresponding to the interface boundary at $x_* = 0$. From the known boundary flux

$$U_{*0}(t_*) = \sum_{n=0}^{\infty} t_*^{\frac{n-1}{2}} \zeta_{*n}, \tag{14}$$

it follows that

$$\Omega_0(t_*) = \sum_{n=0}^{\infty} t_*^{\frac{n}{2}} \frac{2\zeta_{*n}}{\sqrt{\alpha_i}(n+1)} \equiv \sum_{n=0}^{\infty} t_*^{\frac{n}{2}} \gamma_{0,n}. \tag{15}$$

To track the moving boundary at $x_* = X_*(t_*)$ with our new variables, let $\omega_1 = \Omega_1(t_*)$, and $\omega_2 = \Omega_2(t_*)$ at the solidification front. Our adopted method of linearisation is somewhat complicated by the fact that $\Omega_1(t_*) \neq \Omega_2(t_*)$, an issue that was avoided by the linearisation path taken in [14]. However, we will see that this merely introduces an extra phase-front flux boundary condition to satisfy in our transformed problem.

From an understanding of the similar order-by-order solution procedure of [12], we expect to have to determine sets of constants $\{\gamma_{i,n}\}$ where

$$\Omega_i(t_*) = \sum_{n=0}^{\infty} t_*^{\frac{n}{2}} \gamma_{i,n}. \tag{16}$$

We now write the separable transformed equations

$$t_* \frac{\partial u_i}{\partial t_*} = \frac{\omega_i}{2} \frac{\partial u_i}{\partial \omega_i} + \frac{\partial^2 u_i}{\partial \omega_i^2}, \tag{17}$$

and transformed boundary conditions out in full. The surface flux condition takes the form

$$-\frac{\sqrt{\alpha_1}}{u_1} \left. \frac{\partial u_1}{\partial \omega_1} \right|_{\omega_1 = \Omega_0(t_*)} = \sqrt{t_*} U_{*0}(t_*). \tag{18}$$

The phase-front temperature conditions remain straightforward,

$$u_i\big(\Omega_i(t_*), t_*\big) = \frac{1}{\Theta_{ci}}, \tag{19}$$

but we now have two phase-front flux conditions to satisfy, which can be written as two independent equations in a variety of ways. The following forms are considered relatively simple to use

$$I_{c1} = -\sqrt{\alpha_1}\Theta_{c1} \int_0^{t_*} \frac{1}{\sqrt{\tau}} \left. \frac{\partial u_1}{\partial \omega_1} \right|_{\omega_1 = \Omega_1(\tau)} d\tau = \frac{\sqrt{\alpha_1}(\lambda_* - \Theta_{c2})\Omega_1(t_*) + \sqrt{\alpha_2}\,\Theta_{c1}\Omega_2(t_*)}{\lambda_* + \Theta_{c1} - \Theta_{c2}}, \tag{20}$$

$$I_{c2} = -\sqrt{\alpha_2}\Theta_{c2} \int_0^{t_*} \frac{1}{\sqrt{\tau}} \left. \frac{\partial u_2}{\partial \omega_2} \right|_{\omega_2 = \Omega_2(\tau)} d\tau = \frac{\sqrt{\alpha_2}(\lambda_* + \Theta_{c1})\Omega_2(t_*) - \sqrt{\alpha_1}\,\Theta_{c2}\Omega_1(t_*)}{\lambda_* + \Theta_{c1} - \Theta_{c2}}. \tag{21}$$

Finally we denote Θ_{20} as the value of Θ_2 at $t_* = 0$, and must have

$$u_2(\omega_2, t_*) \to \frac{1}{\Theta_{20}} \quad \text{as} \quad t_* \to 0. \tag{22}$$

While this system is larger than Eqs. (24) and (26) of [12], both systems are amenable to iterative series solution methods. The linearisation technique detailed above was tailored to reproduce Eq. (24) of [12] so that the iterative techniques developed in [1, 12] can be more directly utilised.

3 Iterative Solution Procedure

Solutions of (17) are confluent hypergeometric functions, and we have some choice as to which two independent solutions we choose. As in [12], the following choices have tidy power series forms convenient for iterative solution

$$u_i(\omega_i, t_*) = \sum_{m=0}^{\infty} t_*^{\frac{m}{2}} \left\{ C_{i,m} G^{-}\left(-\frac{m}{2}; \frac{1}{2}; -\frac{\omega_i^2}{4}\right) + D_{i,m} G^{+}\left(-\frac{m}{2}; \frac{1}{2}; -\frac{\omega_i^2}{4}\right)\right\}, \tag{23}$$

$$G^{\mp}\left(-\frac{m}{2}; \frac{1}{2}; -\frac{\omega_i^2}{4}\right) \equiv \sqrt{\pi} \sum_{p=0}^{\infty} \frac{(\mp\omega_i)^p}{p!\,\Gamma(1+\frac{m}{2}-\frac{p}{2})} \tag{24}$$

$$= \frac{\sqrt{\pi}}{\Gamma(1+\frac{m}{2})}\, {}_1F_1\left(-\frac{m}{2}; \frac{1}{2}; -\frac{\omega_i^2}{4}\right) \mp \frac{\sqrt{\pi}\,\omega_i}{\Gamma(\frac{1}{2}+\frac{m}{2})}\, {}_1F_1\left(\frac{1}{2}-\frac{m}{2}; \frac{3}{2}; -\frac{\omega_i^2}{4}\right).$$

Note that G^{-} is a just a Kummer confluent hypergeometric function of the second kind, and that we have introduced four sets of separation constants $\{C_{i,m}\}$ and $\{D_{i,m}\}$. Any series solutions above with arbitrary constants $\{C_{i,m}\}$ and $\{D_{i,m}\}$ correspond to exact solutions of the initial nonlinear heat equations (1), our task now is to find the members of this family that satisfy the boundary conditions (18)–(22) by determining $\{C_{i,m}\}$ and $\{D_{i,m}\}$.

Given the power series representations of $G^{\pm}$, and Eq. (16), it is clear that evaluation of our transformed boundary conditions involves manipulating terms of type

$$\left(\sum_{n=0}^{\infty} z^n \gamma_n\right)^p = \sum_{n=0}^{\infty} z^n W_n\big(p; \{\gamma_q : q \le n\}\big), \tag{25}$$

involving an unknown set of coefficients $\{\gamma_n\}$. The $W_n\big(p; \{\gamma_q : q \le n\}\big)$ coefficients naturally take the form of sums over partitions [1], but can be efficiently evaluated iteratively as in [12].

However, boundary condition evaluation to isolate terms of different orders in time does not involve the $W_n\big(p; \{\gamma_q : q \le n\}\big)$ coefficients directly. Consider for example the phase-front temperature boundary condition (19) above:

$$\frac{1}{\Theta_{ci}} = u_i\left(\Omega_i, t_*\right) \tag{26}$$

$$= \sum_{l=0}^{\infty} t_*^{l/2} \sum_{n=1}^{l} \left\{ C_{l-n} \xi_i^{-}(n, l-n) + D_{l-n} \xi_i^{+}(n, l-n)\right\},$$

$$\xi_i^{\mp}(n, j) \equiv \sqrt{\pi} \sum_{m=0}^{\infty} \frac{(\mp 1)^m W_n\big(m; \{\gamma_{i,q} : q < n\}\big)}{m!\,\Gamma(1+\frac{j}{2}-\frac{m}{2})}. \tag{27}$$

Note that while the index i in (26) may take the value 1 or 2, the index i in (27) may also take the value 0, using the known set of coefficients $\{\gamma_{0,n}\}$ introduced in (15). From (24) it follows that the $\xi_i^{\mp}(0, j)$ can be expressed as easily computable ${}_1F_1$ hypergeometric functions.

The key to efficiently calculating terms like those in the series form of (26), is an iterative result generalising equation (A6) of [12]

$$\xi_i^{\mp}(n, j) = \mp\gamma_{i,n}\xi_i^{\mp}(0, j-1) \mp \frac{1}{\gamma_{i,0}n}\xi_i^{\mp}(0, j-1)\sum_{q=1}^{n-1}\gamma_{i,q}\gamma_{i,n-q}(n-q)$$
$$\mp \frac{1}{\gamma_{i,0}n}\sum_{s=1}^{n-1}\xi_i^{\mp}(s, j-1)\sum_{q=s}^{n-1}\gamma_{i,q-s}\gamma_{i,n-q}(n-q) - \frac{1}{\gamma_{i,0}n}\sum_{s=1}^{n-1}\xi_i^{\mp}(s, j)\gamma_{i,n-s}s. \tag{28}$$

This is derived using two iterative identities for the $W_n\big(p, \{\gamma_q : q \le n\}\big)$ coefficients.

We note here a closely related problem [10] valid for a more limited range of materials governed by linear diffusion, there terms of different order were isolated by repeated differentiation, rather than the series rearrangement methodology demonstrated in (26).

It soon becomes clear that the initial condition (22) is exceptional, and can be simply satisfied by fixing the set $\{D_{2,m}\}$

$$D_{2,0} = \frac{1}{\sqrt{\pi}\Theta_{20}}, \qquad D_{2,m} = 0 \quad \text{for} \quad m \ge 1. \tag{29}$$

This leaves five sets of undetermined constants $\{C_{1,m}\}$, $\{D_{1,m}\}$, $\{C_{2,m}\}$, $\{\gamma_{1,m}\}$, $\{\gamma_{2,m}\}$ to be evaluated order-by-order by small-time series expansion of the five remaining boundary conditions (18)–(21).

We find that the leading-order equations are nonlinear

$$\zeta_0 C_{1,0}\sqrt{\pi}\operatorname{erfc}\left(\frac{\gamma_{0,0}}{2}\right) + \zeta_0 D_{1,0}\sqrt{\pi}\left(1 + \operatorname{erf}\left(\frac{\gamma_{0,0}}{2}\right)\right) \tag{30}$$
$$= (\sqrt{\alpha_1}C_{1,0} - \sqrt{\alpha_1}D_{1,0})\exp\left(-\frac{\gamma_{0,0}^2}{4}\right),$$

$$\sqrt{\alpha_2}\Theta_{c2}\exp\left(-\frac{\gamma_{2,0}^2}{4}\right)\left(2C_{2,0} - \frac{1}{\sqrt{\pi}\Theta_{20}}\right) = \frac{\sqrt{\alpha_2}(\lambda_* + \Theta_{c1})\gamma_{2,0} - \sqrt{\alpha_1}\,\Theta_{c2}\gamma_{1,0}}{\lambda_* + \Theta_{c1} - \Theta_{c2}}, \tag{31}$$

$$\sqrt{\alpha_1}\Theta_{c1}\exp\left(-\frac{\gamma_{1,0}^2}{4}\right)(2C_{1,0} - 2D_{1,0}) = \frac{\sqrt{\alpha_1}(\lambda_* - \Theta_{c2})\gamma_{1,0} + \sqrt{\alpha_2}\,\Theta_{c1}\gamma_{2,0}}{\lambda_* + \Theta_{c1} - \Theta_{c2}}, \tag{32}$$

$$\frac{1}{\Theta_{c1}} = C_{1,0}\mathrm{erfc}\left(\frac{\gamma_{1,0}}{2}\right) + D_{1,0}\left(1 + \mathrm{erf}\left(\frac{\gamma_{1,0}}{2}\right)\right), \tag{33}$$

$$\frac{1}{\Theta_{c2}} = C_{2,0}\mathrm{erfc}\left(\frac{\gamma_{2,0}}{2}\right) + \frac{1}{\sqrt{\pi}\Theta_{20}}\left(1 + \mathrm{erf}\left(\frac{\gamma_{2,0}}{2}\right)\right). \tag{34}$$

These are easily reduced to the form $f(\gamma_{2,0}) = 0$ and solved numerically, yielding $C_{1,0}$, $D_{1,0}$, $C_{2,0}$, $\gamma_{1,0}$ and $\gamma_{2,0}$.

As in [12], considering an l'th-order satisfaction of our boundary conditions, we encounter a linear system in the l'th-order constants

$$\begin{pmatrix} a_{11} & a_{12} & 0 & 0 & \\ 0 & 0 & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & 0 & a_{34} & a_{35} \\ a_{41} & a_{42} & 0 & a_{44} & 0 \\ 0 & 0 & a_{53} & 0 & a_{55} \end{pmatrix} \begin{pmatrix} C_{1,l} \\ D_{1,l} \\ C_{2,l} \\ \gamma_{1,l} \\ \gamma_{2,l} \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{pmatrix}. \tag{35}$$

The coefficients of this system are untidy expressions involving $\xi_i^{\mp}(q, j)$ and constants $C_{1,q}, D_{1,q}, C_{2,q}, \gamma_{1,q}$ and $\gamma_{2,q}$ of all orders $q < l$. They are written out explicitly in the Appendix.

Through repeated calculation of this linear system using (28), we can evaluate as many coefficients as necessary either to satisfy our boundary conditions to specified tolerances, or to observe divergent behaviour of our small-time series solution at larger times.

If evaluating up to the Nth-order constants satisfies our termination criterion, setting all higher-order constants equal to zero implies transformed solutions which take the form of sums of $N + 1$ hypergeometric terms from (23), and a phase-front position $X(t)$ specified as a power series with $N + 1$ terms via the $\Omega_i(t_*)$.

In heat-density coordinates solutions take the parametric form

$$x = \sqrt{\alpha_i t_*}\int_{\Omega_i(t_*)}^{\omega_i} u_i(\bar{\omega}_i, t_*)\, d\bar{\omega}_i + \frac{\sqrt{\alpha_i t_*}\Omega_i(t_*)}{\Theta_{ci}} + \sqrt{\alpha_i}\int_0^{t_*} \frac{1}{\sqrt{\tau}} \left.\frac{\partial u_i}{\partial \omega_i}\right|_{\omega_i = \Omega_i(\tau)} d\tau,$$

$$\Theta_i = \frac{1}{u_i(\omega_i, t_*)};$$

and our final step is to obtain θ_{*i} from the Θ_i by inverting (6).

4 An Illustrative Example

We can adopt some simple metal properties as a demonstration model. Assume a phase change at 1400 K, with phase-change latent heat 2000 J/cm^3, and a constant volumetric heat capacity $c(\theta)$ of 4 W s cm^{-3} K^{-1}. Let the nonlinear thermal conductivity $k(\theta)$ be specified according to

$$\alpha_1 = \frac{3+2\sqrt{2}}{50}, \qquad \Theta_{c1} = -\frac{\sqrt{2}+2}{5}, \tag{36}$$

$$\alpha_2 = \frac{3+2\sqrt{2}}{4}, \qquad \Theta_{c1} = \frac{\sqrt{2}+2}{5}; \tag{37}$$

as illustrated in Fig. 2. The metal properties thus chosen have a vague resemblance to those of copper.

We set the initial temperature of the molten metal at $\theta_0 = 1800\,\mathrm{K}$ and consider three simple interface fluxes as shown in Fig. 3. The black curve contains only the dominant term as $t_* \to 0$ such that the corresponding problem can be solved without the aid of series expansions. The blue and red curves have the same leading-order term, but have lower-order terms included that result in more or less heat respectively being extracted over the illustrated period.

For subsequent calculations, we chose the largest terms in our series solutions so that the phase-front temperature error $|\theta_1(X(t),t) - \theta_2(X(t),t)|$ was found to be consistently less than 1 mK. For the examples considered this criterion was sufficient to ensure that the other boundary conditions were satisfied to a high degree of accuracy. Unstable oscillations about the ideal value were observed when tracking satisfaction of boundary conditions for times significantly greater than $t_* = 0.4$. This is interpreted as divergence of our small-time series solutions at relatively large times.

The solution for the leading-order surface flux term predicts a constant interface temperature of about 923 K. Adding terms to promote heat extraction results in the surface temperature dropping for intermediate times, as shown in Figs. 4 and 6. Conversely decreasing the rate of boundary heat flux results in the surface temperature

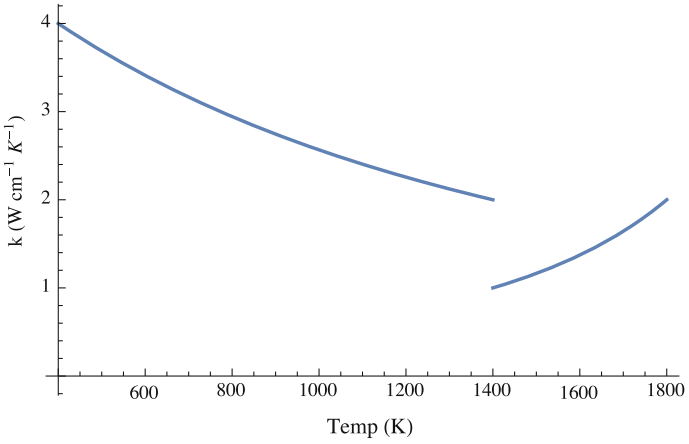


Fig. 2 Example metal conductivity $k(\theta)$ with discontinuity at $\theta_c = 1400\,\mathrm{K}$

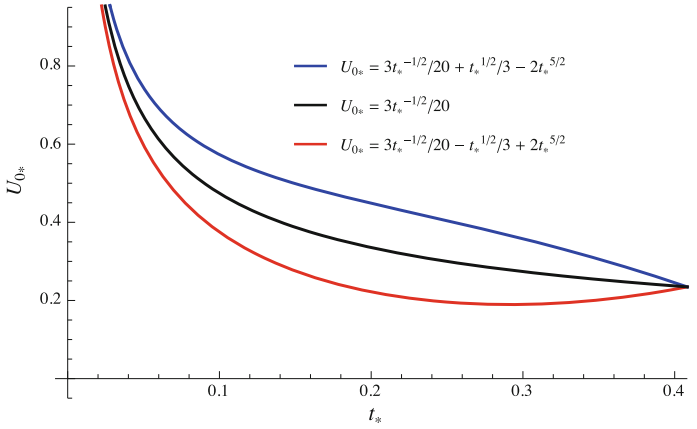


Fig. 3 Three scaled surface fluxes of interest. The *black central curve* shows the canonical surface flux. The flux specified by the *blue curve* extracts more heat, whereas the flux in *red* does not extract as much heat

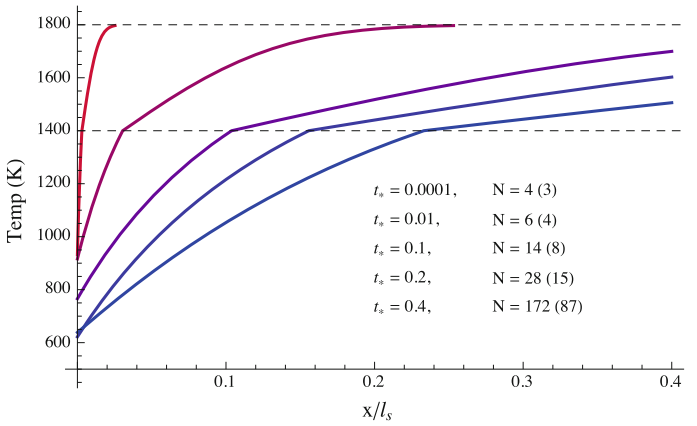


Fig. 4 Full temperature profiles calculated from the boundary flux $U_{*0} = 3\sqrt{t_*}/20 + \sqrt{t_*}/3 - 2t_*^{5/2}$. Here the steep *red curve* is the $t_* = 0.0001$ solution, and the final *blue curve* the $t_* = 0.4$ solution. The N values listed show the suffix of the largest coefficients used to generate each curve, and the number of non-zero terms in the same series expansion, in parentheses

increasing at intermediate times, as shown in Figs. 5 and 6. The location of the moving solidification front is shown in Fig. 7. As expected, extracting more heat at the boundary results in a phase-front position that is further into the body of the production metal at a particular time.

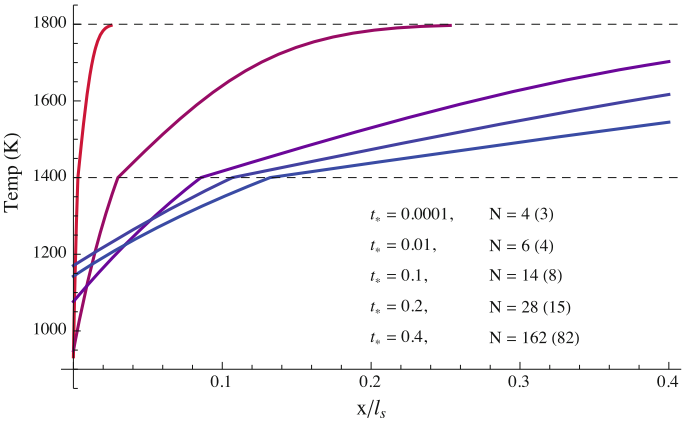


Fig. 5 Full temperature profiles calculated from the boundary flux $U_{*0} = 3\sqrt{t_*}/20 - \sqrt{t_*}/3 + 2t_*^{5/2}$. Here the steep *red curve* is the $t_* = 0.0001$ solution, and the final *blue curve* the $t_* = 0.4$ solution. The N values listed show the suffix of the largest coefficients used to generate each curve, and the number of non-zero terms in the same series expansion, in parentheses

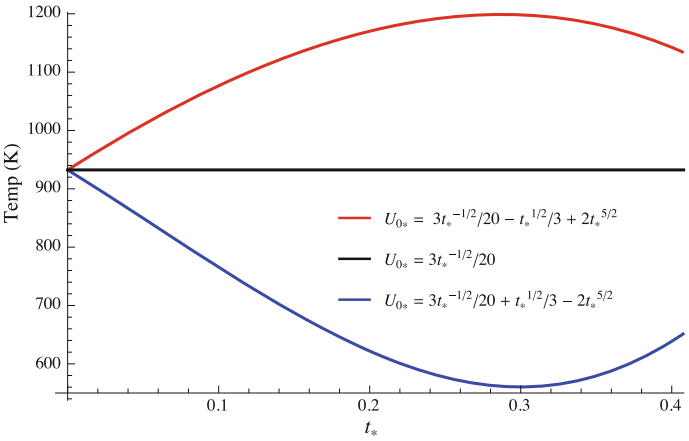


Fig. 6 Surface temperature variation for the three boundary fluxes of interest

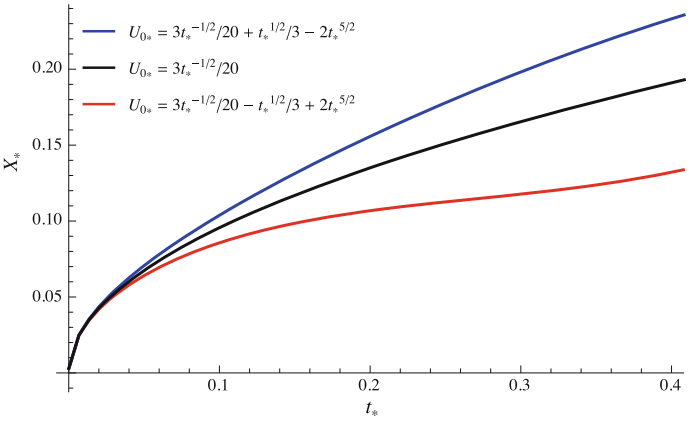


Fig. 7 Solidification front positions for the three boundary fluxes of interest

5 Conclusion

We have now demonstrated that analytical series solutions for nonlinear heat conduction may be produced for a vastly expanded family (5) of boundary fluxes with leading $O(1/\sqrt{t})$ term. Due to efficient iterative algorithms, the full range of validity of solutions can be explored with minimal computational restrictions.

As small-time expansions, these solutions diverge for sufficiently large times. Their convergence at any non-zero time has not been strictly proven. The solution of [12] was found to satisfy boundary conditions accurately at surprisingly large times, with latter soil moisture profiles reasonably approximating the large-time travelling wave solution. In the present context we have shown that series solutions can produce accurate temperature profiles that exhibit a wide range of boundary-surface temperatures. Larger systems that closely match industrially relevant casting processes will need to be solved to ascertain just how much of an impediment large-time divergence is, though future work will also explore the possibility of analytic continuation of the solution series. A closely related issue involves the possibility of crossing of critical phase-boundary temperatures at non-zero times, which may be conversely be viewed as being able to account for non-constant initial conditions. Having the metal-metal interface temperature specified rather than the boundary flux will at times be more useful, and solutions for this alternate boundary condition should be easily produced. Work towards explicit consideration of a finite cast metal layer as illustrated in Fig. 1 remains ongoing.

Many contexts exist where producing series solutions from a leading-order symmetry is possible, and of interest. Related soil water infiltration solutions appear to produce more manageable systems of equations, and many boundary conditions of practical importance remain unsolved. Curvature-dependent surface diffusion in the vicinity of a grain boundary [2] is another intriguing application.

Appendix

Here we show the explicit form of the linear system (35), which results from reordering each boundary condition to isolate terms of different orders in time. From the surface flux condition (18), we have

$$a_{11} = \zeta_0\, \xi_0^-(0,l) - \sqrt{\alpha_1}\, \xi_0^-(0,l-1), \tag{38}$$

$$a_{12} = \zeta_0\, \xi_0^+(0,l) + \sqrt{\alpha_1}\, \xi_0^+(0,l-1), \tag{39}$$

$$\begin{aligned} b_1 = &- \sum_{p=0}^{l-1} \zeta_{l-p}\big[C_{1,p}\, \xi_0^-(0,p) + D_{1,p}\, \xi_0^+(0,p)\big] \\ &- \sum_{n=1}^{l} \sum_{p=n}^{l} \zeta_{l-p}\big[C_{1,p-n}\, \xi_0^-(n,p-n) + D_{1,p-n}\, \xi_0^+(n,p-n)\big] \\ &+ \sqrt{\alpha_1} \sum_{n=1}^{l} C_{1,l-n}\, \xi_0^-(n,l-n-1) + D_{1,l-n}\, \xi_0^+(n,l-n-1). \end{aligned} \tag{40}$$

For notational convenience we define

$$\xi_i^{r\mp}(n,j) \equiv \xi_i^{\pm}(n,j) \pm \gamma_{i,n}\, \xi_i^{\mp}(0,j-1), \tag{41}$$

the remainder when the term with the highest-order coefficient $\gamma_{i,n}$ is removed from $\xi_i^{\mp}(n,j)$. From the phase-front flux boundary conditions (20) and (21) we have

$$a_{23} = \frac{2\sqrt{\alpha_2}\Theta_{c2}}{l+1} \xi_2^-(0,l-1), \tag{42}$$

$$a_{24} = \frac{\sqrt{\alpha_1}\Theta_{c2}}{\lambda + \Theta_{c1} - \Theta_{c2}}, \tag{43}$$

$$a_{25} = -\frac{\sqrt{\alpha_2}(\lambda + \Theta_{c1})}{\lambda + \Theta_{c1} - \Theta_{c2}} - \frac{2\sqrt{\alpha_2}\Theta_{c2}}{l+1}\big[C_{2,0}\, \xi_2^-(0,-2) + D_{2,0}\, \xi_2^+(0,-2)\big], \tag{44}$$

$$b_2 = \frac{2\sqrt{\alpha_2}\Theta_{c2}}{l+1}\left\{- C_{2,0}\, \xi_2^{r-}(l,-1) + D_{2,0}\, \xi_2^{r+}(l,-1) - \sum_{n=1}^{l-1} C_{2,l-n}\, \xi_2^-(n,l-n-1)\right\}; \tag{45}$$

$$a_{31} = \frac{2\sqrt{\alpha_1}\Theta_{c1}}{l+1} \xi_1^-(0,l-1), \tag{46}$$

$$a_{32} = -\frac{2\sqrt{\alpha_1}\Theta_{c1}}{l+1} \xi_1^+(0,l-1), \tag{47}$$

$$a_{34} = -\frac{\sqrt{\alpha_1}(\lambda - \Theta_{c2})}{\lambda + \Theta_{c1} - \Theta_{c2}} - \frac{2\sqrt{\alpha_1}\Theta_{c1}}{l+1}\big[C_{1,0}\, \xi_1^-(0,-2) + D_{1,0}\, \xi_1^+(0,-2)\big], \tag{48}$$

$$a_{24} = -\frac{\sqrt{\alpha_2}\Theta_{c1}}{\lambda + \Theta_{c1} - \Theta_{c2}}, \tag{49}$$

$$b_3 = \frac{2\sqrt{\alpha_1}\Theta_{c1}}{l+1}\Bigg\{-C_{1,0}\,\xi_1^{r-}(l,-1) + D_{1,0}\,\xi_1^{r+}(l,-1) + \sum_{n=1}^{l-1}\Big[-C_{1,l-n}\,\xi_1^{-}(n,l-n-1) + D_{1,l-n}\,\xi_1^{+}(n,l-n-1)\Big]\Bigg\}. \tag{50}$$

Finally the phase-front temperature boundary conditions (19) result in the terms

$$a_{41} = \xi_1^{-}(0,l), \tag{51}$$

$$a_{42} = \xi_1^{+}(0,l), \tag{52}$$

$$a_{44} = -C_{1,0}\,\xi_1^{-}(0,-1) + D_{1,0}\,\xi_1^{+}(0,-1), \tag{53}$$

$$b_4 = -C_{1,0}\,\xi_1^{r-}(l,0) - D_{1,0}\,\xi_1^{r+}(l,0) - \sum_{n=1}^{l-1}\Big[C_{1,l-n}\,\xi_1^{-}(n,l-n) + D_{1,l-n}\,\xi_1^{+}(n,l-n)\Big]; \tag{54}$$

$$a_{53} = \xi_2^{-}(0,l), \tag{55}$$

$$a_{55} = -C_{2,0}\,\xi_2^{-}(0,-1) + \frac{\xi_2^{+}(0,-1)}{2\sqrt{\pi}\Theta_{20}}, \tag{56}$$

$$b_5 = -C_{2,0}\,\xi_2^{r-}(l,0) - \frac{\xi_2^{r+}(l,0)}{2\sqrt{\pi}\Theta_{20}} - \sum_{n=1}^{l-1} C_{2,l-n}\,\xi_2^{-}(n,l-n). \tag{57}$$

References

1. Broadbridge, P., Triadis, D., Hill, J.M.: Infiltration from supply at constant water content: an integrable model. J. Eng. Math. **64**, 193–206 (2009)
2. Broadbridge, P., Tritscher, P.: An integrable fourth-order nonlinear evolution equation applied to thermal grooving of metal surfaces. IMA J. Appl. Math. **53**(3), 249–265 (1994)
3. Broadbridge, P., Tritscher, P., Avagliano, A.: Free-boundary problems with nonlinear diffusion. Math. Comput. Modell. **18**(10), 15–34 (1993)
4. Green, W.H., Ampt, G.A.: Studies on soil physics part I – the flow of air and water through soils. J. Agric. Sci. **4**, 1–24 (1911)
5. Kingston, J.G., Rogers, C.: Reciprocal Bäcklund-transformations of conservation-laws. Phys. Lett. A **92**(6), 261–264 (1982)
6. Mazumdar, D., Chakraborty, S., Bhambure, S., Patil, S., Chowdhury, S.: An experimental and computational study of casting of large round steel ingots. In: Asia Steel 2015 Proceedings, pp. 244–245
7. Rogers, C.: Application of a reciprocal transformation to a two-phase Stefan problem. J. Phys. A - Math. Gen. **18**(3), L105–L109 (1985)
8. Rogers, C.: On a class of moving boundary-problems in nonlinear heat-conduction – application of a Bäcklund transformation. Int. J. Non-Linear Mech. **21**(4), 249–256 (1986)
9. Storm, M.L.: Heat conduction in simple metals. J. Appl. Phys. **22**(7), 940–951 (1951)
10. Tao, L.N.: On free boundary problems with arbitrary initial and flux conditions. J. Appl. Math. Phys. (ZAMP) **30**, 416–426 (1979)
11. Tao, L.N.: The heat-conduction problem with temperature-dependent material properties. Int. J. Heat Mass Transf. **32**(3), 487–491 (1989)

12. Triadis, D., Broadbridge, P.: Analytical model of infiltration under constant-concentration boundary conditions. Water Resour. Res. **46**, W03526 (2010)
13. Triadis, D., Broadbridge, P.: The Green-Ampt limit with reference to infiltration coefficients. Water Resour. Res. **48**, W07515 (2012)
14. Tritscher, P., Broadbridge, P.: A similarity solution of a multiphase Stefan problem incorporating general nonlinear heat-conduction. Int. J. Heat Mass Transf. **37**(14), 2113–2121 (1994)

Index

B. Anderssen et al. (eds.), *The Role and Importance of Mathematics in Innovation*, Mathematics for Industry 25,
DOI 10.1007/978-981-10-0962-4