Springer Proceedings in Mathematics & Statistics

Taher Abualrub
Abdul Salam Jarrah
Sadok Kallel
Hana Sulieman *Editors*

Mathematics Across Contemporary Sciences

AUS-ICMS, Sharjah, UAE, April 2015

Springer

Springer Proceedings in Mathematics & Statistics

Volume 190

Springer Proceedings in Mathematics & Statistics

This book series features volumes composed of selected contributions from workshops and conferences in all areas of current research in mathematics and statistics, including operation research and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

More information about this series at http://www.springer.com/series/10533

Taher Abualrub · Abdul Salam Jarrah
Sadok Kallel · Hana Sulieman
Editors

Mathematics Across Contemporary Sciences

AUS-ICMS, Sharjah, UAE, April 2015

Springer

Editors
Taher Abualrub
Department of Mathematics and Statistics
American University of Sharjah
Sharjah
United Arab Emirates

Abdul Salam Jarrah
Department of Mathematics and Statistics
American University of Sharjah
Sharjah
United Arab Emirates

Sadok Kallel
Department of Mathematics and Statistics
American University of Sharjah
Sharjah
United Arab Emirates

and

Laboratoire Painlevé
Université de Lille 1
Lille
France

Hana Sulieman
Department of Mathematics and Statistics
American University of Sharjah
Sharjah
United Arab Emirates

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-319-46309-4 ISBN 978-3-319-46310-0 (eBook)
DOI 10.1007/978-3-319-46310-0

Library of Congress Control Number: 2016955689

Mathematics Subject Classification (2010): 62P06, 35A99, 05B20, 55M30, 94C99, 60A06

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This volume presents peer-reviewed contributions covering different areas of mathematics and its applications. Many of these articles were presented at the second International Conference on Mathematics and Statistics, which was held on April 2–5, 2015 at the American University of Sharjah in the United Arab Emirates (UAE). This conference was jointly organized by AUS and the American Mathematical Society. It was attended by over 220 participants from around the world.

The main objective of the conference, which is held every 5 years, is to offer a forum where researchers and scientists working in all fields of mathematics and statistics, and from both academia and industry, can come together, exchange ideas, and learn about recent developments in mathematical research. The conference also aims at promoting the image of mathematics and mathematical research in the UAE and the Gulf region.

In addition to keynote lectures delivered by renowned mathematicians and parallel sessions in all areas of mathematics and statistics, the scientific program of the conference included special sessions in the areas of "commutative algebra and co-algebras", "designs, codes, and graphs", "mathematical applications in biology and medicine", "mathematical finance and probability", "number theory", and "topology and geometry". These special sessions attracted a host of specialists from all five continents. Keynote lectures were delivered by Gunnar Carlsson (Stanford University), Samad Hedayat (University of Illinois at Chicago), and Edriss Titi (Texas A&M). A short address was also given by Richard A. Brualdi (University of Wisconsin, Madison) as the representative of the American Mathematical Society.

The authors of the fine contributions in this volume are mathematicians who are working in or have collaborative ties with mathematicians in the Gulf region. This testifies to the rapid advances in mathematical research that the region is experiencing and to the enormous importance bestowed on academia during the last two decades.

The American University of Sharjah (AUS) is a leading institution of higher education in the UAE. It was founded in 1997 by His Highness Sheikh Dr. Sultan

bin Muhammad Al Qasimi, a member of the Supreme Council of the UAE and ruler of Sharjah. Despite being young in age, AUS is acclaimed throughout the region for its academic excellence and multicultural campus life. Its faculty members come from more than 50 countries while its students hail from more than 90 countries.

This conference was sponsored by the American Mathematical Society, the Mediterranean Institute for the Mathematical Sciences, and Wiley publishing.

We warmly thank all contributors to this volume and all participants who made this conference a real success.

Sharjah, United Arab Emirates
May 2016

Taher Abualrub
Abdul Salam Jarrah
Sadok Kallel
Hana Sulieman

Contents

About the Editors

Taher Abualrub is Professor of Mathematics at the American University of Sharjah. He received his Master and Ph.D. degrees in Mathematics from the University of Iowa, USA, in 1994 and 1998, respectively. In 1998 he also joined the American University of Sharjah (AUS) as Assistant Professor in the Department of Mathematics and Statistics. His current research interests include error correcting codes, DNA computing, wavelet theory, and control theory. He serves as a reviewer for several scientific journals including Math Review, IEEE Transactions on Information Theory, Applied Discrete Mathematics, and Journal of the Franklin Institute, and has published over 45 refereed journal articles and conference proceedings papers.

Abdul Salam Jarrah is Professor of Mathematics at the American University of Sharjah (AUS) in the United Arab Emirates. Before joining the AUS, he was a senior research scientist at the Virginia Bioinformatics Institute at Virginia Tech, USA. He received his Ph.D. in Mathematics from New Mexico State University in 2002. His current research interests include discrete dynamical systems, computational algebra, and their applications in biology, especially the modeling and simulation of biological systems. He has made valuable contributions to these fields through many publications in international journals. He also serves as a reviewer for several scientific journals including Mathematical Biosciences, Bulletin of Mathematical Biology, and IEEE Transactions on Computational Biology.

Sadok Kallel holds a Ph.D. in Mathematics from Stanford University (1994) and a Bachelor's degree in Electrical Engineering from Michigan State University. He is currently a faculty member at the American University of Sharjah and a research fellow with the "Laboratoire Painlevé" at the University of Lille 1 in France, where he was a faculty member from 1999 to 2011. His main research interests are in algebraic and differential topology, with a focus on applications to nonlinear analysis and mathematical physics. He is the director of the "Mediterranean Institute for the Mathematical Sciences" in Tunis, Tunisia, and the managing editor for the journal GSMD (Graduate Student Mathematical Diary).

Hana Sulieman is Professor of Statistics and Head of the Department of Mathematics and Statistics at the American University of Sharjah in the United Arab Emirates (UAE). Before joining the AUS, she was an associate researcher at Queen's University, Kingston, Canada, where she also completed her Ph.D. in Applied Statistics. She has many years of experience in statistical consulting in Canada and the UAE. Her primary research interests are sensitivity analysis in nonlinear parameter estimation and statistical design of experiments. Through consulting activities, she has also developed other research interests related to statistical applications and modeling in the health sciences and education. She has authored numerous publications in international peer-reviewed journals and given many presentations at international and regional conferences. Further, she is a regular reviewer for international journals in the fields of applied statistics and data analysis.

The Lusternik-Schnirelmann Category for a Differentiable Stack

Samirah Alsulami, Hellen Colman and Frank Neumann

Abstract We introduce the notion of Lusternik-Schnirelmann category for differentiable stacks and establish its relation with the groupoid Lusternik-Schnirelmann category for Lie groupoids. This extends the notion of Lusternik-Schnirelmann category for smooth manifolds and orbifolds.

Keywords Differentiable stacks · Lie groupoids · Orbifolds · LS-category

2010 Mathematics Subject Classication. 55M30 · 14A20 · 14D23 · 22A22

1 Introduction

The Lusternik-Schnirelmann category or LS-category of a manifold is a numerical invariant introduced by Lusternik and Schnirelmann [21] in the early 1930s as a lower bound on the number of critical points for any smooth function on a compact smooth manifold. Later it was shown that the Lusternik-Schnirelmann category is in fact a homotopy invariant and it became an important tool in algebraic topology and especially homotopy theory. For an overview and survey on the importance of LS-category in topology and geometry we refer the reader to [12, 18, 19].

Fundamental in the definition of LS-category of a smooth manifold or topological space is the concept of a categorical set. A subset of a space is said to be categorical

S. Alsulami · F. Neumann (✉)
Department of Mathematics, University of Leicester, University Road
Leicester LE1 7RH, England, UK
e-mail: fn8@le.ac.uk

S. Alsulami
e-mail: shba2@le.ac.uk

H. Colman
Department of Mathematics, Wilbur Wright College, 4300 N. Narragansett Avenue,
Chicago, IL 60634, USA
e-mail: hcolman@ccc.edu

T. Abualrub et al. (eds.), *Mathematics Across Contemporary Sciences*,
Springer Proceedings in Mathematics & Statistics 190,
DOI 10.1007/978-3-319-46310-0_1

if it is contractible in the space. The Lusternik-Schnirelmann category $\mathrm{cat}(X)$ of a smooth manifold X is defined to be the least number of categorical open sets required to cover X, if that number is finite, otherwise the category $\mathrm{cat}(X)$ is said to be infinite.

In this article, we generalize the notion of Lusternik-Schnirelmann category to differentiable stacks with the intention of providing a useful tool and invariant to study homotopy theory, the theory of geodesics and Morse theory of differentiable stacks. Differentiable stacks naturally generalize smooth manifolds and orbifolds and are therefore of interest in many areas of geometry, topology and mathematical physics. They are basically generalized smooth spaces where its points also have automorphism groups. For example, they often appear as an adequate replacement of quotients for general Lie group actions on smooth manifolds, especially when the naive quotient does not exist as a smooth manifold. Many moduli and classification problems like for example the classification of Riemann surfaces or vector bundles on Riemann surfaces naturally lead to the notion of a differentiable stack. It can be expected that the stacky LS-category will be a very useful topological invariant for these kind of generalized smooth spaces which appear naturally in geometry and physics. We aim to study the geometrical and topological aspects of the stacky LS-category and its applications in a follow-up article [11]. Many of these constructions can also be presented in a purely homotopical manner [2] by employing the homotopy theory of topological stacks [24, 25].

The new notion of stacky LS-category for differentiable stacks presented here employs the notion of a categorical substack and is again an invariant of the homotopy type of the differentiable stack, in fact of the underlying topological stack. It generalizes the classical LS-category for manifolds [21] and we show that it is directly related with the groupoid LS-category for Lie groupoids as defined by Colman [9] for Lie groupoids.

The material of this article is organised as follows: In the first section we collect the basic definitions of differentiable stacks and Lie groupoids and establish some fundamental properties. In particular we exhibit the various connections between differentiable stacks and Lie groupoids. The second section recalls the foundations of Lusternik-Schnirelmann category for groupoids and its Morita invariance. In the third section we introduce the new notion of stacky Lusternik-Schnirelmann category and establish its relationship with the groupoid LS-category of the various groupoids introduced in the Sect. 2.

2 Differentiable Stacks and Lie Groupoids

In this section we will collect in detail the notions and some of the fundamental properties of differentiable stacks and Lie groupoids, which we will use later. We refer the reader to various resources on differentiable stacks [3–6, 13, 16] and on Lie groupoids [14, 20, 23, 28] for more details and specific examples and their interplay.

Differentiable stacks are defined over the category of smooth manifolds. A smooth manifold here will always mean a finite dimensional second countable smooth

manifold, which need not necessarily be Hausdorff. We denote the category of smooth manifolds and smooth maps by $\mathfrak{S}$.

A *submersion* is a smooth map $f : U \to X$ such that the derivative $f_* : T_u U \to T_{f(u)} X$ is surjective for all points $u \in U$. The dimension of the kernel of the linear map f_* is a locally constant function on U and called the *relative dimension* of the submersion f. An *étale morphism* is a submersion of relative dimension 0. This means that a morphism f between smooth manifolds is *étale* if and only if f is a local diffeomorphism.

The *étale site* $\mathfrak{S}_{et}$ on the category $\mathfrak{S}$ is given by the following Grothendieck topology on $\mathfrak{S}$. We call a family $\{U_i \to X\}$ of morphisms in $\mathfrak{S}$ with target X a *covering family* of X, if all smooth maps $U_i \to X$ are étale and the total map $\coprod_i U_i \to X$ is surjective. This defines a pretopology on $\mathfrak{S}$ generating a Grothendieck topology, the *étale topology* on $\mathfrak{S}$ (see [1], Exposé II).

As remarked in [5], not all fibre products for two morphisms $U \to X$ and $V \to X$ exist in $\mathfrak{S}$, but if at least one of the two morphisms is a submersion, then the fibre product $U \times_X V$ exists in $\mathfrak{S}$, which will be enough, while dealing with differentiable stacks.

Definition 2.1 A *category fibred in groupoids* over $\mathfrak{S}$ is a category $\mathfrak{X}$, together with a functor $\pi_{\mathfrak{X}} : \mathfrak{X} \to \mathfrak{S}$ such that the following axioms hold:

(i) For every morphism $V \to U$ in $\mathfrak{S}$, and every object x of $\mathfrak{X}$ lying over U, there exists an arrow $y \to x$ in $\mathfrak{X}$ lying over $V \to U$.
(ii) For every commutative triangle $W \to V \to U$ in $\mathfrak{S}$ and morphisms $z \to x$ lying over $W \to U$ and $y \to x$ lying over $V \to U$, there exists a unique arrow $z \to y$ lying over $W \to V$ such that the composition $z \to y \to x$ is the morphism $z \to x$.

The axiom (ii) ensures that the object y over V, which exists after (i) is unique up to a unique isomorphism. Any choice of such an object y is called a *pullback* of x via the morphism $f : V \to U$. We will write as usual $y = x|V$ or $y = f^*x$.

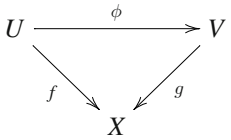
Let $\mathfrak{X}$ be a category fibred in groupoids over $\mathfrak{S}$. Occasionally we will denote by X_0 the collection of objects and X_1 the collection of arrows of the category $\mathfrak{X}$. The subcategory of $\mathfrak{X}$ of all objects lying over a fixed object U of $\mathfrak{S}$ with morphisms being those lying over the identity morphism id_U is called the *fibre* or *category of sections* of $\mathfrak{X}$ over U, which will be denoted by $\mathfrak{X}_U$ or $\mathfrak{X}(U)$. By definition all fibres $\mathfrak{X}_U$ are discrete groupoids.

Categories fibred in groupoids form a 2-category, denoted by CFG. The 1-morphisms are given by functors $\phi : \mathfrak{X} \to \mathfrak{Y}$ respecting the projection functors i.e. $\pi_{\mathfrak{Y}} \circ \phi = \pi_{\mathfrak{X}}$ and the 2-morphisms are given by natural transformations between these functors preserving projection functors. Fibre products exist in CFG (see [15]).

We will say that the categories fibred in groupoids $\mathfrak{X}$ and $\mathfrak{Y}$ are *isomorphic* if there are 1-morphisms $\phi : \mathfrak{X} \to \mathfrak{Y}$ and $\psi : \mathfrak{Y} \to \mathfrak{X}$ and 2-morphisms T and T' such that $T : \phi \circ \psi \Rightarrow \mathrm{id}_{\mathfrak{Y}}$ and $T' : \psi \circ \phi \Rightarrow \mathrm{id}_{\mathfrak{X}}$.

Example 2.2 (Identity) Let $\mathfrak{X}$ be the fixed category $\mathfrak{S}$. Let $\pi = \mathrm{id}_{\mathfrak{S}} : \mathfrak{S} \to \mathfrak{S}$ be the projection functor. Then $\mathfrak{X} = \mathfrak{S}$ together with the identity map is a category fibred in groupoids.

Example 2.3 (Object) Given a fixed object $X \in \mathfrak{S}$, i.e. a smooth manifold, consider the category $\underline{X}$ whose objects are (U, f) where $f : U \to X$ is a morphism in $\mathfrak{S}$ and U an object in $\mathfrak{S}$, and whose arrows are diagrams



The projection functor is $\pi : \underline{X} \to \mathfrak{S}$ with $\pi((U, f)) = U$ and $\pi((U, f), \phi, (V, g)) = \phi$. We have that $\underline{X}$ is a category fibred in groupoids.

In particular, in case that X is a point, $X = *$, we have that $\underline{*} = \mathfrak{S}$.

Example 2.4 (Sheaves) Let $F : \mathfrak{S} \to (Sets)$ be a presheaf, i.e. a contravariant functor. We get a category fibred in groupoids $\mathfrak{X}$, where the objects are pairs (U, x), with U a smooth manifold and $x \in F(U)$ and a morphism $(U, x) \to (V, y)$ is a smooth map $f : U \to V$ such that $x = y|_F U$, i.e. $x = F(f)(y)$. The projection functor is given by

$$\pi : \mathfrak{X} \to \mathfrak{S}, \ (U, x) \mapsto U.$$

Especially any sheaf $F : \mathfrak{S} \to (Sets)$ gives therefore a category fibred in groupoids over $\mathfrak{S}$ and in particular we see again that every smooth manifold X gives a category fibred in groupoids $\underline{X}$ over $\mathfrak{S}$ as the sheaf represented by X, i.e. where

$$\underline{X}(U) = Hom_{\mathfrak{S}}(U, X).$$

To simplify notation, we will sometimes freely identify $\underline{X}$ with the smooth manifold X.

Now let us recall the definition of a stack [5]. In the following let $\mathfrak{S}$ always be the category of smooth manifolds equipped with the étale topology as defined above. We could of course replace the étale topology with any other Grothendieck topology on $\mathfrak{S}$, but in this article we are mainly interested in stacks over the étale site $\mathfrak{S}_{et}$.

Definition 2.5 A category fibred in groupoids $\mathfrak{X}$ over $\mathfrak{S}$ is a *stack* over $\mathfrak{S}$ if the following gluing axioms hold:

(i) For any smooth manifold X in $\mathfrak{S}$, any two objects x, y in $\mathfrak{X}$ lying over X and any two isomorphisms $\phi, \psi : x \to y$ over X, such that $\phi|U_i = \psi|U_i$ for all U_i in a covering $\{U_i \to X\}$ it follows that $\phi = \psi$.

(ii) For any smooth manifold X in $\mathfrak{S}$, any two objects $x, y \in \mathfrak{X}$ lying over X, any covering $\{U_i \to X\}$ and, for every i, an isomorphism $\phi_i : x|U_i \to y|U_i$, such that $\phi|U_{ij} = \phi_j|U_{ij}$ for all i, j, there exists an isomorphism $\phi : x \to y$ with $\phi|U_i = \phi$ for all i.

(iii) For any smooth manifold X in $\mathfrak{S}$, any covering $\{U_i \to X\}$, any family $\{x_i\}$ of objects x_i in the fibre $\mathfrak{X}_{U_i}$ and any family of morphisms $\{\phi_{ij}\}$, where $\phi_{ij} : x_i|U_{ij} \to x_j|U_{ij}$ satisfying the cocycle condition $\phi_{jk} \circ \phi_{ij} = \phi_{ik}$ in $\mathfrak{X}(U_{ijk})$

there exist an object x lying over X with isomorphisms $\phi_i : x|U_i \to x_i$ such that $\phi_{ij} \circ \phi_i = \phi_j$ in $\mathfrak{X}(U_{ij})$.

The isomorphism ϕ in (ii) is unique by (i) and similar from (i) and (ii) it follows that the object x whose existence is asserted in (iii) is unique up to a unique isomorphism. All pullbacks mentioned in the definitions are also only unique up to isomorphism, but the properties do not depend on choices.

In order to do geometry on stacks, we have to compare them with smooth manifolds.

A category $\mathfrak{X}$ fibred in groupoids over $\mathfrak{S}$ is *representable* if there exists a smooth manifold X such that $\underline{X}$ is isomorphic to $\mathfrak{X}$ as categories fibred in groupoids over $\mathfrak{S}$.

A morphism of categories fibred in groupoids $\mathfrak{X} \to \mathfrak{Y}$ is *representable*, if for every smooth manifold U and every morphism $\underline{U} \to \mathfrak{Y}$ the fibred product $\mathfrak{X} \times_{\mathfrak{Y}} \underline{U}$ is representable.

A morphism of categories fibred in groupoids $\mathfrak{X} \to \mathfrak{Y}$ is a *representable submersion*, if it is representable and the induced morphism of smooth manifolds $\mathfrak{X} \times_{\mathfrak{Y}} U \to U$ is a submersion for every smooth manifold U and every morphism $\underline{U} \to \mathfrak{Y}$.

Definition 2.6 A stack $\mathfrak{X}$ over $\mathfrak{S}$ is a *differentiable* or *smooth stack* if there exists a smooth manifold X and a surjective representable submersion $x : X \to \mathfrak{X}$, i.e. there exists a smooth manifold X together with a morphism of stacks $x : X \to \mathfrak{X}$ such that for every smooth manifold U and every morphism of stacks $U \to \mathfrak{X}$ the fibre product $X \times_{\mathfrak{X}} U$ is representable and the induced morphism of smooth manifolds $X \times_{\mathfrak{X}} U$ is a surjective submersion.

If $\mathfrak{X}$ is a differentiable stack, a surjective representable submersion $x : X \to \mathfrak{X}$ as before is called a *presentation of* $\mathfrak{X}$ or *atlas for* $\mathfrak{X}$. It need not be unique, i.e. a differentiable stack can have different presentations.

Example 2.7 All representable stacks are differentiable stacks. Let X be a smooth manifold. The category fibred in groupoids $\underline{X}$ is in fact a differentiable stack over $\mathfrak{S}$ since it is representable. A presentation is given by the identity morphism id_X, which is in fact a diffeomorphism.

Example 2.8 (Torsor) Let G be a Lie group. Consider the category $\mathfrak{B}G$ which has as objects principal G-bundles (or G-torsors) P over S and as arrows commutative diagrams

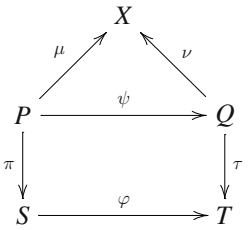
$$\begin{array}{ccc} P & \xrightarrow{\psi} & Q \\ {\scriptstyle\pi}\downarrow & & \downarrow{\scriptstyle\tau} \\ S & \xrightarrow{\varphi} & T \end{array}$$

where the map $\psi : P \to Q$ is equivariant.

The category $\mathfrak{B}G$ together with the projection functor $\pi : \mathfrak{B}G \to \mathfrak{S}$ given by $\pi(P \to S) = S$ and $\pi((\psi, \varphi)) = \varphi$ is a category fibred in groupoids, in fact

a differentiable stack, the *classifying stack* of G whose atlas presentation is given by the representable surjective submersion $* \to \mathfrak{B}G$.

Example 2.9 (Quotient Stack) Let X be a smooth manifold with a smooth (left) action $\rho : G \times X \to X$ by a Lie group G. Let $[X/G]$ be the category which has as objects triples (P, S, μ), where S is a smooth manifold of $\mathfrak{S}$, P a principal G-bundle (or G-torsor) over S and $\mu : P \to X$ a G-equivariant smooth map. A morphism $(P, S, \mu) \to (Q, T, \nu)$ is a commutative diagram



where $\psi : P \to Q$ is a G-equivariant map. Then $[X/G]$ together with the projection functor $\pi : [X/G] \to \mathfrak{S}$ given by $\pi((P, S, \mu)) = S$ and $\pi((\psi, \varphi)) = \varphi$ is a category fibred in groupoids over $\mathfrak{S}$. $[X/G]$ is in fact a differentiable stack, the *quotient stack* of G. An atlas is given by the representable surjective submersion $x : X \to [X/G]$.

If $X = *$ is just a point, we simply recover the differentiable stack $\mathfrak{B}G$ as defined in the previous example, i.e. $[*/G] = \mathfrak{B}G$.

In some way, quotient stacks encode in a non-equivariant and systematic way various equivariant data of general Lie group actions, which need not to be free.

Differentiable stacks are basically incarnations of Lie groupoids.

Definition 2.10 A *Lie groupoid* $\mathcal{G}$ is a groupoid in the category $\mathfrak{S}$ of smooth manifolds, i.e.

$$G_1 \rightrightarrows G_0$$

such that the space of arrows G_1 and the space of objects G_0 are smooth manifolds and all structure morphisms

$$m : G_1 \times_{G_0} G_1 \to G_1,\ s, t : G_1 \to G_0,\ i : G_1 \to G_0,\ e : G_0 \to G_1$$

are smooth maps and additionally the source map s and the target map t are submersions.

Morphisms between Lie groupoids are given by functors (ϕ_1, ϕ_0) where $\phi_i : G_i \to G_i'$ is a smooth map. We will call them *strict morphisms*.

Lie groupoids, strict morphisms and natural transformations form a 2-category that we denote LieGpd.

We will show now a construction of a Lie groupoid associated to a differentiable stack. Let $\mathfrak{X}$ be a differentiable stack with a given presentation $x : X \to \mathfrak{X}$. We can

associate to $(\mathfrak{X}, x)$ a Lie groupoid $\mathcal{G}(x) = G_1 \rightrightarrows G_0$ as follows: Let $G_0 = X$ and $G_1 = X \times_{\mathfrak{X}} X$. The source and target morphisms $s, t : X \times_{\mathfrak{X}} X \rightrightarrows X$ of $\mathcal{G}$ are given as the two canonical projection morphisms. The composition of morphisms m in $\mathcal{G}$ is given as projection to the first and third factor

$$X \times_{\mathfrak{X}} X \times_{\mathfrak{X}} X \cong (X \times_{\mathfrak{X}} X) \times_X (X \times_{\mathfrak{X}} X) \to X \times_{\mathfrak{X}} X.$$

The morphism which interchanges factors $X \times_{\mathfrak{X}} X \to X \times_{\mathfrak{X}} X$ gives the inverse morphism i and the unit morphism e is given by the diagonal morphism $X \to X \times_{\mathfrak{X}} X$. Because the presentation $x : X \to \mathfrak{X}$ of a differentiable stack is a submersion, it follows that the source and target morphisms $s, t : X \times_{\mathfrak{X}} X \rightrightarrows X$ are submersions as induced maps of the fibre product.

The Lie groupoid associated to the differentiable stack $\mathfrak{X}$ in this way is also part of a simplicial smooth manifold $X_\bullet$, whose homotopy type encodes the homotopy type of $\mathfrak{X}$ [14, 25, 26].

Given instead a Lie groupoid we can associate a differentiable stack to it. Basically this is a generalization of Example 2.8 where we associate to a Lie group G the classifying stack $\mathfrak{B}G$.

Let's introduce first the notions of *groupoid action* and *groupoid torsor*.

Definition 2.11 (Action of a groupoid on a manifold) Let $\mathcal{G}$ be a Lie groupoid $G_1 \rightrightarrows G_0$ and P a manifold in $\mathfrak{S}$. An *action of* $\mathcal{G}$ *on* P is given by

(i) an anchor map $a : P \to G_0$,
(ii) an action map $\mu : G_1 \times_{s,a} P \to P$

such that $t(k) = a(k \cdot p)$ for all $(k, p) \in G_1 \times P$ with $s(k) = a(p)$, satisfying the standard action properties: $e(a(p)) \cdot p = p$ and $(k \cdot p) \cdot h = (gh) \cdot p$ whenever the operations are defined.

Definition 2.12 (Groupoid torsor) Let $\mathcal{G}$ be a Lie groupoid $G_1 \rightrightarrows G_0$ and S a manifold in $\mathfrak{S}$. A $\mathcal{G}$*-torsor over* S is given by

(i) a manifold P together with
(ii) a surjective submersion $\pi : P \to S$ and
(iii) an action of $\mathcal{G}$ on P

such that for all $p, p' \in P$ with $\pi(p) = \pi(p')$ there exists a unique $k \in G_1$ such that $k \cdot p = p'$.

Let $\pi : P \to S$ and $\rho : Q \to T$ be $\mathcal{G}$-torsors. A *morphism* of $\mathcal{G}$-torsors is given by a commutative diagram

$$\begin{array}{ccc} P & \xrightarrow{\psi} & Q \\ {\scriptstyle\pi}\downarrow & & \downarrow{\scriptstyle\tau} \\ S & \xrightarrow{\varphi} & T \end{array}$$

such that ψ is a $\mathcal{G}$-equivariant map.

Example 2.13 ($\mathcal{G}$-torsors) Let $\mathcal{G}$ be a fixed Lie groupoid $G_1 \rightrightarrows G_0$. Consider the category $\mathfrak{B}\mathcal{G}$ which has as objects $\mathcal{G}$-torsors P over S and as arrows morphisms of $\mathcal{G}$-torsors as described above.

Then $\mathfrak{B}\mathcal{G}$ is a category fibred in groupoids over $\mathfrak{S}$ with projection functor $\pi : \mathfrak{B}\mathcal{G} \to \mathfrak{S}$ given by $\pi(P \to S) = S$ and $\pi((\psi, \varphi)) = \varphi$. Moreover, $\mathfrak{B}\mathcal{G}$ is a differentiable stack.

We have the following fundamental property (see for example [5, Prop. 2.3]) of $\mathfrak{B}\mathcal{G}$.

Theorem 2.14 *For every Lie groupoid $\mathcal{G} = G_1 \rightrightarrows G_0$ the category fibred in groupoids $\mathfrak{B}\mathcal{G}$ of $\mathcal{G}$-torsors is a differentiable stack with a presentation $\tau_0 : G_0 \to \mathfrak{B}\mathcal{G}$.*

The stack $\mathfrak{B}\mathcal{G}$ is also called the *classifying stack* of $\mathcal{G}$-torsors. It follows from this also that the Lie groupoid $\mathcal{G} = G_1 \rightrightarrows G_0$ is isomorphic to the Lie groupoid $\mathcal{G}(\tau_0)$ associated to the atlas $\tau_0 : G_0 \to \mathfrak{B}\mathcal{G}$ of the stack $\mathfrak{B}\mathcal{G}$.

Under this correspondence between differentiable stacks and Lie groupoids, the quotient stack $[X/G]$ of an action of a Lie group G on a smooth manifold X as described in Example 2.9 corresponds to the action groupoid $G \times X \rightrightarrows X$ [5, 16]. In particular, if X is just a point the classifying stack $\mathfrak{B}G$ of a Lie group G corresponds to the Lie groupoid $G \rightrightarrows *$.

As the presentations of a differentiable stack are not unique, the associated Lie groupoids might be different. In order to define algebraic invariants, like cohomology or homotopy groups for differentiable stacks they should however not depend on a chosen presentation of the stack. Therefore it is important to know, when two different Lie groupoids give rise to isomorphic stacks. This will be the case when the Lie groupoids are Morita equivalent.

Definition 2.15 Let $\mathcal{G} = G_1 \rightrightarrows G_0$ and $\mathcal{H} = H_1 \rightrightarrows H_0$ be Lie groupoids. A morphism of Lie groupoids is a smooth functor $\phi : \mathcal{G} \to \mathcal{H}$ given by two smooth maps $\phi = (\phi_1, \phi_0)$ with

$$\phi_0 : G_0 \to H_0,\ \phi_1 : G_1 \to H_1$$

which commute with all structure morphisms of the groupoids. A morphism $\phi : \mathcal{G} \to \mathcal{H}$ of Lie groupoids is a *Morita morphism* or *essential equivalence* if

(i) $\phi_0 : G_0 \to H_0$ is a surjective submersion,
(ii) the diagram

$$\begin{array}{ccc} G_1 & \xrightarrow{(s,t)} & G_0 \times G_0 \\ {\scriptstyle \phi_1}\downarrow & & \downarrow{\scriptstyle \phi_0 \times \phi_0} \\ H_1 & \xrightarrow{(s,t)} & H_0 \times H_0 \end{array}$$

is cartesian, i.e. $G_1 \cong H_1 \times_{H_0 \times H_0} G_0 \times G_0$.

Two Lie groupoids $\mathcal{G}$ and $\mathcal{H}$ are *Morita equivalent*, if there exists a third Lie groupoid $\mathcal{K}$ and Morita morphisms

$$\mathcal{G} \overset{\phi}{\leftarrow} \mathcal{K} \overset{\psi}{\rightarrow} \mathcal{H}$$

We have the following main theorem concerning the relation of the various Lie groupoids associated to various presentations of a differentiable stack [5, Theorem 2.24].

Theorem 2.16 *Let $\mathcal{G}$ and $\mathcal{H}$ be Lie groupoids. Let $\mathfrak{X}$ and $\mathfrak{Y}$ be the associated differentiable stacks, i.e. $\mathfrak{X} = \mathfrak{B}\mathcal{G}$ and $\mathfrak{Y} = \mathfrak{B}\mathcal{H}$. Then the following are equivalent:*

(i) the differentiable stacks $\mathfrak{X}$ and $\mathfrak{Y}$ are isomorphic,
(ii) the Lie groupoids $\mathcal{G}$ and $\mathcal{H}$ are Morita equivalent.

As a special case we have the following fundamental property concerning different presentations of a differentiable stack $\mathfrak{X}$.

Proposition 2.17 *Let $\mathfrak{X}$ be a differentiable stack with two given presentations $x : X \to \mathfrak{X}$ and $x' : X' \to \mathfrak{X}$. Then the associated Lie groupoids $\mathcal{G}(x)$ and $\mathcal{G}(x')$ are Morita equivalent.*

Therefore Lie groupoids present isomorphic differentiable stacks if and only if they are Morita equivalent or in other words differentiable stacks correspond to Morita equivalence classes of Lie groupoids.

We now recall the fundamental notion of a smooth morphism between differentiable stacks (see for example [16]).

Definition 2.18 (*Smooth morphism*) An arbitrary morphism $\mathfrak{X} \to \mathfrak{Y}$ of differentiable stacks is *smooth*, if there are atlases $X \to \mathfrak{X}$ and $Y \to \mathfrak{Y}$ such that the induced morphism from the fibered product $X \times_{\mathfrak{Y}} Y \to Y$ in the diagram below is a smooth map between manifolds.

$$\begin{array}{ccc} X \times_{\mathfrak{Y}} Y & \longrightarrow & X \\ \downarrow & & \downarrow \\ & & \mathfrak{X} \\ & & \downarrow \\ Y & \longrightarrow & \mathfrak{Y} \end{array}$$

Let $\mathfrak{U}$ be a subcategory of $\mathfrak{X}$. Recall that a subcategory is called *saturated* if whenever it contains an object x then it contains the entire isomorphism class $\bar{x}$ of that object and is called *full* if whenever it contains an arrow between x and y, it contains the entire set $\mathrm{Hom}(x, y)$ of arrows.

Let $\pi : \mathfrak{X} \to \mathfrak{S}$ be a differentiable stack and $x : X \to \mathfrak{X}$ be an atlas. Let $U \subset X$ be an open subset and consider the saturation U_0 of the image $x(U)$ in X_0, i.e.

$$U_0 = \{z \in X_0 |\ z \in \bar{x} \text{ for some } x \in U\}.$$

The full subcategory $\mathfrak{U}$ on U_0 is $U_1 \rightrightarrows U_0$ where $U_1 = \{g \in X_1 |\ s(g), t(g) \in U_0\}$.

Definition 2.19 (*Restricted substack*) Let $\pi : \mathfrak{X} \to \mathfrak{S}$ be a differentiable stack with atlas $x : X \to \mathfrak{X}$ and $U \subset X$ be an open set. Consider the full subcategory $\mathfrak{U}$ on U_0 and let $\pi' := \pi \circ i$, where $i : \mathfrak{U} \to \mathfrak{X}$ is the inclusion. We say that $\mathfrak{U}$ with the projection $\pi' : \mathfrak{U} \to \mathfrak{S}$ is the *restricted substack* of $\mathfrak{X}$ to $\mathfrak{U}$.

Definition 2.20 (*Constant morphism*) Let $c : \mathfrak{X} \to \mathfrak{Y}$ be a smooth morphism between differentiable stacks. We say that c is a *constant morphism* if there are presentations $X \to \mathfrak{X}$ and $Y \to \mathfrak{Y}$ such that the induced morphism from the fibered product $X \times_{\mathfrak{Y}} Y \to Y$ is a constant map.

For instance, any smooth morphism $\mathfrak{X} \to \mathfrak{Y}$ where $\mathfrak{Y}$ admits a presentation by a point $* \to \mathfrak{Y}$ is a constant morphism.

Example 2.21 Let S^1 act on S^1 by rotation and consider the quotient stack $\mathfrak{X}$ associated to this action, $\mathfrak{X} = [S^1/S^1]$. We will show that the identity map $\mathrm{id}_{[S^1/S^1]} : [S^1/S^1] \to [S^1/S^1]$ is a constant map. The groupoid $S^1 \times S^1 \rightrightarrows S^1$ is Morita equivalent to a point groupoid $* \rightrightarrows *$, therefore the stacks $\mathfrak{X} = [S^1/S^1]$ and $\underline{*}$ are isomorphic. Since $* \to \underline{*}$ is a presentation for $\underline{*}$ it follows that $* \to [S^1/S^1]$ is a presentation for $\mathfrak{X}$. Hence any map with codomain $\mathfrak{X} = [S^1/S^1]$ is a constant morphism of stacks.

Let us finish this section by remarking that it is also possible to define stacks and groupoids over the more general category of diffeological spaces instead of using the category of smooth manifolds as we have done here [7, 17, 27].

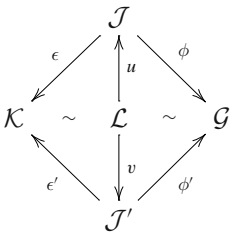
3 Lusternik-Schnirelmann Category for Lie Groupoids

We will first recall the definition and fundamental properties for the notion of Lusternik-Schnirelmann category for Lie groupoids. The most important property here is that the Lusternik-Schnirelmann category of a Lie groupoid is in fact Morita invariant, which means that it is in fact an invariant of the associated differentiable stack. We will follow here the general approach of [9], where the notion of Lusternik-Schnirelmann category for Lie groupoids was first introduced.

Our context will be the Morita bicategory of Lie groupoids $\mathsf{LieGpd}(E^{-1})$ obtained from LieGpd by formally inverting the essential equivalences E. Objects in this bicategory are Lie groupoids, 1-morphisms are *generalized maps*

$$\mathcal{K} \xleftarrow{\epsilon} \mathcal{J} \xrightarrow{\phi} \mathcal{G}$$

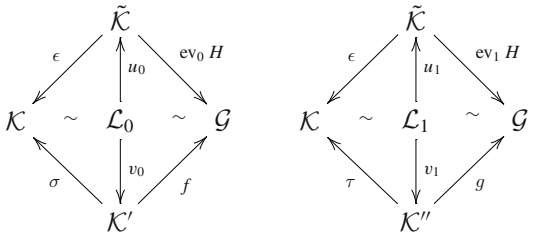
such that ϵ is an essential equivalence and 2-morphisms from $\mathcal{K} \xleftarrow{\epsilon} \mathcal{J} \xrightarrow{\phi} \mathcal{G}$ to $\mathcal{K} \xleftarrow{\epsilon'} \mathcal{J}' \xrightarrow{\phi'} \mathcal{G}$ are given by classes of diagrams:



where $\mathcal{L}$ is a topological groupoid, and u and v are essential equivalences.

The *path groupoid* of $\mathcal{G}$ is defined as the mapping groupoid in this bicategory, $P\mathcal{G} = \mathcal{M}ap(\mathcal{I}, \mathcal{G})$.

Let $(\sigma, f) : \mathcal{K} \xleftarrow{\sigma} \mathcal{K}' \xrightarrow{f} \mathcal{G}$ and $(\tau, g) : \mathcal{K} \xleftarrow{\tau} \mathcal{K}'' \xrightarrow{g} \mathcal{G}$ be generalized maps. The map (σ, f) is *groupoid homotopic* to (τ, g) if there exists $(\epsilon, H) : \mathcal{K} \xleftarrow{\epsilon} \tilde{\mathcal{K}} \xrightarrow{H} P\mathcal{G}$ and two commutative diagrams up to natural transformations:



where $\mathcal{L}_i$ is an action groupoid, and u_i and v_i are equivariant essential equivalences for $i = 0, 1$.

Similarly as for differentiable stacks we also have the concept of a restricted groupoid over a given invariant subset and that of a generalized constant map, which we will need to define LS-category for Lie groupoids.

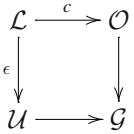
Definition 3.1 Let $\mathcal{G} = G_1 \rightrightarrows G_0$ be a Lie groupoid. An open set $U \subset G_0$ is *invariant* if $t(s^{-1}(U)) \subset U$. The *restricted groupoid* $\mathcal{U}$ to an invariant set $U \subset G_0$ is the full groupoid over U. In other words, $U_0 = U$ and $U_1 = \{g \in G_1 : s(g), t(g) \in U\}$. We write $\mathcal{U} = \mathcal{G}|_U \subset \mathcal{G}$.

Definition 3.2 We say that the map $(\epsilon, c) : \mathcal{K} \xleftarrow{\epsilon} \mathcal{K}' \xrightarrow{c} \mathcal{G}$ is a *generalized constant map* if for all $x \in K_0'$ there exists $g \in G_1$ with $s(g) = x_0$ such that $c(x) = t(g)$ for a fixed $x_0 \in G_0$.

In other words, the image of the generalized map (ϵ, c) is contained in a fixed orbit $\mathcal{O}$, the orbit of x_0.

Definition 3.3 For an invariant open set $U \subset G_0$, we will say that the restricted groupoid $\mathcal{U}$ is $\mathcal{G}$-*categorical* if the inclusion map $i_{\mathcal{U}} : \mathcal{U} \to \mathcal{G}$ is groupoid homotopic to a generalized constant map (ϵ, c).

In other words, the diagram



is commutative up to groupoid homotopy where ϵ is an equivariant essential equivalence and $\mathcal{O}$ an orbit.

Now we can make the following definition (see [10]).

Definition 3.4 Let $\mathcal{G} = G_1 \rightrightarrows G_0$ be a Lie groupoid. The *groupoid Lusternik-Schnirelmann* or *groupoid LS-category*, $\mathrm{cat}(\mathcal{G})$, is the least number of invariant open sets U needed to cover G_0 such that the restricted groupoid $\mathcal{U}$ is $\mathcal{G}$-categorical.

If G_0 cannot be covered by a finite number of such open sets, we will say that $\mathrm{cat}(\mathcal{G}) = \infty$.

We have the following important property of the groupoid Lusternik-Schnirelmann category (see [9]).

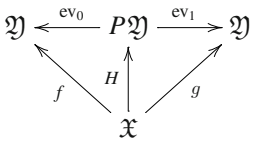
Theorem 3.5 *The Lusternik-Schnirelmann category of a Lie groupoid is invariant under Morita equivalence of Lie groupoids, i.e. if $\mathcal{G}$ is a Lie groupoid which is Morita equivalent to a Lie groupoid $\mathcal{G}'$, then we have*

$$\mathrm{cat}(\mathcal{G}) = \mathrm{cat}(\mathcal{G}').$$

The groupoid Lusternik-Schnirelmann category also generalizes the ordinary Lusternik-Schnirelmann category of a smooth manifold. In fact, if $\mathcal{G} = u(M)$ is the unit groupoid, then we have $\mathrm{cat}(\mathcal{G}) = \mathrm{cat}(M)$, where cat on the right hand side means the ordinary Lusternik-Schnirelmann category of a smooth manifold.

4 Lusternik-Schnirelmann Category of a Differentiable Stack

Using homotopical properties of differentiable stacks we will now introduce the Lusternik-Schnirelmann category of a differentiable stacks. Let $\mathfrak{X}$ be a differentiable stack. Consider the path stack $P\mathfrak{X} = \mathfrak{Map}([0, 1], \mathfrak{X})$ of $\mathfrak{X}$ as defined by Noohi in [24] for general topological stacks. We will say that the morphisms $f : \mathfrak{X} \to \mathfrak{Y}$ and $g : \mathfrak{X} \to \mathfrak{Y}$ between differentiable stacks are *homotopic* if there exists a morphism of stacks $H : \mathfrak{X} \to P\mathfrak{Y}$ such that the following diagram of stack morphisms is commutative up to natural transformations:



Definition 4.1 Let $\pi : \mathfrak{X} \to \mathfrak{S}$ be a differentiable stack with atlas $x : X \to \mathfrak{X}$ and $U \subset X$ be an open set. We will say that the restricted substack $\mathfrak{U}$ is $\mathfrak{X}$-*categorical* if the inclusion map $i_{\mathfrak{U}} : \mathfrak{U} \to \mathfrak{X}$ is homotopic to a constant morphism $c : \mathfrak{U} \to \mathfrak{X}$ between differentiable stacks.

For instance, in Example 2.21 for the stack $\mathfrak{X} = [S^1/S^1]$ let U be the set of triples (P, S, μ), where S is a smooth manifold, P a S^1-torsor over S and $\mu : P \to S^1$ a S^1-equivariant smooth map. That is, $\mathfrak{U} = [S^1/S^1]$. We have that the stack $\mathfrak{U}$ is $\mathfrak{X}$-categorical since the identity map $\mathrm{id} : [S^1/S^1] \to [S^1/S^1]$ is homotopic to a constant morphism of stacks.

Definition 4.2 Let $\pi : \mathfrak{X} \to \mathfrak{S}$ be a differentiable stack with atlas $x : X \to \mathfrak{X}$. The *stacky Lusternik-Schnirelmann* or *stacky LS-category*, $\mathrm{cat}(\mathfrak{X})$, is the least number of open sets U needed to cover X such that the restricted substack $\mathcal{U}$ is $\mathfrak{X}$-categorical.

If X cannot be covered by a finite number of such open sets, we will say that $\mathrm{cat}(\mathfrak{X}) = \infty$.

Example 4.3 Let X be a smooth manifold. It follows immediately from the above definition that the stacky LS-category $\mathrm{cat}(\underline{X})$ is equal to the classical LS-category $\mathrm{cat}(X)$ of the manifold.

Example 4.4 From Example 2.21, we get immediately for the LS-category of the quotient stack $[S^1/S^1]$ that $\mathrm{cat}([S^1/S^1]) = 1$.

The following theorems establish the relationship between Lusternik-Schnirelmann category of differentiable stacks and Lie groupoids (see also [11]).

Theorem 4.5 *Let $\mathfrak{X}$ be a differentiable stack with a given presentation $x : X_0 \to \mathfrak{X}$ and associated Lie groupoid $\mathcal{G}(x) = X_0 \times_{\mathfrak{X}} X_0 \rightrightarrows X_0$. Then*

$$\mathrm{cat}(\mathfrak{X}) = \mathrm{cat}(\mathcal{G}(x)).$$

Proof This follows from the definition of LS-category for Lie groupoids. The fact that LS-category of Lie groupoids is Morita invariant, implies now by using Proposition 2.17 that it does not depend on the chosen presentation for the differentiable stack $\mathfrak{X}$, which gives the result. □

Theorem 4.6 *Let $\mathcal{G}$ be a Lie groupoid and $\mathfrak{B}\mathcal{G}$ be the associated differentiable stack. Then*

$$\mathrm{cat}(\mathfrak{B}\mathcal{G}) = \mathrm{cat}(\mathcal{G}).$$

Proof This follows from the explicit construction of the classifying stack $\mathfrak{B}\mathcal{G}$ of $\mathcal{G}$-torsors for the given Lie groupoid $\mathcal{G}$. Now the associated Lie groupoid of the differentiable stack $\mathfrak{B}\mathcal{G}$ is Morita equivalent to the given Lie groupoid $\mathcal{G}$ following Theorem 2.16 above. □

Example 4.7 Consider the action groupoid $\mathcal{G} = S^1 \times S^3 \rightrightarrows S^3$ defined by the action of the circle S^1 on the 3-sphere S^3 given by $(v, w) \mapsto (z^3 v, zw)$.

If we think of S^3 as the union of two solid tori, we have that the orbits of this action are circles of length 2π for points on the two cores C_1 and C_2 and circles of length $2\pi\sqrt{3\|v\|^2 + \|w\|^2}$ elsewhere. We construct a $\mathcal{G}$-categorical covering $\{\mathcal{U}_1, \mathcal{U}_2\}$ by subgroupoids of $\mathcal{G}$ given by considering the open sets $U_1 = S^3 - C_2$ and $U_2 = S^3 - C_1$. We have that each inclusion $i_{\mathcal{U}_i} : \mathcal{U}_i \to \mathcal{G}$ is groupoid homotopic to a generalized constant map with image in each respective core C_i. Moreover, this groupoid is not $\mathcal{G}$-contractible and we have that $\mathrm{cat}(\mathcal{G}) = 2$.

Now for the differentiable stack $[S^3/S^1]$ associated to this Lie groupoid, which in fact describes the teardrop orbifold, we get from the last theorem that $\mathrm{cat}([S^3/S^1]) = 2$.

More generally, given any quotient stack $[X/G]$ of a Lie group action on a smooth manifold, we see that the stacky LS-category of $[X/G]$ is equal to the groupoid LS-category of the action groupoid $G \times X \rightrightarrows X$, i.e. $\mathrm{cat}([X/G]) = \mathrm{cat}(G \times X \rightrightarrows X)$. This stacky LS-category of a quotient stack in fact generalizes also the notion of equivariant LS-category $\mathrm{cat}_G(X)$ for Lie group actions on manifolds as introduced before by Marzantowicz [8, 22]. We also aim to study the relationship between stacky LS-category and equivariant topology in forthcoming work.

Many of the particular examples and properties of LS-category for Lie groupoids as discussed in [9], especially concerning Lie group actions on smooth manifolds have interesting stacky analogs and will be discussed in detail in [11]. For example, orbifolds in general can naturally be seen as particular differentiable stacks associated to proper étale Lie groupoids and therefore give rise to a variety of interesting examples for LS-category of differentiable stacks and its relation with Morse theory, which we will also explore in detail in [11].

Acknowledgements The first and third author like to thank Sadok Kallel and the American University of Sharjah, UAE for financial support where part of this work was presented at the Second International Conference on Mathematics and Statistics. Both also like to thank the University of Leicester for additional support. The second author is supported in part by the Simons Foundation. Finally, the second and third authors also like to thank the Centro de Investigación en Matemáticas (CIMAT) in Guanajuato for the kind hospitality and support while this project was pursued.

References

1. Artin, M., Grothendieck, A., Verdier, J.-L.: Théorie des topos et cohomolgie étale des schémas, Séminaire de géometrie algébrique du Bois-Marie (SGA 4). Lecture Notes in Mathematics, vol. 269, 270, 305. Springer-Verlag, Berlin-New York, 1972–1973

2. Alsulami, S.: Homotopy types of topological groupoids and Lusternik-Schnirelmann category of topological stacks. Ph.D. thesis, University of Leicester, May 2016
3. Behrend, K.: On the de Rham cohomology of differential and algebraic stacks. Adv. Math. **198**, 583–622 (2005)
4. Behrend, K.: Cohomology of stacks, Intersection theory and moduli. ICTP Lect. Notes XIX, Abdus Salam Int. Cent. Theor. Physics, Trieste 249–294 (2004)
5. Behrend, K., Xu, P.: Differentiable stacks and gerbes. J. Symp. Geom. **9**, 285–341 (2011)
6. Biswas, I., Neumann, F.: Atiyah sequences, connections and characteristic forms for principal bundles over groupoids and stacks. Comp. Ren. Math. Acad. Sci. Paris **352**, 59–64 (2014)
7. Collier, B., Lerman, E., Wolbert, S.: Parallel transport on principal bundles over stacks, preprint. arXiv:math.DG/1509.05000
8. Colman, H.: Equivariant LS-category for finite group actions. Lusternik-Schnirelmann category and related topics (South Hadley, MA) Contemp. Math. **316**(2002), 35–40 (2001)
9. Colman, H.: The Lusternik-Schnirelmann category of a Lie groupoid. Trans. AMS **362**(10), 5529–5567 (2010)
10. Colman, H.: On the 1-homotopy type of Lie groupoids. Appl. Categor. Struct. **19**, 393–423 (2011)
11. Colman, H., Neumann, F.: Lusternik-Schnirelmann theory for stacks. In preparation
12. Cornea, O., Lupton, G., Oprea, J., Tanré, D.: Lusternik-Schnirelmann category, Mathematical Surveys and Monographs, 103. American Mathematical Society, Providence, RI (2003)
13. Fantechi, B.: Stacks for everybody. European Congress of Mathematics. Barcelona 2000, vol. **I**, pp. 349–359. Birkhäuser, Verlag (2001)
14. Felisatti, M., Neumann, F.: Secondary theories for étale groupoids. Regulators (Barcelona). Contemp. Math. **571**(2012), 135–151 (2010)
15. Grothendieck, A.: Revetements étale et groupe fundamental, Séminaire de géometrie algébrique du Bois-Marie 1960–1961 (SGA 1). Lecture Notes in Mathematics, vol. 224. Springer-Verlag, Berlin, New York (1971)
16. Heinloth, J.: Notes on differentiable stacks. In: Y. Tschinkel (ed.) Mathematisches Institut Seminars 2004-2005, pp. 1-32. Georg-August Universität Göttingen
17. Iglesias-Zemmour, P.: Diffeology, Mathematical Surveys and Monographs 185. American Mathematical Society, Providence, RI (2013)
18. James, I.M.: On category, in the sense of Lusternik-Schnirelmann. Topology **17**(4), 331–348 (1978)
19. James, I.M.: Lusternik-Schnirelmann category. In: Handbook of Algebraic Topology, pp. 1293–1310. Elsevier Science, Amsterdam (1995)
20. Laurent-Gengoux, C., Tu, J.-L., Xu, P.: Chern-Weil map for principal bundles over groupoids. Math. Zeit. **255**, 451–491 (2007)
21. Lusternik, L., Schnirelmann, L.: Méthodes topologiques dans les Problèmes Variationnels. Hermann, Paris (1934)
22. Marzantowicz, W.: A G-Lusternik-Schnirelman category of space with an action of a compact Lie group. Topology **28**(4), 403–412 (1989)
23. Moerdijk, I., Mrčun, J.: Introduction to foliations and Lie groupoids. Cambridge Studies in Advanced Mathematics, vol. 91. Cambridge University Press (2003)
24. Noohi, B.: Mapping stacks of topological stacks. J. Reine Angew. Math. **646**, 117–133 (2010)
25. Noohi, B.: Homotopy types of topological stacks. Adv. Math. **230**, 2014–2047 (2012)
26. Segal, G.: Classifying spaces and spectral sequences. Inst. Hautes Études Sci. Publ. Math. No. **34**, 105–112 (1968)
27. Souriau, J.M.: Groupes differentiels. Lecture Notes in Mathematics, vol. 836. Springer-Verlag, Berlin-New York (1980)
28. Tu, L., Xu, P., Laurent-Gengoux, C.: Twisted K-theory of differentiable stacks. Ann. Scient. Éc. Norm. Sup **37**, 841–910 (2004)

Centrosymmetric, and Symmetric and Hankel-Symmetric Matrices

Richard A. Brualdi and Shi-Mei Ma

Abstract We formulate and solve existence questions concerning centrosymmetric matrices and symmetric, Hankel-symmetric matrices which are nonnegative, nonnegative and integral, and (0, 1)-matrices.

Keywords Symmetric matrices · Centrosymmetric matrices · Hankel symmetric matrices · Palindromic vector · Hankel transpose · Interchange

Subject Classication: 05B20 · 15B05

1 Introduction

Let $A = [a_{ij}]$ be an $m \times n$ matrix, and let $A^\dagger = [a_{ij}^\dagger]$ denote the $m \times n$ matrix obtained from A by a rotation of 180°. Thus $a_{ij}^\dagger = a_{n+1-i,n+1-j}$ for all i and j. The matrix A is *centrosymmetric* provided $A^\dagger = A$, that is, provided

$$a_{n+1-i,n+1-j} = a_{ij} \text{ for all } i \text{ and } j.$$

If the $m \times n$ matrix A is centrosymmetric, then if m (resp., n) is odd, row $(m+1)/2$ (resp. column $(n+1)/2$) is *palindromic*, that is, is the same read forward or backward. The centrosymmetric matrix A is determined by its first $\lceil n/2 \rceil$ columns. Centrosymmetric matrices have occured in many investigations; see e.g. [1, 2, 6].

R.A. Brualdi (✉)
Department of Mathematics, University of Isconsin, Madison, WI 53706, USA
e-mail: brualdi@math.wisc.edu

S.-M. Ma
School of Mathematics and Statistics, Northeastern University at Qinhuangdao, Hebei 066004, People's Republic of China
e-mail: shimeimapapers@163.com

T. Abualrub et al. (eds.), *Mathematics Across Contemporary Sciences*,
Springer Proceedings in Mathematics & Statistics 190,
DOI 10.1007/978-3-319-46310-0_2

As usual $A^t = [a_{ij}^t]$ denotes the usual $n \times m$ *transpose* of A so that $a_{ij}^t = a_{ji}$ for all i and j. In addition, we consider the matrix $A^h = [a_{ij}^h]$ to denote the *Hankel transpose* [5] of A. This is the $n \times m$ matrix obtained from A by interchanging rows and columns as with the transpose, but using the reverse order in both cases. Thus $a_{ij}^h = a_{n+1-j,n+1-i}$ for all i and j. For example, if

$$A = \begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix},$$

then

$$A^h = \begin{bmatrix} f & c \\ e & b \\ d & a \end{bmatrix}.$$

If A is a square matrix, then A^h is obtained from A by transposing across the *Hankel diagonal*, that is, across the diagonal of A running from upper right to lower left.

A matrix A is *symmetric* provided that $A^t = A$. We say that a matrix is *Hankel-symmetric* provided that $A^h = A$. Symmetric and Hankel-symmetric matrices must be square matrices. An example of a Hankel-symmetric matrix is

$$\begin{bmatrix} 2 & 1 & 0 \\ 4 & 5 & 1 \\ 3 & 4 & 2 \end{bmatrix}.$$

If a square matrix is both symmetric and Hankel-symmetric, then it is centrosymmetric. This is because, consecutive reflections about two perpendicular lines (the main diagonal and the antidiagonal) is a rotation by 180°. More precisely, we have the following basic result.

Proposition 1 *Let A be an $n \times n$ matrix. Then any two of the following three properties implies the other:*

(s) *A is symmetric: $A^t = A$.*
(hs) *A is Hankel-symmetric: $A^h = A$.*
(cs) *A is centrosymmetric: $A^\dagger = A$.*

Proof (i) (s) and (hs) $\Rightarrow$ (cs): $a_{ij} \overset{(\mathrm{hs})}{=} a_{n+1-j,n+1-i} \overset{(\mathrm{s})}{=} a_{n+1-i,n+1-l}$.
(ii) (s) and (cs) $\Rightarrow$ (hs): $a_{ij} \overset{(cs)}{=} a_{n+1-i.n+1-j} \overset{(s)}{=} a_{n+1-j,n+1-i}$.
(iii) (hs) and (cs) $\Rightarrow$ (s): $a_{ij} \overset{(cs)}{=} a_{n+1-i,n+1-j} \overset{(hs)}{=} a_{ji}$.

□

Centrosymmetric permutation matrices are studied in [1, 2]. Centrosymmetric graphs are considered in [6].

A matrix, even a permutation matrix, may be centrosymmetric but not Hankel symmetric, or Hankel symmetric but not centrosymmetric. For example,

$$\begin{bmatrix} & & 1 & \\ & & & 1 \\ & 1 & & \\ 1 & & & \end{bmatrix}$$ (Hankel symmetric but not centrosymmetric),

and

$$\begin{bmatrix} & 1 & & \\ & & & 1 \\ 1 & & & \\ & & 1 & \end{bmatrix}$$ (centrosymmetric but not Hankel symmetric)

Let $R = (r_1, r_2, \ldots . r_m)$ and $S = (s_1, s_2, \ldots, s_n)$ be two nonnegative vectors with the sum of components:

$$\tau := r_1 + r_2 + \cdots + r_m = s_1 + s_2 + \cdots + s_n.$$

We denote by $\mathcal{T}(R, S)$ the set of all nonnegative real matrices with row sum vector R and column sum vector S. Matrices in $\mathcal{T}(R, S)$ are often called *transportation matrices* because of their connection to the well-known *transportation problem* of transporting goods from m sources with supplies of sizes given by R to n sources with demands given by S. $\mathcal{T}(R, S)$ is a convex polytope, and is nonempty since the $m \times n$ matrix $T = [t_{ij}]$ with

$$t_{ij} = \frac{r_i s_j}{\tau} \text{ for all } i \text{ and } j.$$

is in $\mathcal{T}(R, S)$. If R and S are also integral vectors, then $\mathcal{T}(R, S)$ is an *integral transportation polytope*. An integral transportation polytope always contains an integral matrix. In fact, the following *transportation matrix algorithm* always produces such a matrix $A = [a_{ij}]$ (see e.g. [4], pp. 26–27):

(1) Choose any i and j and set $a_{ij} = \min\{r_i, s_j\}$.

   (a) If $\min\{r_i, s_j\} = r_i$, set $a_{il} = 0$ for all $l \neq j$.
   (b) If $\min\{r_i, s_j\} = s_j$, set $a_{kj} = 0$ for all $k \neq i$.

(2) Reduce r_i and s_j by $\min\{r_i, s_j\}$, and proceed inductively.

In fact, the matrices produced by this algorithm carried out in all possible ways gives all the extreme points of the convex polytope $\mathcal{T}(R, S)$. Note that if R and S are integral vectors, then the algorithm always produces integral matrices. We denote the class of integral matrices in $\mathcal{T}(R, S)$ by $\mathcal{T}_Z(R, S)$. The class $\mathcal{T}_Z(R, S)$ may or may not contain a (0, 1)-matrix even if the components of R and S are small enough. For instance, if $R = (4, 3, 2, 1)$ and $S = (4, 4, 1, 1)$, then there does not exists a (0, 1)-matrix in $\mathcal{Z}(R, S)$.

Again with the assumption that R and S are nonnegative integral vectors, a much studied class of matrices is the class $\mathcal{A}(R, S)$ consisting of all $m \times n$ (0, 1)-matrices

in $\mathcal{T}_Z(R, S)$. The Gale-Ryser theorem (see e.g. [4], p. 27) characterizes the nonemptiness of the class $\mathcal{A}(R, S)$ as follows.

Let $R = (r_1, r_2, \ldots, r_m)$ and $S = (s_1, s_2, \ldots, s_n)$ be nonnegative integral vectors with

$$r_1 + r_2 + \cdots + r_m = s_1 + s_2 + \cdots + s_n.$$

Assume without loss of generality (by permuting rows and columns if necessary) that

$$r_1 \geq r_2 \geq \cdots \geq r_m \text{ and } s_1 \geq s_2 \geq \cdots \geq s_n.$$

Let $R^* = (r_1^*, r_2^*, \ldots, r_n^*)$ be the *conjugate* of R, that is,

$$r_j^* = |\{i : r_i \geq j\}.$$

Then $\mathcal{A}(R, S) \neq \emptyset$ if and only if S is *majorized* by R, that is,

$$s_1 + s_2 + \cdots + s_j \leq r_1^* + r_2^* + \cdots + r_j^* \quad (j = 1, 2, \ldots, n), \text{ with equality for } j = n. \tag{1}$$

Assuming that $R = S = (r_1, r_2, \ldots, r_n)$, one can also consider the existence of symmetric matrices in the classes $\mathcal{T}(R, R)$ and, if R is integral, $\mathcal{T}_Z(R, R)$ and $\mathcal{A}(R, R)$. The classes $\mathcal{T}(R, R)$ and $\mathcal{T}_Z(R, R)$ always contain a symmetric matrix since they contain the diagonal matrix with $r_1, r_2, \ldots, r_n$ on the main diagonal. It is a consequence of a theorem of Fulkerson, Hoffman, and McAndrew that the class $\mathcal{A}(R)$ contains a symmetric matrix if and only if it is nonempty (see e.g. [3], pp. 179–182).

In this note, we consider certain subsets of the above matrix classes defined by imposing the structural conditions of centrosymmetry, and symmetry and Hankel-symmetry, and we obtain analogous results to those described above.

2 Existence Theorems

In this section we obtain nonemptiness criteria for the classes introduced in the previous section. Since the property of centrosymmetry does not require that the matrix be square, we need not assume our matrices are square in this case; but symmetry and Hankel-symmetry do require that the matrix be square. We shall adapt three well-known algorithms to the centrosymmetric, and symmetric and Hankel-symmetric cases.

Let $R = (r_1, r_2, \ldots, r_m)$ and $S = (s_1, s_2, \ldots, s_n)$ be nonnegative vectors. Let $\mathfrak{C}(R, S)$ denote the class of all centrosymmetric, nonnegative matrices with row sum vector R and column sum vector S. If R and S are also integral, let $\mathfrak{C}_Z(R, S)$ denote the class of all centrosymmetric, nonnegative integral matrices with row sum vector

R and column sum vector S. Recall that a vector $(d_1, d_2, \ldots, d_k)$ is *palindromic* provided that $d_i = d_{k+1-i}$ for all i.

Theorem 2 *The class $\mathfrak{C}(R, S)$ is nonempty if and only if*

$$\sum_{i=1}^{m} r_i = \sum_{j=1}^{n} s_j. \tag{2}$$

and

$$R \text{ and } S \text{ are palindromic.} \tag{3}$$

If R and S are integral vectors, the class $\mathfrak{C}_Z(R, S)$ is nonempty if and only if (2) and (3) hold.

Proof The conditions (2) and (3) are certainly necessary in order that $\mathfrak{C}(R, S) \neq \emptyset$ and, if R and S are integral, in order that $\mathfrak{C}_Z(R, S) \neq \emptyset$. Now assume that (2) and (3) hold. We modify the transportation matrix algorithm to show the two classes are nonempty.

Since R and S are palindromic, it follows from (2) that if m is even and n is odd, then $s_{(n+1)/2}$ is even. Similarly. if m is odd and n is even, then $r_{(m+1)/2}$ is even. If m and n are both odd, then $r_{(m+1)/2}$ and $s_{(n+1)/2}$ have the same parity. We proceed as in the transportation algorithm but with the modifications as given below. If R and S are integral, the result will be an integral matrix.

If m and n are both odd and e.g. $r_{(m+1)/2} \geq s_{(n+1)/2}$, then we put $a_{(m+1)/2,(n+1)/2} = s_{(n+1)/2}$, and put the remaining entries in column $(n+1)/2$ equal to zero and adjust $r_{(m+1)/2}$ to $r_{(m+1)/2} - s_{(n+1)/2}$. Then we are left to construct a centrosymmetric $m \times (n-1)$ matrix where m is odd and $n-1$ is even with palindromic row and column sum vectors the sum of whose entries are equal.

If m is even and n is odd, then there are two possibilities (first we use row 1 although any row $i \leq m/2$ can be used): If $2r_1 \leq s_{(n+1)/2}$, then we put $a_{1,(n+1)/2} = a_{m,(n+1)/2} = r_1$ and set the remaining entries in rows 1 and m equal to zero. Then, adjusting the sum of column $(n+1)/2$, we are left to construct a centrosymmetric $(m-2) \times n$ matrix with palindromic row and column sum vectors the sum of whose entries are equal. If $2r_1 > s_{(n+1)/2}$, then we set $a_{1,(n+1)/2} = a_{m,(n+1)/2} = \frac{s_{(n+1)/2}}{2}$, and set the remaining entries in column $(n+1)/2$ equal to zero. If needed, we next consider row 2 and continue until column $(n+1)/2$ is specified. We are then left to construct a centrosymmetric $m \times (n-1)$ matrix with palindromic row and column sum vectors the sum of whose entries are equal.

If m is odd and n is even, we proceed in a similar way.

Finally, if m and n are both even, then we proceed as in the transportation matrix algorithm with additionally setting $a_{n+1-i,n+1-j} = a_{ij}$, and adjusting two row or two column sums as needed. □

Example 3 Let $R = (2, 4, 5, 4, 2)$ and $S = (5, 2, 3, 2, 5)$. Then one way to carry out the above procedure to obtain a matrix in $\mathfrak{C}(R, S)$ is the following:

$$\left[\begin{array}{|c|c|c|c|c|}\hline & & 0 & & \\ \hline & & 0 & & \\ \hline & & 3 & & \\ \hline & & 0 & & \\ \hline & & 0 & & \\ \hline\end{array}\right] \rightarrow \left[\begin{array}{|c|c|c|c|c|}\hline & & 0 & & \\ \hline & & 0 & & \\ \hline 1 & 0 & 3 & 0 & 1 \\ \hline & & 0 & & \\ \hline & & 0 & & \\ \hline\end{array}\right] \rightarrow \left[\begin{array}{|c|c|c|c|c|}\hline 2 & 0 & 0 & 0 & 0 \\ \hline & & 0 & & \\ \hline 1 & 0 & 3 & 0 & 1 \\ \hline & & 0 & & \\ \hline 2 & 0 & 0 & 0 & 2 \\ \hline\end{array}\right] \rightarrow$$

$$\left[\begin{array}{|c|c|c|c|c|}\hline 2 & 0 & 0 & 0 & 2 \\ \hline 2 & & 0 & & 0 \\ \hline 1 & 0 & 3 & 0 & 1 \\ \hline 0 & & 0 & & 2 \\ \hline 0 & 0 & 0 & 0 & 2 \\ \hline\end{array}\right] \rightarrow \left[\begin{array}{|c|c|c|c|c|}\hline 2 & 0 & 0 & 0 & 0 \\ \hline 2 & 2 & 0 & 0 & 0 \\ \hline 1 & 0 & 3 & 0 & 1 \\ \hline 0 & 0 & 0 & 2 & 2 \\ \hline 0 & 0 & 0 & 0 & 2 \\ \hline\end{array}\right].$$

□

We now consider the possibility of the existence of an $m \times n$ $(0, 1)$-matrix $A = [a_{ij}]$ in $\mathfrak{C}(R, S)$. Let $\mathfrak{C}_{(0,1)}(R, S)$ denote the set of $(0, 1)$-matrices in $\mathfrak{C}(R, S)$. Then

$$\mathfrak{C}_{(0,1)}(R, S) = \mathfrak{C}(R, S) \cap \mathcal{A}(R, S).$$

Hence, necessary conditions for the nonemptiness of $\mathfrak{C}_{(0,1)}(R, S)$ are that both $\mathfrak{C}(R, S)$ and $\mathcal{A}(R, S)$ are nonempty. For $\mathfrak{C}(R, S) \neq \emptyset$, conditions (2) and (3) must hold and hence R and S must be palindromic with the same sum of entries. Thus if $A = [a_{ij}] \in \mathfrak{C}_{(0,1)}(R, S)$ and m and n are both odd, then $r_{(m+1)/2}$ and $s_{(n+1)/2}$ have the same parity, and $a_{(m+1)/2,(n+1)/2} = 1$ if this parity is odd and $a_{(m+1)/2,(n+1)/2} = 0$ if this parity is even.

Before showing $\mathfrak{C}(R, S) \neq \emptyset$ and $\mathcal{A}(R, S) \neq \emptyset$ are also sufficient for the class $\mathfrak{C}_{(0,1)}(R, S)$ to be nonempty, we consider a property of this class analogous to a basic property of $\mathcal{A}(R, S)$. This property of $\mathcal{A}(R, S)$ is the following.

Let $A = [a_{ij}] \in \mathcal{A}(R, S)$. Let i, j, k, l be indices where $i < j$ and $k < l$ such that the 2×2 submatrix of A determined by rows i and j and columns k and l is

$$A[i, j|k, l] = \begin{array}{c|cc} & k & l \\ \hline i & 1 & 0 \\ j & 0 & 1 \end{array}. \tag{4}$$

Replacing this 2×2 submatrix equal to I_2 with

$$L_2 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

gives another matrix in $\mathfrak{C}_{(0,1)}(R, S)$. Similarly, replacing an L_2 with an I_2 in a matrix in $\mathcal{A}(R, S)$ always gives another matrix in $\mathcal{A}(R, S)$. Either of these replacements is called an *interchange*. It is a basic fact (see e.g. [4], pp. 52–57) that any matrix in $\mathcal{A}(R, S)$ can transformed into any other by a sequence of interchanges.

Now suppose that A is also centrosymmetric, Then if (4) holds in A, we also have

$$A[n+1-j, n+1-i | n+1-l, n+1-k] = \begin{array}{c|cc} & n+1-l & n+1-k \\ \hline n+1-j & 1 & 0 \\ n+1-i & 0 & 1 \end{array}.$$

Replacing each of these 2×2 submatrices equal to I_2 with

$$L_2 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

gives another matrix in $\mathfrak{C}_{(0,1)}(R, S)$. We call this pair of substitutions and the one going in the reverse direction (so two submatrices equal to L_2 are replaced with matrices equal to I_2) a *centrosymmetric double-interchange*. Note that if $i + j = n + 1$ and $k + l = n + 1$, only one 2×2 submatrix is involved in a centrosymmetric double-interchange.

Lemma 4 *Assume that at least one of m and n is odd and* $A = [a_{ij}] \in \mathfrak{C}_{(0,1)}(R, S) \neq \emptyset$. *Then by a sequence of centrosymmetric double-interchanges we can obtain a matrix* $C = [c_{ij}] \in \mathfrak{C}_{(0,1)}(R, S)$ *such that, except for* $c_{(m+1)/2,(n+1)/2}$ *in the case that both m and n are odd, the* 1*s in row* $(m + 1)/2$ *if m is odd occur in the positions corresponding to the largest column sums, and the* 1*s in column* $(n + 1)/2$ *if n is odd occur in the positions corresponding to the largest row sums.*

Proof This lemma follows easily using centrosymmetric double-interchanges. First, if m and n are both odd, then $a_{(m+1)/2,(n+1)/2}$ equals 1 if $r_{(m+1)/2}$ and $s_{(n+1)/2}$ have odd parity, and equals 0 if they have even parity. If there are two columns k and l with $1 \leq k, l \leq (n + 1)/2$ such that $s_k < s_l$ but $a_{(m+1)/2,k} = 1$ and $a_{(m+1)/2,l} = 0$, then for some i we must have $a_{ik} = 0$ and $a_{il} = 1$. Then there is a centrosymmetric double-interchange that replaces $a_{(m+1)/2,k}$ with 0 and $a_{(m+1)/2,l}$ with 1. A similar argument works for rows. It follows that by centrosymmetric double-interchanges we can arrive at C with the desired properties. □

Lemma 5 *Let A and B be any two matrices in* $\mathfrak{C}_{(0,1)}(R, S)$. *Then there is a sequence of centrosymmetric double-interchanges which transforms A into B with all intermediate matrices in* $\mathfrak{C}_{(0,1)}(R, S)$.

Proof We may assume that both $A = [a_{ij}]$ and $B = [b_{ij}]$ have the properties of C specified in Lemma 4. By deleting row $(m + 1)/2$ if m is odd and column $(n + 1)/2$ if n is odd, we may assume that both m and n are even. By permutations of rows and of columns in a way that preserves centrosymmetry, we may also assume that $r_1 \geq r_2 \geq \cdots \geq r_{m/2}$ and that $s_1 \geq s_2 \geq \cdots \geq s_{n/2}$. Consider columns 1 and n of A, and suppose they differ from columns 1 and n of B, respectively. Then there exists k and l with $k \neq l$ such that $a_{kn} = 1$ and $a_{ln} = 0$, and $b_{kn} = 0$ and $b_{ln} = 1$ (or the other way around). Suppose there did not exist a p such that either $a_{kp} = 0$ and $a_{lp} = 1$, or $b_{kp} = 1$ and $b_{lp} = 0$. Since A and B have the same row sum vector R, from A we see that $r_k > r_l$, and from B we see that $r_k < r_l$, a contradiction. Without loss of generality, assume that $a_{kp} = 0$ and $a_{lp} = 1$ for some p. Then a centrosymmetric

double-interchange applied to A results in a matrix C whose column n (and column 1) has more in common with the corresponding columns of B. Replacing A with C and proceeding recursively, we see that there is a sequence of centrosymmetric double-interchanges applied to A and a sequence of centrosymmetric double-interchanges applied to B such that the resulting matrices A' and B' are in $\mathfrak{C}_{(0,1)}(R, S)$ and their corresponding columns n and corresponding columns 1 agree. Proceeding recursively, we see that there is a sequence of centrosymmetric double-interchanges applied to A and a sequence of centrosymmetric double-interchanges applied to B which result in the same matrix. Since centrosymmetric double-interchanges are reversible, this completes the proof. □

Since by Lemma 4, if at least one of m and n is odd, we can reduce the nonemptiness of $\mathfrak{C}_{(0,1)}(R, S)$ to the case where both m and n are even, we now assume that both m and n are even. By Theorem 2, $\mathfrak{C}(R, S) \neq \emptyset$ if and only if (2) and (3) are satisfied. Necessary and sufficient conditions for $\mathcal{A}(R, S) \neq \emptyset$ are given by the Gale-Ryser theorem (see e.g. [4]). Assuming without loss of generality that the monotonicity conditions $r_1 \geq r_2 \geq \cdots \geq r_{m/2}$ and $s_1 \geq s_2 \geq \cdots \geq s_{n/2}$ hold, the Gale-Ryser conditions applied to the monotone rearrangements of R and S, that is, to

$$(r_1, r_1, r_2, r_2, \ldots, r_{m/2}, r_{m/2}) \text{ and } (s_1, s_1, s_2, s_2, \ldots, s_{n/2}, s_{n/2})$$

reduce to

$$\sum_{j=1}^{k} s_j \leq \sum_{j=1}^{k} r_j^* \quad (j = 1, 2, \ldots, n/2), \tag{5}$$

with equality for $k = n/2$. Here $\tilde{R}^* = (r_1^*, r_2^*, \ldots, r_{m/2}^*)$ is the conjugate of $\tilde{R} = (r_1, r_2, \ldots, r_{m/2})$.

We then have the following theorem.

Theorem 6 *Let m and n be even. The class $\mathfrak{C}_{(0,1)}(R, S)$ is nonempty if and only if both of the classes $\mathfrak{C}(R, S)$ and $\mathcal{A}(R, S)$ are nonempty, thus if and only if the conditions (2), (3), and (5) hold.*

Proof If $\mathfrak{C}_{(0,1)}(R, S)$ is nonempty, then clearly $\mathfrak{C}(R, S)$ and $\mathcal{A}(R, S)$ are nonempty, and (2), (3), and (1) hold.

Now assume that (2), (3), and (5) hold. We prove the existence (along with an algorithm for construction) of a matrix in $\mathfrak{C}_{(0,1)}(R, S)$ by modifying the Gale-Ryser algorithm to construct a matrix in $\mathcal{A}(R, S)$ (again see [4]), and thus we do not make explicit use of (1). In the Gale-Ryser algorithm, the row and column sums are assumed to be monotone, but this is just for ease of description to establish that the algorithm produces a matrix in $\mathcal{A}(R, S)$ or that at least one of the Gale-Ryser conditions fails. The Gale-Ryser algorithm is recursive and proceeds as follows.

(i) Choose a column with the smallest prescribed column sum.
(ii) Put the prescribed number of 1s in that column in those rows with the largest row sum (there may be ties in which case the row can be chosen arbitrarily among those rows with the same sum), and put 0s in all other positions of column n.
(iii) Adjust the prescribed row sums and proceed recursively with a column with the next smallest sum.

Since R and S are palindromic, we have that

$$R = (r_1, r_2, \ldots, r_2, r_1) \text{ and } S = (s_1, s_2, \ldots, s_2, s_1).$$

By reordering the first half of the entries of R (resp., S) with the corresponding reordering of the second half of the entries, as above we assume that

$$r_1 \geq r_2 \geq \cdots \geq r_{\lfloor n/2 \rfloor} \text{ and } s_1 \leq s_2 \leq \cdots \leq s_{\lfloor n/2 \rfloor}. \tag{6}$$

We now adjust the Gale-Ryser algorithm in the case that R and S are palindromic, staying within the constraints of the algorithm, in order to produce a centrosymmetric matrix.

Let $c_0, c_1, c_2, \ldots, c_k$ where $c_0 = 0$ and $c_k = n/2$ be defined by

$$r_1 = \cdots = r_{c_1} > r_{c_1+1} = \cdots = r_{c_1+c_2} > \cdots > r_{c_1+\cdots+c_{k-1}+1} = \cdots = r_{n/2}.$$

Let s_1 satisfy $2(c_0 + c_1 + \cdots + c_p) \leq s_1 < 2(c_0 + c_1 + \cdots + c_p + c_{p+1})$. In column 1 we put 1 s in rows $\{1, 2, \ldots, c_1 + \cdots + c_p\}$ and in rows $\{n + 1 - 1, n + 1 - 2, \ldots, n + 1 - (c_1 + \cdots + c_p)\}$. We also put 1 s in rows $c_1 + \cdots + c_p + 1, \ldots, c_1 + \cdots + c_{p+1}, n + 1 - c_1, \cdots, n + 1 - c_{p+1}$ in the order listed until a total of s_1 1s have been placed in column 1. This is in agreement with a possible way to carry out the Gale-Ryser algorithm to produce a matrix in $\mathcal{A}(R, S)$. Adjusting the needed row sums as a result of these 1s in column 1, we see that if we rotate this column 1 by 180° and take it as column n, then this is also a next possible step in the Gale-Ryser algorithm. Adjusting the needed row sums again, we see that they form a palindromic sequence. We now delete columns 1 and n, and proceed recursively. In order to keep the assumed monotonicty conditions on row and column sums, we may have to reorder the rows keeping as we do so, the palindromic property. We conclude that this way of carrying out the Gale-Ryser algorithm produces a centrosymmetric (0, 1)-matrix with row sum vector R and column sum vector S. □

Example 7 Let $R = (4, 4, 3, 3, 3, 3, 4, 4)$ and $S = (5, 3, 3, 3, 3, 3, 3, 5)$. Then the algorithm in the proof of Theorem 6 produces the following matrix, with the resulting row sum vectors after each pair of steps, including the initial row sum vector, indicated on the right.

1	1		1				1	4	2	1	1	0
1		1	1				1	4	2	2	1	0
1		1			1			3	2	2	0	0
	1		1			1		3	3	1	1	0
	1			1		1		3	3	1	1	0
		1			1		1	3	2	2	0	0
1				1	1		1	4	2	2	1	0
1				1		1	1	4	2	1	1	0

□

We know that a centrosymmetric matrix need not be symmetric or Hankel-symmetric. So Theorem 6 does not directly address the question of the existence of a symmetric and Hankel-symmetric matrix with a specified row sum vector (which by symmetry equals its column sum vector). The existence of a symmetric (or Hankel-symmetric) nonnegative integral matrix with a prescribed row sum vector follows from the theorem of Erdös and Gallai (the (0, 1)-case with all zeros on the main diagonal) and its generalizations (see again [4]). We show using a technique of Fulkerson, Hoffman, and McAndrew (see e.g. [3], pp. 179–182) that the existence of a symmetric and Hankel symmetric (0, 1) matrix (so centrosymmetric) can be gotten from Theorem 6. The row and column sum vector of such a matrix are equal to the same palindromic vector.

We now make use of the digraph associated with a (0, 1)-matrix. Let Γ be a digraph with vertices $\{1, 2, \ldots, n\}$. The *outdegree sequence* (resp., *indegree sequence*) of Γ records the number of edges leaving (resp., entering) each of its vertices. The adjacency matrix of Γ is the $n \times n$ (0, 1)-matrix $A = [a_{ij}]$ where $a_{ij} = 1$ if and only if there is an edge from vertex i to vertex j. We write $\Gamma = D(A)$ to indicate that the digraph associated with A equals Γ. A digraph is a *centrosymmetric digraph* provided after possible reordering of its vertices its adjacency matrix is centrosymmetric.

Recall that $\mathcal{A}(R, R)$ denotes the class of all (0, 1)-matrices with both row and column sum vectors equal to R.

Theorem 8 *Let $R = (r_1, r_2, \ldots, r_n)$ be a vector of nonnegative integers.*

(i) If the class $\mathcal{A}(R, R)$ contains a centrosymmetric matrix, then it contains a Hankel-symmetric, symmetric matrix.

(ii) Necessary and sufficient conditions that $\mathcal{A}(R, R)$ contains a Hankel-symmetric, symmetric matrix are that R is palindromic and the Gale-Ryser conditions (1) are satisfied.

Proof The assertion (ii) follows from Theorem 6 and assertion (i).

To prove assertion (i), let $A = [a_{ij}]$ be a centrosymmetric matrix in $\mathcal{A}(R, R)$ with digraph $D(A)$. The indegree and outdegree sequences of $D(A)$ are both equal to R. If A is symmetric, then by Proposition 1, A is also Hankel-symmetric. Now assume that A is not symmetric. Let $A^* = [a^*_{ij}]$ be the matrix obtained from A by replacing all pairs of symmetrically opposite 1s (including 1 s on the main diagonal) with

zeros. The matrix A^* is also centrosymmetric; moreover, for each i, $a^*_{i,n+1-i} = 0$, for otherwise we would have that $a^*_{i,n+1-i} = 1$ and, by centrosymmetry, $a^*_{n+1-i,i} = 1$, a contradiction. The digraph $D(A^*)$ has vertex set $\{1, 2, \ldots, n\}$ with an edge $i \to j$ from vertex i to vertex j if and only if $a_{ij} = 1$ but $a_{ji} = 0$. Since A is centrosymmetric, $i \to j$ is an edge of $D(A^*)$ if and only if $n + 1 - i \to n + 1 - j$ is also an edge. Since R is both the indegree and outdegree sequence of $D(A)$, the indegree of each vertex of $D(A^*)$ also equals its outdegree. Since A is not symmetric, $D(A^*)$ has at least one edge. These facts imply that $D(A^*)$ has a simple cycle of distinct vertices

$$\gamma : i_1 \to i_2 \to \cdots \to i_k \to i_1$$

for some $k \geq 3$. Since A is palindromic,

$$\gamma^\dagger : n + 1 - i_1 \to n + 1 - i_2 \to \cdots \to n + 1 - i_k \to n + 1 - i_1$$

is also a cycle of $D(A^*)$, which we call the *palindromic mate* of γ. No edge of γ can also be an edge of $\gamma^\dagger$ for that would imply that $a^*_{i,n+1-i} = a^*_{n+1-i,i} = 1$ for some i. a contradiction. It follows that the edges of $D(A^*)$ can be partitioned into cycles and these cycles come in palindromic pairs of cycles without common edges, or are self-palindromic cycles (that is, they are cycles equal to their palindromic mates).

Consider a palindromic pair γ and $\gamma^\dagger$ of these cycles. If the length of γ and $\gamma^\dagger$ is even, we delete the first, third, fifth, ... edges of these cycles and add the reverse of the second, fourth, sixth, ... edges. If the length of γ and $\gamma^\dagger$ is odd, then there are two possibilities. The first possibility is that some vertex a of γ (and the corresponding vertex $n + 1 - a$ of $\gamma^\dagger$) does not contain a loop in $D(A)$ ($D(A^*)$ does not contain any loops since A^* has only zeros on its main diagonal). We then put a loop at vertices a and $n + 1 - a$ and delete every other edge of γ starting with an edge meeting vertex a, and similarly delete every other edge of $\gamma^\dagger$ starting with the corresponding edge at vertex $n + 1 - a$; and we also insert the reverse of the remaining edges of γ and $\gamma^\dagger$. If every vertex of γ has a loop at it in $D(A)$ (and then so does every vertex of $\gamma^\dagger$), we remove the loop at some vertex a of γ (and remove the loop at the corresponding vertex $n + 1 - a$ of $\gamma^\dagger$), insert the reverse of every other edge of γ starting with an edge at vertex a, and similarly insert the reverse of every other edge of $\gamma^\dagger$ starting with the corresponding edge at vertex $n + 1 - a$, and delete all other edges of γ and $\gamma^\dagger$. If $\gamma = \gamma^\dagger$, then we follow a similar procedure as above but we have only to consider γ. In the case of n even, a self-palindromic cycle has even length and so we can follow the procedure above for pairs of cycles of even length. In case of n odd, a self-palindromic cycle may have even or odd length; if even length, we follow the above procedure; if odd length, then $(n + 1)/2$ must be a vertex of the cycle, and we follow the above procedure taking $(n + 1)/2$ as the first vertex of the cycle. The resulting digraph has a centrosymmetric adjacency matrix with fewer nonsymmetric arcs.

Repeating for each pair of palindromic cycles in the partition of the edges of $D(A^*)$, we obtain a digraph whose adjacency matrix B is symmetric and centrosymmetric. Putting the symmetric 1 s into B that were deleted from A to get A^*, we

obtain a symmetric, centrosymmetric $(0, 1)$-matrix with row sum vector R, and hence a Hankel-symmetric, symmetric matrix in $\mathcal{A}(R, R)$. □

Example 9 Consider the 11×11 centrosymmetric $(0, 1)$-matrix with palindromic $R = S = (1, 3, 3, 2, 1, 2, 1, 2, 3, 3, 1)$:

$$A = \begin{bmatrix}
 & & & & & 1 & & & & & \\
 & 1 & 1 & & & & & & 1 & & \\
 & 1 & 1 & 1 & & & & & & & \\
 & & & 1 & & & & & & 1 & \\
 & & & & & & & & & & 1 \\
 & & & & 1 & & 1 & & & & \\
1 & & & & & & & & & & \\
 & 1 & & & & & & 1 & & & \\
 & & & & & & & 1 & 1 & 1 & \\
 & & 1 & & & & & & 1 & 1 & \\
 & & & & & 1 & & & & &
\end{bmatrix}.$$

The matrix A^* is centrosymmetric and is given by

$$A^* = \begin{bmatrix}
 & & & & & 1 & & & & & \\
 & & & & & & & & 1 & & \\
 & & & & 1 & & & & & & \\
 & & & & & & & & & 1 & \\
 & & & & & & & & & & 1 \\
 & & & & 1 & & 1 & & & & \\
1 & & & & & & & & & & \\
 & 1 & & & & & & & & & \\
 & & & & & & & 1 & & & \\
 & & 1 & & & & & & & & \\
 & & & & & 1 & & & & &
\end{bmatrix}.$$

The decomposition of $D(A^*)$ into palindromic pairs of cycles is

$$1 \to 6 \to 7 \to 1,\ 11 \to 6 \to 5 \to 11 \text{ and } 2 \to 9 \to 8 \to 2,\ 10 \to 3 \to 4 \to 10,$$

and this corresponds to the following the decomposition of A^*:

$$A^* = \left[\begin{array}{cccccccccc}
 & & & & 1 & & & & & \\
 & & & & & & & & & \\
 & & & & & & & & & \\
 & & & & & & & & & \\
 & & & & & & & & & 1 \\
 & & & 1 & & 1 & & & & \\
1 & & & & & & & & & \\
 & & & & & & & & & \\
 & & & & & & & & & \\
 & & & & & & & & & \\
 & & & & 1 & & & & &
\end{array}\right] + \left[\begin{array}{cccccccccc}
 & & & & & & & & & \\
 & & & & & & & & 1 & \\
 & & & 1 & & & & & & \\
 & & & & & & & & & 1 \\
 & & & & & & & & & \\
 & & & & & & & & & \\
 & 1 & & & & & & & & \\
 & & & & & & & & & \\
 & & & & & & & 1 & & \\
 & & 1 & & & & & & & \\
 & & & & & & & & &
\end{array}\right].$$

Replacing A^* with

$$\left[\begin{array}{cccccccccc}
1 & & & & & & & & & \\
 & & & & & & & & & \\
 & & & & & & & & & \\
 & & & & & & & & & \\
 & & & & 1 & & & & & \\
 & & & 1 & & 1 & & & & \\
 & & & & 1 & & & & & \\
 & & & & & & & & & \\
 & & & & & & & & & \\
 & & & & & & & & & \\
 & & & & & & & & & 1
\end{array}\right] + \left[\begin{array}{cccccccccc}
 & & & & & & & & & \\
 & -1 & & & & & 1 & 1 & & \\
 & & & & & & & & 1 & \\
 & & & 1 & & & & & & \\
 & & & & & & & & & \\
 & & & & & & & & & \\
 & & & & & & 1 & & & \\
 & 1 & & & & & & & & \\
 & & 1 & 1 & & & & & -1 & \\
 & & & & & & & & &
\end{array}\right],$$

(the -1 corresponds to a loop that is deleted but is not part of the cycle), we then get the centrosymmetric, symmetric (0, 1)-matrix with row and column sum vector (1, 2, 2, 2, 1, 2, 1, 2, 2, 2, 1):

$$A = \left[\begin{array}{cccccccccc}
1 & & & & & & & & & \\
 & & 1 & & & & & 1 & & \\
 & 1 & & & & & & & 1 & \\
 & & & 1 & & & & & & \\
 & & & & & 1 & & & & \\
 & & & & 1 & & 1 & & & \\
 & & & & & 1 & & & & \\
 & & & & & & & 1 & & \\
 & 1 & & & & & & 1 & 1 & \\
 & & 1 & & & & & 1 & 1 & \\
 & & & & & & & & & 1
\end{array}\right].$$

□

Let $\mathcal{H}(R)$ denote the class of all Hankel-symmetric, symmetric (0, 1)-matrices with row and column sum vectors equal to $R = (r_1, r_2, \ldots, r_2, r_1)$. Theorem 8 contains necessary and sufficient conditions on R in order that $\mathcal{H}(R) \neq \emptyset$. We now show how to generate all matrices in a nonempty class $\mathcal{H}(R)$ from any one matrix in $\mathcal{H}(R)$. This is the analogue of Lemma 5 for $\mathcal{H}(R)$. According to Corollary 7.2.4 in [4], any two symmetric (0, 1)-matrices with row and column sum vector R can be obtained from one another by a sequence of *symmetric interchanges*. These symmetric interchanges are of three types and transform a 2×2, 3×3, and 4×4 principal submatrix into another as shown below:

(i) $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \leftrightarrow \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$.

(ii) $\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \leftrightarrow \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$.

(iii) $\begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \leftrightarrow \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$.

In terms of the graph whose adjacency matrix is a symmetric (0, 1)-matrix, these transformations are:

(i) Interchange (a) a configuration consisting of two distinct vertices u and v with a loop at u and a loop at v with (b) the configuration consisting of an edge $\{u, v\}$ joining u and v.

(ii) Interchange (a) a configuration consisting of an edge joining distinct vertices u and v and a loop at a third vertex w with (b) the configuration consisting of an edge joining u and w and a loop at v.

(iii) Interchange (a) a configuration consisting of four distinct vertices u, v, w, z and an edge joining u and v and one joining w and z with (b) the configuration consisting of an edge joining u and w and one joining v and z.

Every symmetric interchange has a corresponding complementary interchange with each index i of the corresponding principal submatrix (so a vertex of the associated graph) replaced by the complementary index $n + 1 - i$ (the complementary vertex of the associated graph).

We can use these symmetric interchanges to transform each matrix in $\mathcal{H}(R)$ to any other matrix in $\mathcal{H}(R)$. First we note that if n is odd, then all matrices in $\mathcal{H}(R)$ agree in position $(n + 1)/2, (n + 1)/2)$ and hence we never have to consider loops at vertex $(n + 1)/2$. Whenever a symmetric interchange is required, we also perform the corresponding complementary symmetric interchange, unless the symmetric interchange is self-complementary in which case only one symmetric interchange is performed. We call the simultaneous application of both of these a *symmetric double-interchange*. This then gives the following result.

Theorem 10 *Let A and B be any two matrices in $\mathcal{H}(R)$. Then there is a sequence of symmetric double-interchanges which transforms A into B with all intermediate matrices in $\mathcal{H}(R)$.*

3 Concluding Remarks

We have investigated two natural subclasses, namely the centrosymmetric subclass and the symmetric, Hankel-symmetric subclass, of several well-known classes of matrices. We have presented existence theorems, construction algorithms, and a means to generate all matrices in a subclass starting from any one matrix contained in it. Combinatorial questions concerning the number (or a good and simple upper bounds) of matrices in these subclasses are of interest but presumably very difficult. Formulas for the number of matrices in nonempty classes $\mathcal{A}(R, S)$ and $\mathcal{T}_Z(R, S)$ are available in terms of the Kostka numbers for the number of Young tableaux of given shape and size (see [4], p. 147). It may be possible to express the number of centrosymmetric, or symmetric and Hankel-symmetric, matrices in these classes using Kostka numbers. There is also a (basically symbolic) generating polynomial in $m + n$ variables for the number of matrices in a class $\mathcal{A}(R, S)$. It would be of interest to have useful generating polynomials for any of the classes and subclasses considered in this paper.

References

1. Barnabei, M., Bonetti, F., Silimbani, M.: The Eulerian distribution on centrosymmetric involutions. Discrete Math. Theor. Comput. Sci. **11**, 95–115 (2009)
2. Barnabei, M., Bonetti, F., Silimbani, M.: The Eulerian numbers on restricted centrosymmetric permutations. Pure Math. Appl. **21**, 99–118 (2010)
3. Brualdi, R.A., Ryser, H.J.: Combinatorial Matrix Theory, Encyclopedia of Math, and its Applics, vol. 39. Cambridge Univ. Press, Cambridge (1991)
4. Brualdi, R.A.: Combinatorial Matrix Classes, Encyclopedia of Math, and its Applics, vol. 108. Cambridge Univ. Press, Cambridge (2006)
5. Brualdi, R.A., Fritscher, E.: Loopy, Hankel, and combinatorially skew-Hankel tournaments. Disc. Applied. Math. to appear
6. Katona, G.Y., Faghani, M., Morteza, A., Ashrafi, A.A.R.: Centrosymmetric graphs and a lower bound for graph energy of fullerenes. Discuss. Math. Graph Theory **34**(4), 751–768 (2014)

Partially Independent Random Variables

Costanza Catalano and Alberto Gandolfi

Abstract We study collections of random variables characterized by independence requirements assigned for only a fraction of the joint values of the variables. Such classes of random variables generalize various known models, like one-dependent processes. We determine existence conditions, showing that existence is decidable, and then interpret such conditions in terms of Dutch Books. As an example, a new low density based independence model is being developed, exhibiting a phase transition in the vacuum probability.

Keywords Independent random variables · Partial independence · Existence conditions · Dutch book · Decidability

MSC 2010 codes: 60C05 · 60K35

1 Introduction

Independence is one of the key concepts in probability theory, and yet its exploitation has been somewhat limited; independent random variables, for instance, have been mainly considered when their existence poses no difficulty, either by considering collection of independent random variables, or adding independent variables to models whose existence has been already ascertained. In this paper we consider collections of random variables with mixed prescribed distributions and independence

C. Catalano
Gran Sasso Science Institute, viale F. Crispi 7, 67100 L'aquila, Italy
e-mail: costanza.catalano@gssi.infn.it

A. Gandolfi (✉)
NYU Abu Dhabi, P.O Box 129188, Abu Dhabi, UAE
e-mail: ag189@nyu.edu

A. Gandolfi
Dipartimento di Matematica e Informatica U. Dini, viale G. B. Morgagni 67/A,
50134 Firenze, Italy

T. Abualrub et al. (eds.), *Mathematics Across Contemporary Sciences*,
Springer Proceedings in Mathematics & Statistics 190,
DOI 10.1007/978-3-319-46310-0_3

requirements, in such a way that the existence or nonexistence of any such collection needs to be carefully determined depending on the details. We call such collections *partially independent random variables*.

Although we do not directly develop applications here, the possibility to analyze existence conditions and basic properties of partially independent random variables offers the availability of a new range of stochastic models which could be successfully applied to various contexts. The range of these models could parallel that of Markov random fields [6, 11] or Gibbs distributions [14]; in fact, Markov random fields are determined purely by assumptions about the conditional probabilities, and partially independent random variables are determined purely by assumptions on independence, both with additional assignments of unconditional probabilities.

One instance of the partially independent random variables has already been considered in the literature, namely the one-dependent (also called one-independent) processes (see, for instance, [12] or [18]). In fact, such case stems from a very important application [16] and constitute one instance of the so called probabilistic method [2]. The one-dependent case has then been connected to statistical mechanics in [17]. Our results here are a wide generalization of those for one-dependent processes.

In the first part of the paper we study the finite case, in which n random variables are considered, each with finite range. For each possible subcollection of the variables and each possible joint value of the variables in the subcollection, we either require independence (according to some further subdivision of the subcollection) or an assigned probability. For an assigned selection of subcollections and of the possible joint values, it is not at all obvious whether the model exists; detailed conditions are in fact needed for it to be the case or to verify nonexistence. We determine necessary and sufficient conditions for such collections of random variables to exist or not; we actually show that, assuming a technical and natural assumption, the problem is decidable, in the sense that there exists a finite time algorithm which allows to decide if the collection of random variables exists or not; unfortunately, the algorithm is exponential time, so decidability is, for practical purposes, only a preliminary step. The existence conditions that we find in the context of partially independent random variables are a particular case of a general theory about existence of probability spaces and random variables in [10], in which also decidability is treated in wider generality. Here, however, we get a self contained and more explicit characterization, amenable to a more concrete use in practice.

Examples of partially independent random variables are introduced in Sect. 4, and then analyzed in some details by means of the existence conditions. The random variables are taken to be independent as long as few of them are considered. This is a simple model, named mean field in the physics literature, which nonetheless exploits some of the power of the definition of partially independent random variables, as independence is assigned only for some carefully chosen collections of events. We call this a low density independence model, and show that it exhibits a phase transition in the vacuum probability, i.e. the probability that all the random variables take the value 0. At the time of printing, we learned that this concepts appears in the literature as K-wise independence, see [4], for instance, in which there are results

about the above mentioned phase transition, and [13] for a survey of applications in derandomization of computer algorithms.

Next, we consider the case of countably many random variables, each with finite range. We show that the conditions found in the finite case, repeated for all finite subsets of the countable set of indices, form an equivalent set of conditions for the existence of countably many partially independent random variables.

In the next part, following [10], we develop a dual interpretation of the existence problem: if there are no random variables satisfying the independence and distributional requirements then there exists a Dutch Book against the believer of such contradictory assumptions. A Dutch Book [19] is a rigging strategy that an external player might devise to rig the believer of the above unsatisfiable requirements; the strategy consists of a game in which the believer computes that (s)he has a nonnegative average gain, while the game result is a loss for each possible realization. The duality between existence of the random variables and Dutch Book is related to De Finetti's coherence (see [9] or, in general, [19]), and to the existence of a martingale measure in arbitrage free markets [15].

2 Partially Independent (p.i.) Random Variables

For $j \in \mathbb{N}$ let S_j be a finite set and let:

$$S^J = \prod_{j \in J} S_j \qquad \forall\, J \subset \mathbb{N},\ J \neq \emptyset \tag{1}$$

$$S = \bigcup_{\substack{J \subseteq \mathbb{N} \\ |J| < \infty}} S^J. \tag{2}$$

If it exists, S_j is the range of the random variable X_j. In general, if $J \neq \emptyset$ then we denote by $\mathbf{x}^J$ an element of S^J, while, given $I \subseteq J$, we let $\mathbf{x}^J_I$ indicate the components of the vector $\mathbf{x}^J$ with indices in I. If $I, J \subset \mathbb{N}$ with $I \cap J = \emptyset$ then $\mathbf{z} = (\mathbf{x}^J, \mathbf{y}^I)$ is the vector in $S^{I \cup J}$ such that $\mathbf{z}_J = \mathbf{x}^J$ and $\mathbf{z}_I = \mathbf{y}^I$. We also let $[n] = \{1, \ldots, n\}$ for every $n \in \mathbb{N}$.

Consider next:

- subsets $\Omega, \Omega' \subseteq S$ such that $\Omega \cap \Omega' = \emptyset$ (3a)
- values $\{\pi_{\mathbf{x}}\}_{\mathbf{x} \in \Omega'}$ in $[0, 1]$ (3b)
- a function ϕ defined on Ω such that $\forall\, \mathbf{x}^J \in \Omega$,
  $\phi(\mathbf{x}^J) = \{I_1, \ldots, I_{m_J}\}$, (3c)
  with $I_i \subsetneq J$ and $I_i \neq \emptyset\ \forall\, i = 1, \ldots, m_J = m_J(\mathbf{x}^J)$.

If $|J| \leq 1$, then $\mathbf{x}^J$ cannot belong to Ω as it is not possible to assign subsets $\{I_1, \ldots, I_{m_J}\}$ as in (3c); $\mathbf{x}^J$ can then only possibly belong to Ω'.

Definition 1 (π-*family and independence function*) Given Ω, Ω', $\{\pi_{\mathbf{x}}\}_{\mathbf{x}\in\Omega'}$ and ϕ as in (3a–3c), the family $\{\pi_{\mathbf{x}}\}_{\mathbf{x}\in\Omega'}$ is called π-*family* and ϕ is called *independence function*.

Given Ω, Ω', a π-family and an independence function ϕ, we want to define partially independent random variables as a collection of random variables with given properties.

Definition 2 (*Partially independent random variables*) Given S_j, $j \in \mathbb{N}$, Ω, Ω', a π-family and an independence function ϕ, partially independent random variables, also denoted by p.i. random variables, are any collection of random variables $\{X_j\}_{j\in\mathbb{N}}$ such that

$$\begin{cases} X_j S_j & \forall\, j \in \mathbb{N} \\ P(\mathbf{X}^J = \mathbf{x}^J) = \pi_{\mathbf{x}^J} & \forall\, \mathbf{x}^J \in \Omega' \\ P(\mathbf{X}^J = \mathbf{x}^J) = P\left(\mathbf{X}^{I_i} = \mathbf{x}^J_{I_i}\right) P(\mathbf{X}^{J\setminus I_i} = \mathbf{x}^J_{J\setminus I_i}) & \forall \mathbf{x}^J \in \Omega, \forall i = 1, \ldots, m_J \end{cases} \tag{4}$$

where $P(\mathbf{X}^J = \mathbf{x}^J)$ indicates $P(X_j = \mathbf{x}^J_j,\, j \in J)$.

Ω e Ω' are the sets of joint values of the random variables for which there is a requirement: in particular, for the values in Ω' we require a specific value of the joint distribution, while Ω indicates the joint values for which there is some factorization of the probabilities. The independence function ϕ indicates how the subcollection of random variables can be split up to factorize the probabilities; for this there is a compatibility condition, namely that

$$P\left(\mathbf{X}^{I_i} = \mathbf{x}^J_{I_i}\right) P(\mathbf{X}^{J\setminus I_i} = \mathbf{x}^J_{J\setminus I_i}) = P(\mathbf{X}^{I_j} = \mathbf{x}^J_{I_j}) P(\mathbf{X}^{J\setminus I_j} = \mathbf{x}^J_{J\setminus I_j})$$

for each $i, j \in \{1, \ldots, m_J\}$, that is automatically satisfied in (4) as they are both equal to $P(\mathbf{X}^J = \mathbf{x}^J)$.

Notice that we have taken conditions which are far less restrictive than assuming that the factorization takes place on all realizations of a given subcollection (in which case ϕ in (3c) would depend only on J), or than assuming that the probability $P(\mathbf{X}^J = \mathbf{x}^J)$ in (4) fully factorizes (in which case it would equal $\prod_{i=1}^{m_J} P\left(\mathbf{X}^{I_i} = \mathbf{x}^J_{I_i}\right)$ where $\phi(\mathbf{x}^J)$ is a partition of J). These more restricted assumptions are special cases of the p.i. random variables.

The first question raised by the definitions above is whether, given the π's, Ω, Ω' and ϕ, there exists at least one collection of random variables satisfying the requirements (4) based on the given elements. If at least one such collection of random variables exists, then the requirements are not contradictory; one can then use the assumptions as a model and draw consequences from them. If no such collection of

random variables exists, then the requirements are contradictory, and no consequence can be reasonably drawn; it is actually the case that a kind of dual consequence can be drawn, namely a Dutch Book, as seen in Sect. 6 below.

3 Finitely Many p.i. Random Variables

We consider now the case of finitely many variables.

3.1 Existence Conditions

Let $n \in \mathbb{N}$ be fixed. We are making a technical assumption, namely that we consider only values of the random variables which are always different from a fixed assignment of values $\mathbf{y}^{[n]} \in S^{[n]}$; assume thus that

$$\Omega \cup \Omega' = \bigcup_{\substack{J \subseteq [n] \\ J \neq \emptyset}} \{\mathbf{x}^J \in S^J : \mathbf{x}^J_j \neq \mathbf{y}^{[n]}_j \ \forall j \in J\} \cup \{\mathbf{x}^\emptyset\}, \tag{5}$$

where it is convenient to take $\mathbf{x}^\emptyset = S^{[n]}$; we always assume that $\mathbf{x}^\emptyset \in \Omega'$.

With this assumption, we can further assume that, after a possible relabelling, $\mathbf{y}^{[n]}$ is identically 0; we then have that $0 \in S_j$ for each $j \in [n]$ and $\mathbf{y}^{[n]} \equiv 0$. Let us denote

$$\hat{S}_j = S_j \setminus \{0\} \qquad \forall j \in [n] \tag{6a}$$

$$\hat{S}^J = \prod_{j \in J} \hat{S}_j \qquad \forall J \subseteq [n] \tag{6b}$$

$$\hat{S} = \bigcup_{J \subseteq [n]} \hat{S}^J \tag{6c}$$

so that $\Omega \cup \Omega' = \hat{S} \cup S^{[n]}$.

The configuration $\mathbf{y}^{[n]} \equiv 0$ can be considered the vacuum state. Its importance in the binary case is underlined by several results (see [16, 18]). It can also be considered a desirable state, as in [8].

We consider next a π-family $\{\pi_{\mathbf{x}}\}_{\mathbf{x} \in \Omega'}$ such that $\pi_{\mathbf{x}^\emptyset} = 1$ and an independence function ϕ. It follows from (5) that

$$\mathbf{x}^J \in \Omega \cup \Omega' \quad \Rightarrow \quad \mathbf{x}^J_{J'} \in \Omega \cup \Omega' \quad \forall J' \subseteq J.$$

For every $\mathbf{x}^J \in \Omega$, it is possible to continue decomposing $P(\mathbf{X}^J = \mathbf{x}^J)$ using the independence function ϕ and the rules in (4) till it is a product of π's. In fact,

the independence function indicates how the probability of a joint assignment $\mathbf{x}^J$ is expressed as a product of probabilities of assignments in subsets of J; these are either themselves decomposed, or expressed in terms of π's, as they are always in $\Omega \cup \Omega'$. More formally, for every $\mathbf{x}^J \in \Omega$ it is possible to find $m'_J(\mathbf{x}^J) = m'_J$ partitions of J (in general $m'_J \leq m_J$), indicated by $\mathscr{X}_i^J = \{J_{i,1}, \ldots, J_{i,l}\}$ for $i = 1, \ldots, m'_J$, where $l = l(i, \mathbf{x}^J)$ is the number of elements of the partition $\mathscr{X}_i^J$, such that the last condition in (4) can be expressed as:

$$P(\mathbf{X}^J = \mathbf{x}^J) = \prod_{k=1}^{l} P\left(\mathbf{X}^{J_{i,k}} = \mathbf{x}^J_{J_{i,k}}\right) = \prod_{k=1}^{l} \pi_{\mathbf{x}^J_{J_{i,k}}} \tag{7}$$

for every $i = 1, \ldots, m'_J$. Clearly, the set of partitions $\{\mathscr{X}_i^J\}_{i \in [m'_J]}$ depends on the whole of $\mathbf{x}^J$ and not only on J: the dependency is not explicitly indicated to simplify the notation. If for some $J \subseteq [n]$ there exist $i, j \in [m'_J]$ such that $\prod_{k=1}^{l} \pi_{\mathbf{x}^J_{J_{i,k}}} \neq \prod_{k=1}^{l'} \pi_{\mathbf{x}^J_{J_{j,k}}}$, then clearly no p.i. random variables can satisfy (4). If

$$\prod_{k=1}^{l} \pi_{\mathbf{x}^J_{J_{i,k}}} = \prod_{k=1}^{l'} \pi_{\mathbf{x}^J_{J_{j,k}}} \qquad \forall\, i, j \in [m'_J], \quad \forall\, J \subseteq [n] \tag{8}$$

holds we say that the independence function ϕ is *compatible* with the $\pi-$family $\{\pi_{\mathbf{x}}\}_{\mathbf{x} \in \Omega'}$.

Compatibility between the independence function and the π-family allows to fix one partition $\mathscr{X}^J = \{J_1, \ldots, J_l\}$ for every $\mathbf{x}^J \in \Omega$ such that

$$P(\mathbf{X}^J = \mathbf{x}^J) = \prod_{k=1}^{l} \pi_{\mathbf{x}^J_{J_k}}$$

and thus define a function $f : \Omega \cup \Omega' \to [0, 1]$ such that:

$$f(\mathbf{x}^J) = \begin{cases} \pi_{\mathbf{x}^J} & \text{if } \mathbf{x}^J \in \Omega' \\ \prod_{k=1}^{l} \pi_{\mathbf{x}^J_{J_k}} & \text{if } \mathbf{x}^J \in \Omega, \text{ with } \{J_1, \ldots, J_l\} = \mathscr{X}^J. \end{cases} \tag{9}$$

We can then rewrite (4) as

$$\begin{cases} X_j \text{ takes value in } S_j & \forall\, j \in [n] \\ P(\mathbf{X}^J = \mathbf{x}^J) = f(\mathbf{x}^J) & \forall\, \mathbf{x}^J \in \Omega \cup \Omega'; \end{cases} \tag{10}$$

we further assume that

$$\pi_{\mathbf{x}^\emptyset} = 1. \tag{11}$$

The following theorem expresses existence conditions. Notice that Eq. (12) below are linear in the probabilities: a priori they were not so, as we are talking about independence; they are not linear in terms of the parameters $\{\pi_{\mathbf{x}}\}_{\mathbf{x}\in\Omega'}$, though.

Theorem 1 *Let Ω and Ω' be as in (5), $\{\pi_{\mathbf{x}}\}_{\mathbf{x}\in\Omega'}$ a $\pi-$family satisfying (11) and ϕ an independence function compatible with the π-family. Let f be the function defined in (9).*
Then there exist p.i. random variables $\mathbf{X} = \{X_1, \ldots, X_n\}$ satisfying (4) if and only if

$$\sum_{J\subseteq I\subseteq[n]} (-1)^{|I\setminus J|} \sum_{\substack{\mathbf{x}^I\in\hat{S}^I:\\ \mathbf{x}^I_J=\hat{\mathbf{z}}^J}} f(\mathbf{x}^I) \geq 0 \qquad \forall\, \hat{\mathbf{z}}^J \in \hat{S}^J, \forall\, J \subseteq [n]. \tag{12}$$

Proof If $\mathbf{X} = \{X_1, \ldots, X_n\}$ are p.i. random variables satisfying (4), then for some probability measure P (10) holds. We now show that, as P is a probability measure,

$$\begin{aligned}\sum_{J\subseteq I\subseteq[n]} (-1)^{|I\setminus J|} \sum_{\substack{\mathbf{x}^I\in\hat{S}^I\\ \mathbf{x}^I_J=\hat{\mathbf{z}}^J}} f(\mathbf{x}^I) &= \sum_{J\subseteq I\subseteq[n]} (-1)^{|I\setminus J|} \sum_{\substack{\mathbf{x}^I\in\hat{S}^I\\ \mathbf{x}^I_J=\hat{\mathbf{z}}^J}} P(\mathbf{X}^I = \mathbf{x}^I)\\ &= \sum_{I\subseteq J^c} (-1)^{|I|} \sum_{\mathbf{x}^I\in\hat{S}^I} P(\mathbf{X}^J = \hat{\mathbf{z}}^J, \mathbf{X}^I = \mathbf{x}^I)\\ &= P(\mathbf{X}^J = \hat{\mathbf{z}}^J, \mathbf{X}^{J^c} = \mathbf{0}) \geq 0, \end{aligned} \tag{13}$$

holds, where J^c indicates $[n] \setminus J$ and $P(\mathbf{X}^{\emptyset} = \mathbf{x}^{\emptyset}) = 1$. We need to verify the last equality, and we proceed by induction on $|J^c|$. If $|J^c| = 1$ then $J^c = \{i\}$ for some $i \in [n]$, and the statement follows by the additivity property of probability measures. Suppose next that the last equality of (13) holds for every $J^c : |J^c| \leq m - 1$, and let $\hat{\mathbf{z}}^J \in \hat{S}^J$; by inclusion-exclusion and induction hypothesis, we have then

$$\begin{aligned} &P(\mathbf{X}^J = \hat{\mathbf{z}}^J, \mathbf{X}^{J^c} = \mathbf{0})\\ =&P(\mathbf{X}^J = \hat{\mathbf{z}}^J) - \sum_{\substack{\mathbf{x}^{J^c}\in S^{J^c}\\ \mathbf{x}^{J^c}\neq\mathbf{0}}} P(\mathbf{X}^J = \hat{\mathbf{z}}^J, \mathbf{X}^{J^c} = \mathbf{x}^{J^c})\\ =&P(\mathbf{X}^J = \hat{\mathbf{z}}^J) - \left(\sum_{\phi\neq I\subsetneq J^c} \sum_{\mathbf{x}^I\in\hat{S}^I} P(\mathbf{X}^{J^c\setminus I} = \mathbf{0}, \mathbf{X}^I = \mathbf{x}^I, \mathbf{X}^J = \hat{\mathbf{z}}^J) \right)\\ &\qquad - \sum_{\mathbf{x}^{J^c}\in\hat{S}^{J^c}} P(\mathbf{X}^{J^c} = \mathbf{x}^{J^c}, \mathbf{X}^J = \hat{\mathbf{z}}^J) \end{aligned}$$

$$= P(\mathbf{X}^J = \mathbf{z}^J)$$
$$- \underbrace{\sum_{\phi \neq I \subsetneq J^c} \sum_{I' \subseteq J^c \setminus I} (-1)^{|I'|} \sum_{\substack{\mathbf{x}^{I'} \in \hat{S}^{I'} \\ \mathbf{x}^I \in \hat{S}^I}} P(\mathbf{X}^{I'} = \mathbf{x}^{I'}, \mathbf{X}^I = \mathbf{x}^I, \mathbf{X}^J = \hat{\mathbf{z}}^J)}_{(a)}$$
$$- \underbrace{\sum_{\mathbf{x}^{J^c} \in \hat{S}^{J^c}} P(\mathbf{X}^J = \hat{\mathbf{z}}^J, \mathbf{X}^{J^c} = \mathbf{x}^{J^c})}_{(b)};$$

we have to show that this equals

$$P(\mathbf{X}^J = \hat{\mathbf{z}}^J) + \sum_{\substack{I \subseteq J^c \\ I \neq \emptyset}} (-1)^{|I|} \sum_{\mathbf{x}^I \in \hat{S}^I} P(\mathbf{X}^I = \mathbf{x}^I, \mathbf{X}^J = \hat{\mathbf{z}}^J). \tag{14}$$

For every $\Lambda \subseteq J^c$, $\Lambda \neq \emptyset$ and every $\mathbf{x}^\Lambda \in \hat{S}^\Lambda$ we sum the coefficients in (a) and (b) of $P(\mathbf{X}^\Lambda = \mathbf{x}^\Lambda, \mathbf{X}^J = \hat{\mathbf{z}}^J)$:

- If $\Lambda = J^c$ then in (a) we have $I' \cup I = J^c$, thus $I' = J^c \setminus I$, and the coefficient is $\sum_{s=1}^{|J^c|-1} \binom{|J^c|}{s} (-1)^s = -1 - (-1)^{|J^c|}$; in (b), the coefficient is 1. Altogether we get $-(-1 - (-1)^{|J|}) - 1 = (-1)^{|J|}$ as in (14).
- If $\Lambda \subset J$ then in (a) we have $I' \cup I = \Lambda$, so $I' = \Lambda \setminus I$, and the coefficient becomes $\sum_{s=0}^{|\Lambda|-1} \binom{|\Lambda|}{s} (-1)^s = -(-1)^{|\Lambda|}$, while in (b) the coefficient is 0. Altogether we get $-(-(-1)^{|\Lambda|}) = (-1)^{|\Lambda|}$ as in (14).

To show the other direction, it is enough to find a collection of random variables satisfying (10). Let $\bar{X}_j \colon S^{[n]} \to S_j$ with $\bar{X}_j(\omega) = \omega_j$. Let $\bar{\mathscr{A}} = \mathscr{P}(S^{[n]})$ be the σ-algebra of all subsets of $S^{[n]}$. For $\omega \in S^{[n]}$, let $0_\omega = \{i : \omega_i = 0\}$, and $1_\omega = 0_\omega^c$. Let

$$\bar{P}(\omega) = \sum_{I : 1_\omega \subseteq I \subseteq [n]} (-1)^{|I \setminus 1_\omega|} \sum_{\substack{\mathbf{x}^I \in \hat{S}^I \\ \mathbf{x}^I_{1_\omega} = \omega_{1_\omega}}} f(\mathbf{x}^I) = \tag{15}$$
$$= \sum_{I \subseteq 0_\omega} (-1)^{|I|} \sum_{\mathbf{x}^I \in \hat{S}^I} f(\omega_{1_\omega}, \mathbf{x}^I)$$

and then extend $\bar{P}$ on $\bar{\mathscr{A}}$ by setting, for $\bar{A} \subseteq S^{[n]}$, $\bar{P}(\bar{A}) = \sum_{\omega \in \bar{A}} \bar{P}(\omega)$. We have that $\bar{P}$ is nonnegative by (12) and (countably) additive. In particular, if $\mathbf{x}^J \in \Omega \cup \Omega'$ then

$$\begin{aligned}\bar{P}(\mathbf{X}^J = \mathbf{x}^J) &= \sum_{\omega:\, \omega_J = \mathbf{x}^J} \bar{P}(\omega) \\ &= \sum_{\omega:\, \omega_J = \mathbf{x}^J} \sum_{I \subseteq 0_\omega} (-1)^{|I|} \sum_{\hat{\mathbf{x}}^I \in \hat{S}^I} f(\omega_{1_\omega}, \hat{\mathbf{x}}^I) \qquad (16) \\ &= f(\mathbf{x}^J) + \sum_{\omega:\, \omega_J = \mathbf{x}^J, 1_\omega \neq J} \sum_{I \subseteq 1_\omega \setminus J} (-1)^{|I|} f(\omega_{1_\omega}) = f(\mathbf{x}^J)\end{aligned}$$

as $\sum_{I \subseteq A} (-1)^{|I|} = 0$ for any set A.

It remains to be shown that $\sum_{\omega \in S^{[n]}} P(\omega) = 1$. By the same calculation as in (16) we have

$$\sum_{\omega \in \bar{\Omega}} \bar{P}(\omega) = \sum_{\omega \in \bar{\Omega}} \sum_{I \subseteq 0_\omega} (-1)^{|I|} \sum_{\hat{\mathbf{x}}^I \in \hat{S}^I} f(\omega_{1_\omega}, \hat{\mathbf{x}}^I) = f(\mathbf{x}^\emptyset) = \pi_{\mathbf{x}^\emptyset} = 1.$$

□

3.2 Decidability

We summarize the results obtained so far by indicating an algorithm which decides whether, given a $\pi-$family and an independent function ϕ, there exists a finite collection of random variables satisfying (4).

In general, it is not necessarily the case that $\Omega, \Omega' \subseteq \hat{S}$, and we can allow simply that $\Omega, \Omega' \subseteq S$. On the other hand, to make use of the previous results let us still assume that $\hat{S} \subseteq \Omega \cup \Omega'$. We can then proceed as follows.

1. Verify if ϕ restricted to $\hat{S}$ satisfies (8) with respect to the family $\{\pi_{\mathbf{x}}\}_{\mathbf{x} \in \hat{S}}$ and if $\pi_{\mathbf{x}^\emptyset} = 1$. If the conditions are not satisfied then no p.i. random variables $\mathbf{X}$ satisfying (4) exist.
2. Define the function f on $\hat{S}$ as in (9). If f does not satisfy (12) then no p.i. random variables $\mathbf{X}$ satisfying (4) exist.
3. Use inclusion-exclusion to compute the probability of all other collection of values in $(\Omega \cup \Omega') \setminus \hat{S}$ from the probabilities of the values in $\hat{S}$. If one of the computed values does not correspond to the one given in (4) for the random variables $\mathbf{X}$, which is to say, it does not equal the corresponding value in $\{\pi_{\mathbf{x}^J}\}_{\mathbf{x}^J \in \Omega' \setminus \hat{S}}$ or in $\phi\,|_{\Omega \setminus \hat{S}}$, then no p.i. random variables $\mathbf{X}$ satisfying (4) exist.

If all the steps above have been satisfied, then we have p.i. random variables $\mathbf{X}$.

Notice that all the three steps above contain a procedure which ends in finite time, so the problems of existence of p.i. random variables $\mathbf{X}$ satisfying (4) is decidable. Notice also that some of the steps, in particular Step 2., require a number of comparisons which depends on the number of subsets of $[n]$, and hence it is exponential

time. Therefore, the determination of existence of p.i. random variables $\mathbf{X}$ satisfying (4) for a given $\pi-$family is a decidable, but exponential time algorithm.

There is an alternative strategy to show decidability which is presented in Sect. 6.

3.3 Partially Independent Events

We consider the case of binary random variables X_i's; this is equivalent to that of considering events A_i, $i = 1, \ldots, n$ with the appropriate independence conditions, which we then call partially independent events. In this case, (12) basically reduces to the formula of inclusion-exclusion. We write it explicitly, as it is used in a later example.

Let $S_i = \{0, 1\}$, $A_i = X_i^{-1}(\{1\})$ and denote $A_i^c = X_i^{-1}(\{0\})$. Let $A_i^1 = A_i$ and $A_i^0 = A_i^c$ so that for $\mathbf{x}^J \in S$ we have $P(\mathbf{X}^J = \mathbf{x}^J) = P(\bigcap_{j\in J} A_j^{\mathbf{x}_j^J})$. Note that in this case $\Omega \cup \Omega' = \{\mathbf{1}^J : J \subseteq [n]\}$ so that the condition on avoiding $\mathbf{y} \equiv 0$ is satisfied. The next corollary follows from Theorem 1.

Corollary 1 *Let $S_j = \{0, 1\}$ for every $j = 1, \ldots, n$. Let Ω and Ω' be as in (5), $\{\pi_{\mathbf{x}}\}_{\mathbf{x}\in\Omega'}$ a $\pi-$family with the property (11), ϕ an independence function compatible with the π-family and f the function defined in (9). Then there exist n events $\mathbb{A} = \{A_1, \ldots, A_n\}$, in a suitable probability space, satisfying*

$$P\left(\bigcap_{j\in J} A_j\right) = f(J) \quad \forall J \subseteq [n] \tag{17}$$

if and only if

$$\sum_{J\subseteq I\subseteq[n]} (-1)^{|I\setminus J|} f(I) \geq 0 \qquad \forall J \subseteq [n]. \tag{18}$$

3.4 Related Works

Clearly, i.i.d. random variables are p.i. random variables. The next natural example of p.i. random variables is block factors: consider i.i.d. random variables Y_j, j in some set of indices Λ and for each $i \in \Lambda$ consider a subset $B_i \subseteq \Lambda$ and a function f_i defined on the range of the joint distribution of the $\mathbf{Y}_i = \{Y_j\}$, $j \in B_i$. The block factor is the collection of random variables $X_i = f_i(\mathbf{Y}_i)$. Clearly, X_{i_1} and X_{i_2} are independent if $B_{i_1} \cap B_{i_2} = \emptyset$. This defines p.i. random variables as follows: let Ω be the collection of ranges of the joint distribution of the X_i's in subsets of Λ of cardinality at least two and such that for any two indices i_1 and i_2 in the subset $B_{i_1} \cap B_{i_2} = \emptyset$ holds; let Ω' be the collection of ranges of the single variables X_i, and $\pi_{\mathbf{x}} := P(X_i = \mathbf{x})$ if $\mathbf{x}$ is in the range of X_i. Then the block factor satisfies the

requirements. Note that in this example, as in the others below, the requirement of independence of an assignment of probability depends effectively on the set of indices, and not on the particular value taken by the random variables with indices in the set: as noted, this is a more restrictive requirement then the general one of p.i. random variables.

A special case of the block factors is when the indices are vertices of some graph $G = (V, E)$ and B_i is the set of neighbors of i in G. Our definition of p.i. random variables matches in this case that of strong dependency graph (see [5, 8]):

Definition 3 Let $G = (V, E)$ be a locally finite graph with $V = \mathbb{N}$ and consider a family of finite valued random variables $\mathbf{X} = \{X_n\}_{n\in\mathbb{N}}$. We say that $\mathbf{X}$ has G as *strong dependency graph* if for every finite $W_1, W_2 \subset V$

$$d(W_1, W_2) > 1 \Rightarrow \mathbf{X}_{W_1} \text{ is independent from } \mathbf{X}_{W_2}$$

where $d(W_1, W_2)$ indicates the graph distance between W_1 and W_2, and $\mathbf{X}_W = \{X_i : i \in W\}$.

In this case the random variables are also called one-dependent (or, occasionally, one-independent). Block factors, with the set of neighbors as block, are one-dependent. It is then natural to wonder if in this case all one-dependent random variables are obtained from a block factor, but this is not the case (see [1]).

As one-dependent processes (and a fortiori p.i. random variables) are more general than block factors, one is now interested in inequalities for certain quantities as others are fixed. For instance, in [12] maximal and minimal correlations of one-dependent processes with 0–1 variables are computed, as function of the probability of 1, for stationary p.i. random variables. The one-dependent processes were also originated in a very important application [16] and connected to model of the so-called *hard-core pair interaction* gasses with negative fugacity in statistical mechanics in [17].

See also the literature on K-wise independence, in [4], and [13] for a survey of applications in derandomization of computer algorithms.

4 Example: Low Density Independence

We present here a class of p.i. random variables which has appeared in the literature only in the binary case (see [4]).

4.1 Binary Variables

We consider first the binary case: we take then 0–1 valued exchangeable random variables (expressed in terms of events). For fixed n, $k \leq n$ and $\alpha \in [0, 1]$, we call low density independent events a collection of events $A_i, i = 1, \ldots, n$ such that:

1. The A_i's are exchangeable.
2. Any sub collection with less than or equal to k events is independent.
3. $P(A_i) = \alpha$.

Call any such distribution a low density p.i. distribution and denote it by $P_{n,k,\alpha}$. Clearly, the assumptions are always consistent, as jointly independent events satisfying 3. do always exist.

Condition (18) can be rewritten as follows. Let $\beta_r = P(A_{i_1} \cap \cdots \cap A_{i_r})$ for some set of distinct indices of cardinality r; β_r does not depend on the selection of indices by exchangeability. Then, taking $|J| = s$, (18) becomes

$$R_{n,k,\alpha}(s) = \alpha^s \sum_{r=0}^{(k-s)\vee 0} \binom{n-s}{r} \alpha^r (-1)^r + \sum_{r=(k-s+1)\vee 0}^{n-s} \binom{n-s}{r} \beta_{r+s} (-1)^r \geq 0 \tag{19}$$

for $s = 0, \ldots, n-1$.
Here is an example with only pairwise independence.

Example 1 For any $n \in \mathbb{N}$, $k = 2$, $\alpha = 1/(n-1)$, let $\beta_r = \alpha^2 = 1/(n-1)^2$ for $r = 3, \ldots, n$. This is a solution of (19) corresponding to $P(A_i) = 1/(n-1)$ and $P(\cap_{j=1}^r A_{i_j}) = 1/(n-1)^2$ for $r = 2, \ldots, n$. In fact, (19) becomes

$$F(s) = \alpha^s \sum_{r=0}^{2-s} \binom{n-s}{r} (-1)^r \alpha^r + \alpha^2 \sum_{r=(2-s+1)\vee 0}^{n-s} \binom{n-s}{r} (-1)^r, \tag{20}$$

where the first sum appears only if $2 - s \geq 0$. Then we have $F(0) = 0$ as $\alpha = 1/(n-1)$, $F(1) = \alpha - \alpha^2 > 0$, and $F(s) = 0$ for all $s \geq 2$.

Notice that in the example, there are only three types of configurations: all A_i^c's, exactly one A_i, or more A_i's. At the other extreme, events could always be independent, except when there are n of them.

Example 2 For any $n \in \mathbb{N}, k = n-1, \alpha \geq 1/2$, let $\beta_n = \alpha^n + (1-\alpha)^n$. These values form a solution of (19). For a verification see the proof of Theorem 4.3 below.

Notice that in these examples we have exploited the full potentialities of p.i. random variables, at least in the binary case, as independence depended upon the entire configuration.

It is interesting to study the vacuum probability $p_{n,k}(\alpha) = \inf P_{n,k,\alpha}(\cap_{i=1}^n A_i^c)$ as the infimum ranges over all low density p.i. distributions for given n, k and α. We consider, in analogy to [8] and ensuing literature, whether $p_{n,k}(\alpha) > 0$ or $p_{n,k}(\alpha) = 0$. Notice that $P_{n,k,\alpha}(\cap_{i=1}^n A_i^c) = R_{n,k,\alpha}(0)$, so this question amounts to add the condition

4. $R_{n,k,\alpha}(0) = 0$

to 1.–3. and verify that there are no probability distributions satisfying all of the four conditions together. As there are always some distributions satisfying the first three conditions, this would imply the strict inequality. We then need to verify that there are no solutions of

$$R_{n,k,\alpha}(0) = 0, \quad R_{n,k,\alpha}(s) \geq 0 \ \text{ for } \ s = 1, \ldots, n-1. \tag{21}$$

The distribution in Example 1 satisfies $R_{n,k,\alpha}(0) = 0$, hence $p_{n,2}(1/(n-1)) = 0$. Similarly in Example 2. We now see that there is a phase transition in α.

Theorem 2 *For all $n, k \leq n-1$:*

(I) If $\alpha \geq 1/2$, then $p_{n,k}(\alpha) = 0$.
(II) If $\alpha < \overline{\alpha}$, where $\overline{\alpha}$ is the smallest solution in $[0, 1]$ of the polynomial equation:

$$\begin{aligned}(1-\alpha)^n + (-1)^k \alpha^k \binom{n}{k+1} {}_2F_1(1, 1+k-n, 2+k, \alpha) \\ = \alpha^k \binom{n}{k+1} {}_2F_1(1, 1+k-n, 2+k, -1)\end{aligned} \tag{22}$$

where ${}_2F_1$ is Gauss hypergeometric function, then $p_{n,k}(\alpha) > 0$.

Proof Part (I) is shown by completing the calculation of Example 2. Since $n-1 \geq k$, we can consider the particular distribution of Example 2 in which the probabilities of up to $n-1$ events factorize; the only remaining probability which does not factorize is that of all n events together, which is taken to be $\beta_n = \alpha^n + (1-\alpha)^n$. We have $\beta_n \geq 0$ and $\beta_n \leq \alpha^{n-1}$, since $\alpha \geq 1/2$, as it should be. For every s,

$$\begin{aligned}R_{n,k,\alpha}(s) &= \alpha^s \sum_{r=0}^{n-s-1} \binom{n-s}{r} (-1)^r \alpha^r + (-1)^{n-s}(\alpha^n - (1-\alpha)^n) \\ &= \alpha^s \big((1-\alpha)^{n-s} + (-1)^{n-s+1}\alpha^{n-s}\big) + (-1)^{n-s}\big(\alpha^n - (1-\alpha)^n\big).\end{aligned}$$

For $(n-s)$ even this is greater or equal than 0 as $\alpha \geq 1-\alpha$. For $(n-s)$ odd, the above expression becomes

$$\big(\alpha^s + (1-\alpha)^s\big)(1-\alpha)^{n-s} \geq 0.$$

Moreover, $R_{n,k,\alpha}(0) = P_\alpha(\cap_{i=1}^n A_i^c) - (1-\alpha)^n = 0$, where P_α is the Binomial distribution with parameter α, under which all events are independent.

To show Part (II) observe that

$$\sum_{r=0}^{k} \binom{n}{r} \alpha^r (-1)^r = (1-\alpha)^n + (-1)^k \alpha^k \binom{n}{k+1} {}_2F_1(1, 1+k-n, 2+k, \alpha) \tag{23}$$

and

$$\sum_{r=k+1}^{n} \binom{n}{r} = \binom{n}{k+1} {}_2F_1(1, 1+k-n, 2+k, -1). \tag{24}$$

We have that $\beta_r < \alpha^k$ for all $r \geq k+1$; therefore,

$$\begin{aligned} p_{n,k}(\alpha) &\geq \sum_{r=0}^{k} \binom{n}{r} \alpha^r (-1)^r + \sum_{r=k+1}^{n} \beta_r \binom{n}{r} (-1)^r \\ &\geq \sum_{r=0}^{k} \binom{n}{r} \alpha^r (-1)^r - \sum_{r=k+1}^{n} \beta_r \binom{n}{r} \\ &\geq \sum_{r=0}^{k} \binom{n}{r} \alpha^r (-1)^r - \alpha^k \sum_{r=k+1}^{n} \binom{n}{r} > 0 \end{aligned}$$

for all $\alpha \leq \overline{\alpha}$, since at $\alpha = 0$ the function equals 1, and $\overline{\alpha}$ is the smallest solution of (22) in $[0, 1]$, written in terms of (23) and (24). □

It makes sense to expect that there is a critical α_c such that $p_{n,k}(\alpha) > 0$ for $\alpha < \alpha_c$ and $p_{n,k}(\alpha) = 0$ for $\alpha > \alpha_c$. Clearly, $\alpha_c = 1$ for $k = n$. Figure 1 shows the numerical values of the bounds we have obtained for α_c.

None of the two bounds is tight:

Example 3 For $n = 6, k = 4$, α_c is not equal to any of the bounds, as by solving all the polynomial relations by cylindrical reduction one gets $\overline{\alpha} \leq 0.377 < \alpha_c \approx 0.419 < 1/2$.

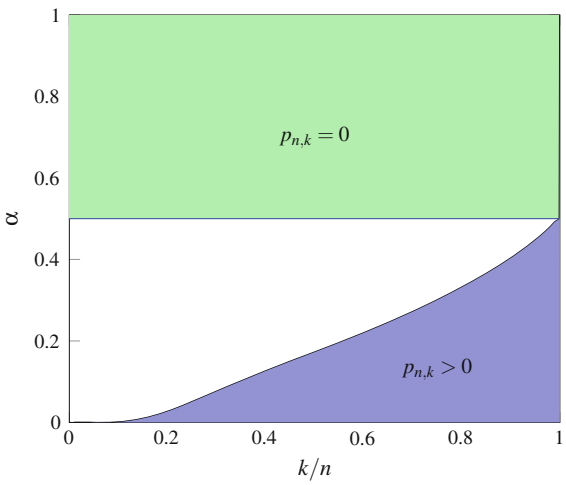


Fig. 1 Bounds for the phase diagram of $p_{n,k}$. None of the bounds is tight; especially notice that $p_{n,n} > 0$ for all $\alpha < 1$

4.2 Ternary Variables

This last example illustrates the full power of the general definition of p.i. random variables, and of our existence conditions. For fixed n, $k \leq n$ and $\alpha_1, \alpha_2 \in [0, 1]$, $\alpha_1 + \alpha_2 \leq 1$, we call low density independent (ternary) random variables a collection of random variables $X_i, i = 1, \ldots, n$ such that:

1. $X_j \in S = \{0, 1, 2\}$ for all $j \in [n]$.
2. The X_j's are exchangeable.
3. For any $J \subseteq [n] : |J| \leq k$ the random variables X_j, $j \in J$ are independent.
4. $P(X_j = \ell) = \alpha_\ell$ for $\ell = 1, 2$.

As before, we call any such distribution a low density p.i. distribution and denote it by $P_{n,k,\alpha_1,\alpha_2}$. Clearly, the assumptions are always consistent.

Condition (12) can now be rewritten as follows. Let $r_1, r_2 \in \mathbb{N}$ be such that $r_1 + r_2 \leq n$, $J \subseteq [n]$ with $|J| = r_1 + r_2$, and $x^J \in S^J$ be such that $|\{j : x_j^J = 1\}| = r_1$, $|\{j : x_j^J = 2\}| = r_2$; then let $\beta_{r_1,r_2} = P_{n,k,\alpha_1,\alpha_2}(X^J = x^J)$, which does not depend on the indices in which x^J takes particular values by exchangeability. Then letting

$$L(k, s_1, s_2) = \{(r_1, r_2) : r_1, r_2 = 0, \ldots, k - s_1 - s_2 \vee 0, r_1 + r_2 \leq k - s_1 - s_2\}$$

and

$$M(k, s_1, s_2) = \{(r_1, r_2) : r_1, r_2 = 0, \ldots, n - s_1 - s_2 \vee 0, k - s_1 - s_2 \leq r_1 + r_2\}$$

(12) becomes

$$\begin{aligned} &R_{n,k,\alpha_1,\alpha_2}(s_1, s_2) \qquad (25)\\ &= \alpha_1^{s_1}\alpha_2^{s_2} \sum_{(r_1,r_2)\in L(k,s_1,s_2)} \binom{n - s_1 - s_2}{r_1, r_2, n - s_1 - s_2 - r_1 - r_2} \alpha_1^{r_1}\alpha_2^{r_2}(-1)^{r_1+r_2} \\ &\quad + \sum_{(r_1,r_2)\in M(k,s_1,s_2)} \binom{n - s_1 - s_2}{r_1, r_2, n - s_1 - s_2 - r_1 - r_2} \beta_{r_1+s_1,r_2+s_2}(-1)^{r_1+r_2} \end{aligned}$$

for $s_1, s_2 = 0, \ldots, n - 1$, $s_1 + s_2 \leq n - 1$.

As in the binary case, it is interesting to study the vacuum probability $p_{n,k}(\alpha_1, \alpha_2) = \inf P_{n,k,\alpha_1,\alpha_2}(X_j = 0 \text{ for all } j)$ as the infimum ranges over all low density p.i. distribution for given n, k and α_1, α_2. The results in the binary case suggest now the following result:

Theorem 3 *For all n, k, with $k \leq n - 1$:*

(I) If $\alpha_1 + \alpha_2 \geq 1/2$, then $p_{n,k}(\alpha_1, \alpha_2) = 0$.
(II) If α_1, α_2 are such that

$$\sum_{r_1,r_2\in L(k,0,0)}\binom{n}{r_1,r_2,n-r_1-r_2}\alpha_1^{r_1}\alpha_2^{r_2}(-1)^{r_1+r_2}$$
$$> \alpha_1^{\lfloor\frac{k+1}{2}\rfloor}\alpha_2^{\lfloor\frac{k+1}{2}\rfloor}\sum_{\lfloor\frac{k+1}{2}\rfloor\le r_1,r_2\le n}\binom{n}{r_1,r_2,n-r_1-r_2} \tag{26}$$
$$+\sum_{r_1=0}^{\lfloor\frac{k+1}{2}\rfloor-1}\alpha_1^{r_1}\alpha_2^{k+1-r_1}\sum_{r_2=k+1-r_1}^{n}\binom{n}{r_1,r_2,n-r_1-r_2}$$
$$+\sum_{r_2=0}^{\lfloor\frac{k+1}{2}\rfloor-1}\alpha_2^{r_2}\alpha_1^{k+1-r_2}\sum_{r_1=k+1-r_1}^{n}\binom{n}{r_1,r_2,n-r_1-r_2}$$

then $p_{n,k}(\alpha_1,\alpha_2)>0$.

Proof Part (*I*) is shown using the result of the binary case, as easily seen by lumping together states 1 and 2.

To show Part (*II*) observe that $\beta_{r_1,r_2}\le\alpha_1^{r_1'}\alpha_2^{r_2'}$ for all pairs r_1',r_2' which satisfy $r_1'\le r_1$, $r_2'\le r_2$ and $r_1'+r_2'\le k$. This implies that

$$p_{n,k}(\alpha_1,\alpha_2)\ge R_{n,k,\alpha_1,\alpha_2}(0,0) \tag{27}$$
$$\ge\alpha_1\alpha_2\sum_{r_1,r_2\in L(k,0,0)}\binom{n}{r_1,r_2,n-r_1-r_2}\alpha_1^{r_1}\alpha_2^{r_2}(-1)^{r_1+r_2}$$
$$-\sum_{r_1,r_2\in M(k,0,0)}\binom{n}{r_1,r_2,n-r_1-r_2}\beta_{r_1,r_2}$$
$$\ge\alpha_1\alpha_2\sum_{r_1,r_2\in L(k,0,0)}\binom{n}{r_1,r_2,n-r_1-r_2}\alpha_1^{r_1}\alpha_2^{r_2}(-1)^{r_1+r_2}$$
$$-\alpha_1^{\lfloor\frac{k+1}{2}\rfloor}\alpha_2^{\lfloor\frac{k+1}{2}\rfloor}\sum_{\lfloor\frac{k+1}{2}\rfloor\le r_1,r_2\le n}\binom{n}{r_1,r_2,n-r_1-r_2} \tag{28}$$
$$-\sum_{r_1=0}^{\lfloor\frac{k+1}{2}\rfloor-1}\alpha_1^{r_1}\alpha_2^{k+1-r_1}\sum_{r_2=k+1-r_1}^{n}\binom{n}{r_1,r_2,n-r_1-r_2}$$
$$-\sum_{r_2=0}^{\lfloor\frac{k+1}{2}\rfloor-1}\alpha_2^{r_2}\alpha_1^{k+1-r_2}\sum_{r_1=k+1-r_1}^{n}\binom{n}{r_1,r_2,n-r_1-r_2}>0$$

if (26) holds. □

It makes sense to expect that there is a critical region R in the $\alpha_1-\alpha_2$ plane such that $p_{n,k}(\alpha_1,\alpha_2)>0$ for $(\alpha_1,\alpha_2)\in R$ and $p_{n,k}(\alpha)=0$ for $(\alpha_1,\alpha_2)\notin R$. Simple numerical calculations give bounds for the phase transition region which are extensions of the bounds shown for the binary case.

5 Countably Many p.i. Random Variables

The results of the Sect. 3 can be extended to the case of countably many variables using Kolmogorov extension theorem.

As before, suppose $0 \in S_j$ for every $j \in \mathbb{N}$, let $\hat{S}_j$ and $\hat{S}^J$ be as in (6),

$$\hat{S} = \bigcup_{\substack{J \subset \mathbb{N} \\ |J| < \infty}} \hat{S}^J$$

and consider

$$\Omega \cup \Omega' = \hat{S} \cup \{\mathbf{x}^{\emptyset}\}. \tag{29}$$

We have again that $\mathbf{x}^J \in \Omega \cup \Omega'$ implies $\mathbf{x}^I_{J'} \in \Omega \cup \Omega'$ for every $J' \subseteq J$. Let $\{\pi_{\mathbf{x}}\}_{\mathbf{x} \in \Omega'}$ be a $\pi-$family satisfying (11) and ϕ an independence function. As $\Omega \cup \Omega'$ contains only finite collection of values, all properties about compatibility between the π-family and the independence function still hold. We then require that (8) holds for every finite $J \subset \mathbb{N}$, i.e. that ϕ is *compatible* with the π-family. As before, we define the function $f\colon \Omega \cup \Omega' \to [0, 1]$ such that formula (9) holds, in which $\mathscr{X}^J$ is a partition of J dependent on $\mathbf{x}^J$ (see Sect. 3). Requirements (4) can be written as

$$\begin{cases} X_j \text{ takes value in } S_j & \forall\, j \in \mathbb{N} \\ P(\mathbf{X}^J = \mathbf{x}^J) = f(\mathbf{x}^J) & \forall\, \mathbf{x}^J \in \Omega \cup \Omega'. \end{cases} \tag{30}$$

Then we have the following theorem.

Theorem 4 *Let Ω and Ω' be as in (29), $\{\pi_{\mathbf{x}}\}_{\mathbf{x} \in \Omega'}$ a $\pi-$family satisfying (11) and ϕ an independence function compatible with the π-family. Let f be the function defined in (9). Then there is a collection of random variables $\mathbf{X} = \{X_i\}_{i \in \mathbb{N}}$ satisfying (30) if and only if*

$$\sum_{J \subseteq I \subseteq [n]} (-1)^{|I \setminus J|} \sum_{\substack{\mathbf{x}^I \in \hat{S}^I \\ \mathbf{x}^I_J = \hat{\mathbf{z}}^J}} f(\mathbf{x}^I) \geq 0 \qquad \forall\, \hat{\mathbf{z}}^J \in \hat{S}^J, \forall\, J \subseteq [n], \forall n \in \mathbb{N}. \tag{31}$$

Proof If there exist p.i. random variables $\mathbf{X} = \{X_i\}_{i \in \mathbb{N}}$ satisfying (30), then there is a probability space $(\Omega, \mathscr{A}, P)$ on which the X_i's are defined. In particular, as P is a probability, by inclusion-exclusion we have that:

$$\sum_{J \subseteq I \subseteq [n]} (-1)^{|I \setminus J|} \sum_{\substack{\mathbf{x}^I \in \hat{S}^I \\ \mathbf{x}^I_J = \hat{\mathbf{z}}^J}} f(\mathbf{x}^I) = \sum_{J \subseteq I \subseteq [n]} (-1)^{|I \setminus J|} \sum_{\substack{\mathbf{x}^I \in \hat{S}^I \\ \mathbf{x}^I_J = \hat{\mathbf{z}}^J}} P(\mathbf{X}^I = \mathbf{x}^I) =$$

$$= P(\mathbf{X}^J = \hat{\mathbf{z}}^J, \mathbf{X}^{[n] \setminus J} = \mathbf{0}) \geq 0$$

as in the proof of Theorem 1.

To show the reversed implication it is, once again, sufficient to find a probability space $(\bar{\Omega}, \bar{\mathscr{A}}, \bar{P})$ and random variables $\bar{\mathbf{X}} = \{\bar{X}_i\}_{i\in\mathbb{N}}$ satisfying (30). Let $\bar{\Omega} = S^{\mathbb{N}}$ and $\bar{X}_i \colon \bar{\Omega} \to S_i$ with $\bar{X}_i(\omega) = \omega_i$. Consider the σ-algebra $\bar{\mathscr{A}}$ generated by cylinders. For $n \in \mathbb{N}$ let

$$\bar{\Omega}_n = S^{[n]}, \qquad \Omega_n = \Omega \cap \overset{*}{\bigcup_{J\subseteq[n]}} \hat{S}^J, \qquad \Omega'_n = \Omega' \cap \overset{*}{\bigcup_{J\subseteq[n]}} \hat{S}^J, \qquad f_n = f\,|_{\Omega_n\cup\Omega'_n}.$$

Note that $\Omega_{n+1} \supseteq \Omega_n$, $\Omega'_{n+1} \supseteq \Omega'_n$ so that $f_{n+1}(\mathbf{x}^J) = f_n(\mathbf{x}^J)$ for $\mathbf{x}^J \in \Omega_n \cup \Omega'_n$.

Note also that (31) is, for fixed $n \in \mathbb{N}$, equivalent to (12) in Theorem 1 with $f = f_n$. Following the proof of the theorem we find a probability $\bar{P}_n$ on $\left(\bar{\Omega}_n, \mathscr{P}(\bar{\Omega}_n)\right)$ such that $\bar{P}_n(\mathbf{x}^J) = f_n(\mathbf{x}^J)$ for every $\mathbf{x}^J \in \Omega_n \cup \Omega'_n$.

We now show that these probabilities are consistent: let $\omega \in \bar{\Omega}_n$. Then we have

$$\bar{P}_{n+1}(\omega \times S_{n+1}) = \sum_{x^{n+1}\in\hat{S}_{n+1}} \bar{P}_{n+1}\big((\omega, x^{n+1})\big) + \bar{P}_{n+1}\big((\omega, 0^{n+1})\big) \tag{32}$$

where we denote by $(\omega, 0^{n+1})$ the vector in S^{n+1} which equals to ω in the first n components and whose $(n+1)-$th component equals 0. Let $0_{\mathbf{x}^J} = \{j \in J : \mathbf{x}^J_j = 0\}$ (note that this is different from 0^{n+1} used before) and $1_{\mathbf{x}^J} = \{j \in J : \mathbf{x}^J_j \neq 0\}$ for every $\mathbf{x}^J \in S$; we have

$$0_{(\omega,x^{n+1})} = 0_\omega, \qquad 1_{(\omega,x^{n+1})} = 1_\omega \cup \{n+1\} \tag{33}$$

$$1_{(\omega,0^{n+1})} = 1_\omega, \qquad 0_{(\omega,0^{n+1})} = 0_\omega \cup \{n+1\}. \tag{34}$$

It follows from (32) that

$$\bar{P}_{n+1}(\omega \times S_{n+1}) = \sum_{x^{n+1}\in\hat{S}_{n+1}} \sum_{I\subseteq 0_\omega} (-1)^{|I|} \sum_{\mathbf{x}^I\in\hat{S}^I} f_{n+1}(\omega_{1_\omega}, x^{n+1}, \mathbf{x}^I) + \\ + \sum_{I\subseteq 0_\omega\cup\{n+1\}} (-1)^{|I|} \sum_{\mathbf{x}^I\in\hat{S}^I} f_{n+1}(\omega_{1_\omega}, \mathbf{x}^I) = \tag{35}$$

$$= \sum_{x^{n+1}\in\hat{S}_{n+1}} \sum_{I\subseteq 0_\omega} (-1)^{|I|} \sum_{\mathbf{x}^I\in\hat{S}^I} f_{n+1}(\omega_{1_\omega}, x^{n+1}, \mathbf{x}^I) + \\ + \sum_{I\subseteq 0_\omega} (-1)^{|I|} \sum_{\mathbf{x}^I\in\hat{S}^I} f_{n+1}(\omega_{1_\omega}, \mathbf{x}^I) + \\ + \sum_{I\subseteq 0_\omega} (-1)^{|I|+1} \sum_{\mathbf{x}^I\in\hat{S}^I} \sum_{x^{n+1}\in\hat{S}_{n+1}} f_{n+1}(\omega_{1_\omega}, x^{n+1}, \mathbf{x}^I) =$$

$$= \sum_{I \subseteq 0_\omega} (-1)^{|I|} \sum_{\mathbf{x}^I \in \hat{S}^I} f_{n+1}(\omega_{1_\omega}, \mathbf{x}^I) =$$
$$= \sum_{I \subseteq 0_\omega} (-1)^{|I|} \sum_{\mathbf{x}^I \in \hat{S}^I} f_n(\omega_{1_\omega}, \mathbf{x}^I) = \bar{P}_n(\omega)$$

where the penultimate equality holds as $(\omega_{1_\omega}, \mathbf{x}^I)$ has at most n components. By Kolmogorov extention theorem (see [7] for instance), there is a probability $\bar{P}$ on $\left(\bar{\Omega}, \bar{\mathscr{A}}\right)$ such that for every $n \in \mathbb{N}$ and for every $\omega_n \in \bar{\Omega}_n$:

$$\bar{P}(\mathbf{X}^n = \omega_n) = \bar{P}(\omega_n) = \bar{P}_n(\omega_n).$$

In particular, let $\mathbf{x}^J \in \Omega \cup \Omega'$. As J is a finite subset of $\mathbb{N}$, there exists $n \in \mathbb{N}$ such that $J \subseteq [n]$ and $\mathbf{x}^J \in \Omega_n \cup \Omega'_n$. Therefore,

$$\bar{P}(\mathbf{X}^J = \mathbf{x}^J) = \bar{P}(\mathbf{x}^J) = \bar{P}_n(\mathbf{x}^J) = f_n(\mathbf{x}^J) = f(\mathbf{x}^J)$$

as required. □

6 Dutch Books

We develop now a dual theory for the case in which no collection of random variables satisfies the requirements imposed by a $\pi-$family and an independent function. We follow [10], repeating several details for self-containedness.

To do this, it is convenient to introduce new real variables $z_{\mathbf{x}}$, indexed by $\mathbf{x} \in S$, with S defined as in (2), in such a way that $z_{\mathbf{x}^J}$ is meant to replace the probability of $P(\mathbf{X}^J = \mathbf{x}^J)$ in the various formulas. Notice that with this replacement, the last two sets of conditions in (4) become the following system of polynomial equations

$$\begin{cases} z_{\mathbf{x}^J} = \pi_{\mathbf{x}^J} & \forall\, \mathbf{x}^J \in \Omega' \\ z_{\mathbf{x}^J} = z_{\mathbf{x}^J_{I_i}} z_{\mathbf{x}^J_{J \setminus I_i}} & \forall \mathbf{x}^J \in \Omega, \forall i = 1, \ldots, m_J \end{cases} \tag{36}$$

in the variables $\{z_{\mathbf{x}}\}_{\mathbf{x} \in S}$. As noticed in [10], existence of solutions of this system is only a necessary condition for existence of p.i. random variables satisfying (4), as the real solutions might be negative, for instance. It is possible, however, to make a change of variables which allows to write necessary and sufficient conditions in terms of polynomial equation and inequalities.

Recall that X_j takes values in S_j for $j = 1, \ldots, n$ and that $S^{[n]} = \prod_{j=1}^n S_j$. We will consider new variables $\{w_{\mathbf{s}}\}_{\mathbf{s} \in S^{[n]}}$ which are meant to represent $P(\mathbf{X} = \mathbf{s})$. It is shown in [10] that

Lemma 1 *Consider a system of relations of the form*

$$g_i\left(P(\mathbf{X}^{J_1} = \mathbf{s}^{J_1}), \ldots, P(\mathbf{X}^{J_k} = \mathbf{s}^{J_k})\right) \triangleright_i 0 \tag{37}$$

for $i \in \mathbb{I}$, where $\mathbb{I}$ is some set of indices, $k = |S|$ and $\triangleright_i$ stands for one of $\{=, \geq, >\}$ and express it in variables using z_ℓ to replace $P(\mathbf{X}^{J_\ell} = \mathbf{s}^{J_\ell})$, to get the polynomial system of equations and inequalities

$$g_i(z_1, \ldots, z_k) \triangleright_i 0 \quad i \in \mathbb{I}. \tag{38}$$

Then consider the following change of variables: for every $J \subseteq [n]$ express $\{\mathbf{X}^J = \mathbf{s}^J\} = \bigcup_{\mathbf{s} \in \Gamma_{\mathbf{s}^J}} \{\mathbf{X} = \mathbf{s}\}$ for a suitable $\Gamma_{\mathbf{s}^J} \subseteq S^{[n]}$. The change of variables is

$$z_\ell(\mathbf{w}) = \sum_{\mathbf{s} \in \Gamma_{\mathbf{s}^{J_\ell}}} w_{\mathbf{s}}$$

in the new variables $\mathbf{w} = \{w_{\mathbf{s}}\}_{\mathbf{s} \in S^{[n]}}$. Then there are random variables $\mathbf{X}$ satisfying (37) if and only if the system formed by the equations:

$$\begin{cases} g_i(z_1(\mathbf{w}), \ldots, z_k(\mathbf{w})) \triangleright_i 0 & \forall\, i \in I \\ w_{\mathbf{s}} \geq 0 & \forall\, \mathbf{s} \in S^{[n]} \\ \sum_{\mathbf{s} \in S^{[n]}} w_{\mathbf{s}} = 1 \end{cases} \tag{39}$$

admits solution in the $\mathbf{w}$ variables.

As these are polynomial relations, existence of a solution is decidable by Tarski-Seidenberg and Sturm's Theorem [3]. As (36) is a special case of (37), this is a more general, and algorithmically more involved, proof of decidability of the existence problem for p.i. random variables.

The above translation of the existence problem into a question of semi-algebraic geometry allows to use results from this field. In particular, the Positivstellensatz (see [3]) asserts the following: rewrite the system (39) as

$$\begin{cases} f_j(\mathbf{w}) = 0 & j = 1, \ldots, l \\ g_r(\mathbf{w}) \geq 0 & r = 1, \ldots, t \\ h_i(\mathbf{w}) \neq 0 & i = 1, \ldots, s \end{cases} \tag{40}$$

for some ℓ, t and s depending on the value of $\triangleright_i$. Note that an equation of the form $f(\mathbf{x}) > 0$ can be rewritten as the system

$$\begin{cases} f(\mathbf{x}) \geq 0 \\ f(\mathbf{x}) \neq 0 \end{cases}.$$

so that every system (39) can be rewritten as (40). Next, let I be the ideal generated by the family $\{f_j\}_{j=1,\dots,l}$, C the positive cone generated by the family $\{g_r\}_{r=1,\dots,t}$ and M the multiplicative monoid generated by the family $\{h_i\}_{i=1,\dots,s}$. Then there is no solution to (40), and hence to (39), if and only if there are $f \in I$, $g \in C$ and $h \in M$ such that $f + g + h^2 \equiv 0$. In particular, if there are no strict inequality, i.e. $\{h_i\}$ is empty, then (40) has no solutions if and only if there exist $f \in I$ and $g \in C$ such that $f + g \equiv -1$.

It is proven in [10], based on the Positivstellensatz, that when there are no random variables fulfilling (37) then it is possible to define a Dutch Book, i.e. a rigging strategy in which a bookmaker can insure a strictly positive gain.

Definition 4 Given n events A_i, $i = 1, \dots, n$ on some probability space $(\overline{\Omega}, \overline{\mathscr{A}}, \overline{P})$, and relations, namely equalities or inequalities, among the probabilities of boolean combinations of the A_i's, a Dutch Book is a $\sigma(A_1, \dots, A_n)$-measurable random variable G such that the mean value $E(G) \geq 0$ if all the relations hold, but $G(\omega) = -1$ for all $\omega \in \overline{\Omega}$.

In our case, the bookmaker offers a believer of (37) a random game in which the believer computes that (s)he has a nonnegative average gain, while instead losing a fixed amount in every single round. Adapting it from [10], we have

Theorem 5 ([10]) *Given a family of requirements of the form* (38) *with no strict inequalities, there are no random variables satisfying all the equations if and only if, assuming that it is possible to realize a finite but sufficiently large number of independent copies of the collection of random variables, it is possible to realize a Dutch Book against any believer of* (38).

Some care must be used in interpreting the content of this theorem. When talking about a believer of (38) we intend that (s)he has determined some events that (s)he believes satisfy (38). One of the assumptions in the theorem is that it is possible to find or produce a finite, but sufficiently large, number of copies of such collection of random variables in such a way that the believer of (38) also thinks that the copy are independent. The need for more copies comes from the nonlinearity of the polynomials in (40).

We now determine the Dutch Book for the p.i. random variables, showing that, although the relations in (36) are not linear, in this case we do not need the assumption about additional copies as the results of the previous sections imply that the Dutch Book actually requires one single copy only of the random variables.

Theorem 6 *If there are no random variables satisfying* (10) *then it is possible to realize a Dutch Book against any believer of* (10).

Proof Rewrite (10) in the variables $\mathbf{z} = \{z_{\mathbf{x}}\}_{\mathbf{x}\in S}$ as in (36) with the $\pi-$family satisfying (8). If (8) fails, it means that the believer of (10) believes also in the equality of two unequal numbers: a not-random Dutch Book can then be produced by offering the smaller amount, augmented by a fraction of the difference, in exchange for the larger amount.

Changing variables into $\mathbf{w}$, the system (39) becomes

$$\begin{cases} z_{\mathbf{x}^J}(\mathbf{w}) = \pi_{\mathbf{x}^J} & \forall\, \mathbf{x}^J \in \Omega' \\ z_{\mathbf{x}^J}(\mathbf{w}) = z_{\mathbf{x}^J_{I_i}}(\mathbf{w}) z_{\mathbf{x}^J_{J\setminus I_i}}(\mathbf{w}) & \forall \mathbf{x}^J \in \Omega, \forall i = 1, \dots, m_J \\ w_{\mathbf{s}} \geq 0 & \forall\, \mathbf{s} \in S^{[n]} \\ \sum\limits_{\mathbf{s}\in S^{[n]}} w_{\mathbf{s}} = 1. \end{cases} \tag{41}$$

By Lemma (1), there exist random variables satisfying (10) if and only if there is a solution to (41), so the believer of (10) also believes that there is a solution to (41). Introducing the function f as in (9), (41) becomes

$$\begin{cases} \sum\limits_{\mathbf{s}:\, \mathbf{s}_J = \mathbf{x}^J} w_{\mathbf{s}} - f(\mathbf{x}^J) = 0 \quad \forall\, J \subseteq [n], \quad \forall\, \mathbf{x}^J \in \hat{S}^J \\ \sum\limits_{\mathbf{s}\in S^{[n]}} w_{\mathbf{s}} - 1 = 0 \\ w_{\mathbf{s}} \geq 0 \qquad\qquad\qquad\qquad \forall\, \mathbf{s} \in S^{[n]} \end{cases} \tag{42}$$

as $z_{\mathbf{x}^J}(\mathbf{w}) = \sum\limits_{\mathbf{s}:\, \mathbf{s}_J = \mathbf{x}^J} w_{\mathbf{s}}$. The believer of (10) also believes that (42) has a solution.

By Theorem 1, if there are no random variables satisfying (10) then there exist $\bar{J} \subseteq [n]$ and $\mathbf{z}^{\bar{J}} \in \hat{S}^{\bar{J}}$ such that

$$\sum_{I:\bar{J}\subseteq I\subseteq[n]} (-1)^{|I\setminus\bar{J}|} \sum_{\substack{\mathbf{x}^I\in\hat{S}^I \\ \mathbf{x}^I_{\bar{J}} = \mathbf{z}^{\bar{J}}}} f(\mathbf{x}^I) = -D < 0 \tag{43}$$

with D a positive constant.

Consider then the following linear combination of the equations in (42):

$$\begin{aligned} &\sum_{I:\bar{J}\subseteq I\subseteq[n]} (-1)^{|I\setminus\bar{J}|} \sum_{\substack{\mathbf{x}^I\in\hat{S}^I \\ \mathbf{x}^I_{\bar{J}} = \mathbf{z}^{\bar{J}}}} \left(f(\mathbf{x}^I) - \sum_{\mathbf{s}:\, \mathbf{s}_I = \mathbf{x}^I} w_{\mathbf{s}} \right) = \\ &= -D - \sum_{I:\bar{J}\subseteq I\subseteq[n]} (-1)^{|I\setminus\bar{J}|} \sum_{\substack{\mathbf{x}^I\in\hat{S}^I \\ \mathbf{x}^I_{\bar{J}} = \mathbf{z}^{\bar{J}}}} \left(\sum_{\mathbf{s}:\, \mathbf{s}_I = \mathbf{x}^I} w_{\mathbf{s}} \right) = \end{aligned} \tag{44}$$

$$= -D - \sum_{I \subseteq \bar{J}^c} (-1)^{|I|} \sum_{\mathbf{x}^I \in \hat{S}^I} \left(\sum_{\substack{\mathbf{s}: \mathbf{s}_{\bar{J}} = \mathbf{z}^{\bar{J}} \\ \mathbf{s}_I = \mathbf{x}^I}} w_{\mathbf{s}} \right).$$

This polynomial belongs to the ideal generated by the equations in (42), as can be seen in the first line. However, the coefficient of each $w_{\mathbf{s}}$ is zero, as it equals

$$\sum_{I \subseteq \bar{J}^c} (-1)^{|I|} = 0.$$

We have then found the relation

$$\frac{1}{D} \left\{ \sum_{\bar{J} \subseteq I \subseteq [n]} (-1)^{|I \setminus \bar{J}|} \sum_{\substack{\mathbf{x}^I \in \hat{S}^I \\ \mathbf{x}^I_{\bar{J}} = \mathbf{z}^{\bar{J}}}} \left(f(\mathbf{x}^I) - \sum_{\mathbf{s}: \mathbf{s}_I = \mathbf{x}^I} w_{\mathbf{s}} \right) \right\} \equiv -1. \tag{45}$$

In order to generate the Dutch Book substitute $w_{\mathbf{s}}$ in (45) by the indicator function $\mathbb{I}_{\mathbf{s}}$ that $\mathbf{X} = \mathbf{s}$. Then consider the game

$$G = \frac{1}{D} \left\{ \sum_{I: \bar{J} \subseteq I \subseteq [n]} (-1)^{|I \setminus \bar{J}|} \sum_{\substack{\mathbf{x}^I \in \hat{S}^I \\ \mathbf{x}^I_{\bar{J}} = \mathbf{z}^{\bar{J}}}} \left(f(\mathbf{x}^I) - \sum_{\mathbf{s}: \mathbf{s}_I = \mathbf{x}^I} \mathbb{I}_{\mathbf{s}} \right) \right\} \tag{46}$$

As $E(\mathbb{I}_{\mathbf{s}}) = P(\mathbf{X} = \mathbf{s})$, we have

$$E(G) = E\left(\frac{1}{D} \left\{ \sum_{I: \bar{J} \subseteq I \subseteq [n]} (-1)^{|I \setminus \bar{J}|} \sum_{\substack{\mathbf{x}^I \in \hat{S}^I \\ \mathbf{x}^I_{\bar{J}} = \mathbf{z}^{\bar{J}}}} \left(f(\mathbf{x}^I) - \sum_{\mathbf{s}: \mathbf{s}_I = \mathbf{x}^I} \mathbb{I}_{\mathbf{s}} \right) \right\} \right)$$

$$= \frac{1}{D} \left\{ \sum_{I: \bar{J} \subseteq I \subseteq [n]} (-1)^{|I \setminus \bar{J}|} \sum_{\substack{\mathbf{x}^I \in \hat{S}^I \\ \mathbf{x}^I_{\bar{J}} = \mathbf{z}^{\bar{J}}}} E\left(f(\mathbf{x}^I) - \sum_{\mathbf{s}: \mathbf{s}_I = \mathbf{x}^I} \mathbb{I}_{\mathbf{s}} \right) \right\}$$

$$= \frac{1}{D} \left\{ \sum_{\bar{J} \subseteq I \subseteq [n]} (-1)^{|I \setminus \bar{J}|} \sum_{\substack{\mathbf{x}^I \in \hat{S}^I \\ \mathbf{x}^I_{\bar{J}} = \mathbf{z}^{\bar{J}}}} \left(f(\mathbf{x}^I) - \sum_{\mathbf{s}: \mathbf{s}_I = \mathbf{x}^I} w_{\mathbf{s}} \right) \right\} \geq 0$$

if one believes that the relations in (42) can all hold. On the other hand, (45) implies that every single game results in a constant negative loss for the player, so G is the required Dutch Book.

Notice that such a Dutch Book is linear in the $\mathbb{I}$'s, and thus requires only one copy of such variables. □

References

1. Aaronson, J., Gilat, D., Keane, M.S.: On the structure of 1-dependent Markov chains. J. Theor. Probab. **5**, 545–561 (1992)
2. Alon, N., Spencer, J.H.: The Probabilistic Method, 2nd edn. Wiley-Interscience (2000)
3. Bochnak, J., Coste, M., Roy, M.F.: Real Algebraic Geometry, vol. 36. Springer (1998)
4. Benjamini, I., Gurel-Gurevich, O., Peled, R.: On K-wise independent distributions and Boolean functions. Preprint. arXiv:1201.3261 [maTheor.PR] (2012)
5. Brightwell, G., Leader, I., Scott, A., Thomason, A.: Combinatorics and Probability. Cambridge University Press (2007)
6. Chellappa, R., Jain, A.: Markov Random Fields: Theory and Applications. Academic Press, Boston (1993)
7. Durrett, R.: Probability: Theory and Examples, 2nd edn. Cambridge University Press (2010)
8. Erdős, P., Lovász, L.: Problems and results on 3-chromatic hypergraphs and some related questions. Colloq. Math. Soc. János Bolyai **10**, 609–627 (1974)
9. De Finetti, B.: Probability, Induction and Statistics. Wiley, New York (1972)
10. Gandolfi, A.: Probability and Hilbert's VI Problem. Preprint
11. Geman, S., Graffigne, C.: Markov random field image models and their applications to computer vision. Proc. Intern. Congr. Math. 1496–1517 (1986)
12. Gandolfi, A., Keane, M., de Valk, V.: Extremal two-correlations of two-valued stationary one-dependent processes. Probab. Theory Relat. Fields **80**, 475-480, Springer-Verlag (1989)
13. Luby, M., Wigderson, A.: Pairwise independence and derandomization. Technical Report TR-95-035. International Computer Science Institute, Berkeley, California (1995)
14. Ruelle, D.: Thermodynamic Formalism: The Mathematical Structure of Equilibrium Statistical Mechanics, 2nd edn. Cambridge University Press (2004)
15. Ross, S.: An Elementary Introduction to Mathematical Finance. Cambridge University Press (2011)
16. Shearer, J.B.: On a problem of Spencer. Combinatorica **5**, 241–245. Springer (1985)
17. Scott, A.D., Sokal, A.D.: The repulsive lattice gas, the independent-set polynomial, and the Lovász local lemma. J. Stat. Phys. **118**, 1151–1261 (2005)
18. Temmel, C.: Shearer's measure and stochastic domination of product measures. J. Theor. Probab. **27**, 22–40 (2014)
19. Vineberg, S.: Dutch Book arguments. The stanford encyclopedia of philosophy. http://plato.stanford.edu/entries/dutch-book/ (2016)

Sobol Sensitivity: A Strategy for Feature Selection

Dmitry Efimov and Hana Sulieman

Abstract In this paper we propose a novel approach for feature selection in machine learning. The approach is based on the Sobol sensitivity analysis, a variance-based technique that determines the contribution of each feature and their interactions to the overall variance of the target variable. Similar to wrappers, Sobol sensitivity is a model-based approach that utilizes the trained model to evaluate feature importances. It uses the full feature set to train the model just as embedded methods do. Based on the trained model, it evaluates importance scores and, similar to filters, identifies the subset of important features with highest scores without retraining the model. The distinctive characteristic of the Sobol sensitivity approach is its computational efficiency compared to the existing feature selection algorithms. This is because importance scores for all individual features and subsets of features are calculated with the same trained model. We apply the proposed algorithm to a simulated data set and to four benchmark data sets used in machine learning literature. The results are compared to those obtained by two of the widely used feature selection algorithms and some computational aspects are also discussed.

Keywords Feature selection · Sobol index · Sensitivity analysis · Machine learning

Mathematics Subject Classification (2010): Primary 62P07 · Secondary 68U04

1 Introduction

The problem of variable (feature) selection in predictive modelling has received considerable attention during the past 10 years in both statistics and machine learning

D. Efimov · H. Sulieman (✉)
Department of Mathematics and Statistics, American University of Sharjah,
P.O. Box 26666, Sharjah, UAE
e-mail: hsulieman@aus.edu

D. Efimov
e-mail: defimov@aus.edu

T. Abualrub et al. (eds.), *Mathematics Across Contemporary Sciences*,
Springer Proceedings in Mathematics & Statistics 190,
DOI 10.1007/978-3-319-46310-0_4

literatures. The aim of feature selection is to identify the subset of predictor variables that provides a reliable and robust model for a given target variable. Feature selection plays a central role in many areas such as natural language processing, gene expression array studies, computational biology, image recognition, information retrieval, temporal modelling, consumer profile analysis and business data analytics. Curse of dimensionality in data collected in these and other areas and the increased level of noise in the associated features have motivated the development of various feature selection techniques. Feature selection is a key mechanism to reduce a large number of variables to relatively few.

In this article, a new approach for feature selection is proposed. The new approach is inspired by the popular Sobol sensitivity measure developed by I.M. Sobol in 1990 [18]. Sobol sensitivity measure is a variance-based sensitivity technique that decomposes the output (target) variable variance into summands of variances of the input variables (features) in an increasing dimensionality. It has been widely used in assessing global sensitivity of models in different fields such as environment, economics, engineering and many others. The approach is a model-based technique that utilizes the fitted model to compute the partial variances or variance contributions by each feature and their interactions to the overall variance of the target variable. It is shown to have lower computational cost than many existing feature selection algorithms but its effectiveness depends on the quality of the fitted model as it is the case for some popular algorithms.

The article is organized as follows. Section 2 provides a review of some existing methods on variable selection. Section 3 proposes Sobol sensitivity approach for variable selection and gives some theoretical foundation of the measure. It also discusses some computational aspects of the method. Section 4 applies the proposed Sobol sensitivity to several data sets used in machine learning literature and provides comparisons with some existing benchmark algorithms.

2 Literature Review

In this section we summarize some popular feature selection techniques, give some computational consideration and provide the motivation for the proposed new technique.

Feature selection techniques can be divided into three general frameworks: *wrappers*, *filters* and *embedded methods* [21, 25]. Wrappers evaluate the predictive power of subsets of features by retraining the model for different feature subsets. Filters evaluate the importance of features before the main prediction algorithm is trained. Embedded methods search for the optimal subset of features simultaneously with minimizing a loss function. What follows is a brief description of each framework:

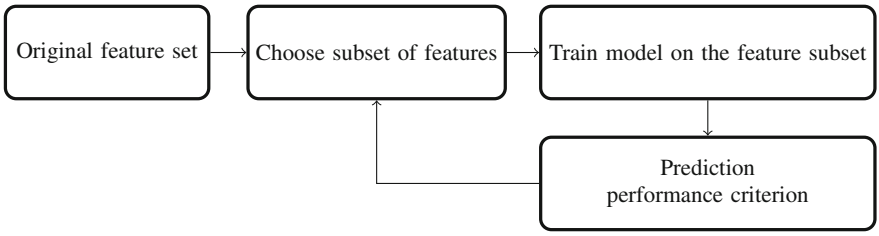


Fig. 1 Wrappers framework

1. Wrappers.
   Wrappers are model-based methods for feature selection and are considered to be the most effective and computationally intractable algorithms. Figure 1 shows the main principle of the wrapper methods' framework.
   Basically, to find the most relevant and informative subset of features, the prescribed model is trained for different subsets of features. The subset with the highest score on a particular prediction performance criterion is selected as best set of features. Because wrapper methods utilize the model algorithm they are considered more effective and hence more desired than filters and embedded methods.
   However, wrapper methods are generally criticized for their potential to overfit the training data and for their computational cost. Overfitting occurs when the complete data set is used for the training and the model becomes excessively complex to fit the data too precisely but still provides poor predictions when applied to new data set. This occurs when there is insufficient data to train and the data does not fully recover the concept learned. Several approaches have been proposed in the literature to overcome model overfitting:

   a. Cross-Validation (CV): in this approach the data set is split into training and validation sets. The model is trained on the training set and predictions are obtained on the validation set. A variety of techniques are developed to determine the fraction of data that should be used for training and that used for validation. These techniques include random sub-sampling, leave-one-out, K-fold and other CV sampling mechanisms.
   b. Probabilistic approach: based on information theory principles. The prediction accuracy of the algorithm is measured using various techniques such as Akaike Information Criteria (AIC, [3, 4]), Bayesian Information Criteria (BIC, [5]) and others.

   As for computational cost, wrappers are deemed computationally expensive. For n features, the number of feature subsets is defined by $O(2^n)$, i.e., the computational needs of wrappers exponentially increase with the number of features in the model. This makes the search for all possible subsets of features impractical for even moderate value of n. The computational cost of wrapper methods can be reduced by using efficient search strategies to find the optimal subset of features.

One of the earliest attempt to improve the computational efficiency is due to Hocking and Leslie in [1]. Their method starts by fitting the full model and then features are eliminated based on the magnitudes of their t-statistics. The efficiency gain of the method lies in the fact that entire subsets of features are eliminated from further consideration when their reduced prediction error is greater than other subsets already evaluated. The method can assume independent features and works well when there is a small number of important features that dominate the target variable and can easily be identified.
Sequential search methods such as forward selection, backward elimination and stepwise regression became popular techniques used to overcome some of the computational demands of wrappers. Forward selection begins with no variables and progressively adds features until maximum reduction in prediction error is reached. The reverse of this strategy is the backward elimination which begins with full model and progressively removes features having smallest contributions. Once a feature is added in forward selection or eliminated in backward elimination, the operation can not be reversed. To overcome this drawback, stepwise selection is used. Stepwise selection starts by adding features until reaching some stopping criteria. Then the algorithm starts dropping features until reaching another stopping criteria and so on. While stepwise selection can reduce the computational cost of the best set of features it does not, however, guarantee the selection of the global optimal set.

2. Filters.
Filters evaluate feature importance as a pre-processing operation to model training as depicted in Fig. 2. The main difference between filters and wrappers is that filters do not use the training procedure to capture the relationship between features. Rather, they use some information metric to calculate feature ranking from the data without direct input from the target. Popular information metrics include t-statistic, p-value, Pearson correlation coefficient, mutual information and other correlation measures. Computationally, filters are more efficient than wrappers as they require only the computation of n scores for n features. They are also more robust against overfitting than wrappers.
By using Pearson correlation, filters can only capture linear effects between features and target variable. The nonlinear effects are left undiscovered. A successful attempt to deal with nonlinear effects has been recently developed. Aliferis et al. [24] described Markov blanket technique that is based on Bayesian network. A Markov blanket of the target variable Y is defined as a minimal set of features on which all other features are conditioned so as they become independent of Y.

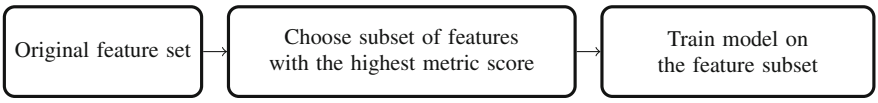


Fig. 2 Filters framework

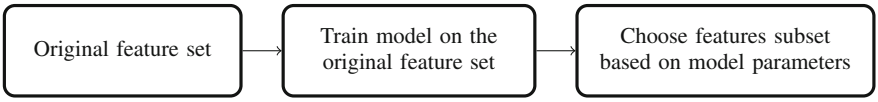


Fig. 3 Embedded methods framework

In terms of relevancy of selected features, Markov blanket is shown by Tsamardinos and Aliferis (2003) to provide the most relevant and optimal features in calibrated classifications in which the probability distribution of the target variable can be perfectly estimated with the smallest number of variables. The authors showed that neither filters nor wrappers are superior to one another in identifying the optimal features because filters lack universal optimality,i.e., independently of the classification algorithm and model-performance metric, and wrappers lack universal efficiency. Markov blanket technique does not suffer from these shortcomings.

3. Embedded methods.
   Embedded methods use training procedure to obtain feature rankings Fig. 3. The aim of embedded methods is two-fold: first, maximizing the prediction accuracy and second, minimizing the number of features in the predictive algorithm.
   Regularization methods such as Ridge Regression [2], Nonnegative Garrote [6], Least Absolute Selection and Shrinkage Operator (LASSO, [8]) are the most common forms of embedded methods. In these methods, the coefficients (weights) of the features are penalized by some regularization terms or forced to be exactly zero. Features with weights close to zero are then eliminated without compromising the prediction performance of the model. Analogy to filters, several developments have been achieved in embedded regularization methods. LASSO [15] and Elastic Net [11] are examples of methods that were developed to measure the importance of subsets of features. Boosted LASSO [13] and Smoothly Clipped Absolute Deviation (SCAD) [14] are examples of methods that use nonlinear regularization terms to produce more sparse and unbiased estimators of coefficients. Other embedded methods are based on decision tree algorithms. In this group of embedded methods, various decision trees are iteratively built using bootstrapping and for each tree, information gain (based on specific information entropy) is calculated for each feature. Features are then ranked based on the average information gain over all trees. Random Forest algorithm [7] is one popular example of decision tree methods. Louppe et al. in [17] provided comparative analysis of feature importance using various decision tree algorithms. Embedded methods can be disadvantaged for the fact that feature weights are often estimated iteratively not explicitly.

The reader is referred to the excellent reviews of feature selection methods found in [22, 23].

3 Sobol Sensitivity Approach

In this section we propose a new technique for feature selection in machine learning and provide some mathematical foundation of the algorithm. The proposed technique is based on variance decomposition principle of model output developed by Sobol (1990) [18] (in Russian) and Sobol (1993) [19] (in English). Sobol sensitivity analysis is intended to determine how much of the variability in model output is dependent upon each of the input variables, either upon a single variable or upon an interaction between different variables. The decomposition of the output variance in a Sobol sensitivity analysis employs the same principal as the classical analysis of variance in factorial experimental designs.

3.1 Theoretical Background

Let the function $f(\mathbf{x})$, where $\mathbf{x} = (x_1, ..., x_n)$ be defined on the unit n-dimensional cube

$$K^n = \{\mathbf{x} \mid 0 \leqslant x_i \leqslant 1, i = 1, ..., n\}.$$

Sobol's main idea is to decompose the function $f(\mathbf{x})$ into summands of increasing dimensionality, namely

$$f(\mathbf{x}) = f_0 + \sum_{i=1}^{n} f_i(x_i) + \sum_{1 \leq i < j \leq n} f_{ij}(x_i, x_j) + ... + f_{12...n}(x_1, ..., x_n). \tag{1}$$

The decomposition in (1) holds true if f_0 is a constant and the integral of every summand over any of its variables is zero, i.e.

$$\int_0^1 f_{i_1...i_s}(x_{i_1}, ..., x_{i_s}) dx_{i_k} = 0, 1 \leqslant i_1 < .... < i_s \leq n, 1 \leqslant k \leqslant s, s = 1, 2, ...n.$$

For independent $x_1, ..., x_n$, all terms in Eq. (1) are orthogonal and f_0 can be calculated as:

$$f_0 = \int_{K^n} f(\mathbf{x}) d\mathbf{x} \tag{2}$$

which represents the average value (or expectation) of the function f. Sobol (1993) [19] showed that decomposition (1) is unique and all of its terms can easily be evaluated through multi-variable integrals.

Because of the orthogonality of the $\mathbf{x}$-space, the total variance D of $f(\mathbf{x})$ can also be partitioned in the same way as the original function, i.e.,

$$D = \sum_{i=1}^{n} D_i + \sum_{1 \le i < j \le n} D_{ij} + ... + D_{12...n} \tag{3}$$

where

$$D = \int_{K^n} f^2(\mathbf{x}) d\mathbf{x} - f_0^2$$

and

$$D_{i_1...i_s} = \int_0^1 ... \int_0^1 f^2_{i_1...i_s}(x_{i_1}, ..., x_{i_s}) dx_{i_1} ... dx_{i_s} \qquad \begin{array}{l} 1 \leqslant i_1 < .... < i_s \leq n, \\ s = 1, 2, ...n \end{array}$$

$D_{i_1...i_s}$ is the partial variance attributed to $x_{i_1}, ..., x_{i_s}$ defined by the variance of the conditional expectation of $f(\mathbf{x})$ conditioned on $x_{i_1}, ..., x_{i_s}$, namely,

$$D_{i_1...i_s} = Var[E(f|x_{i_1}, ..., x_{i_s})]$$

where the conditional expectation is taken over all x_j not in $\{i_1, ..., i_s\}$ and variance is computed over the range of possible values of $x_{i_1}, ..., x_{i_s}$.
The usefulness of $D_{i_1...i_s}$ as a measure of sensitivity is easy to grasp. Influential $x_{i_1}, ..., x_{i_s}$ control f significantly and so $E(f|x_{i_1}, ..., x_{i_s})$ will mimic f. In this case the total variance in f will be matched by the variability in $E(f|x_{i_1}, ..., x_{i_s})$ as $x_{i_1}, ..., x_{i_s}$ vary making $D_{i_1...i_s}$ large compared to the total variance D.

Sobol in [20] proposed the following indices to measure sensitivity of the function with respect to $x_{i_1}, ..., x_{i_s}$:

$$S_{i_1...i_s} = \frac{D_{i_1...i_s}}{D}, 1 \leqslant i_1 < .... < i_s \leqslant n, s = 1, 2, ...n \tag{4}$$

with $\sum S_{i_1...i_s} = 1$.

For $s = 1$, the sensitivity measure $S_{i_1} = S_i$ is called *first-order sensitivity index* which measures the fractional contribution of the individual variable x_i to the total variance of f. For $s = 2$, S_{ij} is called the *second-order sensitivity index* which measures the portion of the variability in f due to the interaction of x_i and x_j and so on. Total sensitivity index, defined as the sum of all sensitivity indices involving x_i up to the n-th order, i.e.,

$$TS_i = S_i + \sum_{j:j \neq i}^{n} S_{ij} + ... + S_{1....i...n} \tag{5}$$

was also proposed to quantify the overall effect of x_i on the model output.

Decomposition (1) or (3) has long history and was given in its general form by Efron and Stein [9]. More concisely, one can think of $f(\mathbf{x})$ as some statistics defined on the independent variables $x_1, ..., x_n$, then $f(\mathbf{x})$ may be decomposed into

a grand mean $f_0 = E[f(\mathbf{x})]$; i-th main effect $f_i(x_i) = E[f(\mathbf{x}|x_i)] - f_0$; ij-th interaction $f_{ij}(x_i, x_j) = E[f(\mathbf{x}|x_i, x_j)] - E[f(\mathbf{x}|x_i)] - E[f(\mathbf{x}|x_j)] + f_0$ and so on. Given these definitions, decomposition (1) follows immediately. The f_i's, f_{ij}'s,....are known in factorial experimental design as ANOVA-HDMR, where HDMR stands for High-Dimensional Model Representation.
For example, when $n = 2, f(\mathbf{x})$ can be decomposed into:

$$\begin{aligned} f(x_1, x_2) &= f_0 + f_1(x_1) + f_2(x_2) + f_{12}(x_1, x_2) = \\ &= f_0 + E[f(\mathbf{x}|x_1)] - f_0 + E[f(\mathbf{x}|x_2)] - f_0 + \\ &+ E[f(\mathbf{x}|x_1, x_2)] - E[f(\mathbf{x}|x_1)] - E[f(\mathbf{x}|x_2)] + f_0. \end{aligned}$$

The individual terms of decomposition (1) can easily be shown to have a zero mean. For example $E[f_i(x_i)] = E[E[f(\mathbf{x}|x_i)] - f_0] = E[f(\mathbf{x})] - f_0 = 0$. Decomposition (1) terms can also be shown to be mutually uncorrelated, implying that the unconditional total variance of $f(\mathbf{x})$, D, can simply be expressed as a sum of variances of these uncorrelated terms giving the variance decomposition (3) where

$$\begin{aligned} D_i &= Var(f_i(x_i)) = Var(E[f(\mathbf{x}|x_i)]) \\ D_{ij} &= Var(f_{ij}(x_i, x_j)) = Var(E[f(\mathbf{x}|x_i, x_j)]) + Var(E[f(\mathbf{x}|x_i)]) + \\ &+ Var(E[f(\mathbf{x}|x_j)]) \end{aligned}$$

and so on. It is noted that decomposition (3) is similar to the classical ANOVA decomposition without the residual error term.

If the relationship between $\mathbf{x}$ and the model output is additive linear, then a straightforward variance decomposition can be provided by regression analysis. It can be shown, in this case, that the first-order sensitivity index, S_i is equal to the squared standardized regression coefficients, i.e., $S_i = \beta_i^2$. That is, the β_i's give the fractional contribution of each predictor to the variance of the response variable. The effectiveness of β_i's as a measure of sensitivity, in this case, depends on the quality of the fitted model and the degree of linearity in the relationship between the response variable and predictors.

The definition of Sobol sensitivity index given in (3) can be extended to include group indices for subsets of variables and their joint sensitivity behaviour. Suppose the variables $x_1, ..., x_n$ are partitioned into r disjoint groups $\mathbf{x}^1, ..., \mathbf{x}^r$, $r < n$, then decomposition (1) can be expressed as:

$$f(\mathbf{x}) = f_0 + \sum_{i=1}^{r} f_i(\mathbf{x}^i) + \sum_{1 \leqslant i < j \leqslant r} f_{ij}(\mathbf{x}^i, \mathbf{x}^j) + ... + f_{1,2,.....,r}(\mathbf{x}^1, ..., \mathbf{x}^r).$$

For $r = 2$ for example, the variables $\mathbf{x}$ are partitioned into two groups $\mathbf{y}$ and $\mathbf{z}$, giving the following decomposition:

$$f(\mathbf{x}) = f_1(\mathbf{y}) + f_2(\mathbf{z}) + f_{12}(\mathbf{y}, \mathbf{z}).$$

The variances D_1 and D_2 for each of $\mathbf{y}$ and $\mathbf{z}$ are calculated as

$$D_1 = \int_0^1 \ldots \int_0^1 f_1^2(\mathbf{y}) d\mathbf{y}, \quad D_2 = \int_0^1 \ldots \int_0^1 f_2^2(\mathbf{z}) d\mathbf{z} \tag{6}$$

and

$$D = \int_0^1 \ldots \int_0^1 f^2(\mathbf{x}) d\mathbf{x} - f_0^2, \quad D_{12} = D - D_1 - D_2. \tag{7}$$

For this two-set decomposition, we define the following sensitivity index:

$$\begin{aligned} SI_{\mathbf{y}} &= \frac{1}{D}(D_1 + D_{12}) \\ SI_{\mathbf{z}} &= \frac{1}{D}(D_2 + D_{12}). \end{aligned} \tag{8}$$

In the next section, $SI_{\mathbf{y}}$ and $SI_{\mathbf{z}}$ will be referred to as Sobol Importance (SI) measure that will be used as the basis for feature selection mechanism.

In practice, variables are usually ranked based on the magnitude of their Sobol sensitivity indices, the higher the magnitude, the more influential respective variables are. Although no distinct cutoff value has been defined, the rather arbitrary value of 0.05 is frequently accepted for this type of analysis for distinguishing important from unimportant variables. It should be noted though that this value of 0.05 is primarily used for more complex models and it may be not stringent enough for relatively simple models that contain only few input variables.

3.2 Sobol Sensitivity for Machine Learning

3.2.1 General Framework

Let m be a number of samples in the dataset and n be a number of features (variables). Denote the set of feature indices as $J = \{1, \ldots, n\}$. For the purpose of feature selection in machine learning, we propose to partition the set of indices J into two subsets $J_1 = \{j_1, \ldots, j_s\}$ and $J_2 = \{j \in J \mid j \notin J_1\}$ and estimate the importance for the features from each group separately using Sobol sensitivity index given in Eq. (8). In many machine learning problem settings, splitting features into two groups is deemed sufficient for identifying important features. In the classical Sobol's sensitivity analysis, variables (features) are assumed independent and uniformly distributed over the interval [0, 1]. In our proposed analysis, we consider normally distributed features following the work of Arwade et al. (2010) [26] and continue to assume independent

features. The Monte-Carlo procedure [20] can be applied to evaluate the quantities from above (6) and (7). Assuming that X is an original design matrix we generate two new matrices Y and Z such that Y is obtained from X by random shuffling each column with index $j \in J_1$, Z is obtained by random shuffling each column with index $j \in J_2$. We denote $\mathbf{x}_i$, $\mathbf{y}_i$, $\mathbf{z}_i$ the *i-th* row of the matrices X, Y, and Z accordingly.

Based on Monte-Carlo exploration of the features' space, the quantities in (6) and (7) can be estimated as follows:

$$\begin{aligned} &f_0 = \tfrac{1}{m}\sum_{i=1}^{m} f(\mathbf{x}_i), && D + f_0^2 \approx \tfrac{1}{m}\sum_{i=1}^{m} f^2(\mathbf{x}_i), \\ &D_1 + f_0^2 \approx \tfrac{1}{m}\sum_{i=1}^{m} f(\mathbf{x}_i) f(\mathbf{z}_i), && D_2 + f_0^2 \approx \tfrac{1}{m}\sum_{i=1}^{m} f(\mathbf{x}_i) f(\mathbf{y}_i), \\ &D_{12} = D - D_1 - D_2, \end{aligned} \tag{9}$$

where summation is taken by all dataset entries.

The suggested permutation procedure of the design matrix X allows the generation of random values from the assumed distribution of features for use in the Monte-Carlo algorithm. Sobol importance scores for the feature subsets J_1 and J_2 are calculated using (8).

3.2.2 Computational Aspects

The general framework described above gives rise to the following algorithm that calculates feature importance for subsets of features. The main output of this algorithm is the Sobol importance (SI) score for a given subset of features.

Algorithm 1 Sobol importance scores for feature subset

1: Let X be an $m \times n$ design matrix for the given dataset, y be a vector of outputs.
2: Train the model M on the original dataset X, y and obtain predictions p on the dataset X
3: Evaluate $f_0 = \dfrac{1}{m}\sum_{i=1}^{m} p_i$ and $D = \dfrac{1}{m}\sum_{i=1}^{m} p_i^2 - f_0^2$
4: Define a feature subset of interest

$$J_1 = \{j_1, \ldots, j_s\}$$

and complimentary feature subset

$$J_2 = \{j \in \{1, 2, \ldots, n\} \,|\, j \notin J_1\}$$

5: Create matrix Y from X by random shuffling columns with indices $j \in J_1$ and matrix Z from X by random shuffling columns with indices $j \in J_2$
6: Use model M with design matrix Z as an input to obtain D_1 using Eq. (9)
7: Use model M with design matrix Y as an input to obtain D_2 using Eq. (9)
8: Evaluate $D_{12} = D - D_1 - D_2$
9: Compute Sobol importance score for the subset J_1 using Eq. (8):

$$SI_{J_1} = \frac{1}{D}(D_1 + D_{12})$$

As mentioned earlier, the main challenge in feature selection algorithms is the high computational cost due to huge number of subsets that need to be investigated. With Sobol sensitivity approach, the importance of both individual features and subsets of features can be computed using the same Monte Carlo integral. The next algorithm utilizes this computational efficiency of the approach and calculates importance score based on the total sensitivity index given in Eq. (5) up to second-order interactions. In many application areas, second order interactions are deemed sufficient to capture the joint sensitivity behaviour of features. To calculate the Sobol importance score SI_i for individual features, the subset J_1 in Algorithm 1 is set to $J_1 = \{i\}, i = 1, 2, \ldots, n$ and for joint importance score SI_{ij}, $J_1 = \{i, j\}, j = 1, 2, \ldots, i-1, i+1, \ldots, n$.

Algorithm 2 Feature selection based on total second order Sobol importances

1: Initialize the $n \times n$ matrix $\mathbf{S}$ of zeros
2: **for** i=1 **to** n **do**
3: **for** j=i **to** n **do**
4: **if** j == i **then**
5: Using Algorithm 1 calculate the individual importance SI_i of feature with index i and assign it to the diagonal elements of $\mathbf{S}$, i.e. $\mathbf{S}_{ii} = SI_i$
6: **else**
7: Using Algorithm 1 calculate the importance SI_{ij} of features with indices i, j and assign it to $\mathbf{S}_{ij}$
8: Sort the features based on the total second order sensitivity indices given by:

$$TSI_i = \sum_{j=1}^{n} \mathbf{S}_{ij}$$

9: For a given k, select the features with the highest k-TSI scores. k depends on the desired accuracy of the model.

Algorithm 2 requires $\frac{n(n+1)}{2}$ score evaluations for n features. All these evaluations are completed using the same Monte Carlo integral.

Similar to wrappers, Sobol sensitivity is a model-based approach that utilizes the trained model to evaluate feature importances. While wrappers select a subset of features to train the model, Sobol sensitivity uses the full feature set to train the model just as embedded methods do. Based on the trained model, it evaluates importance scores and, similar to filters, it identifies the subset of important features with highest scores without retraining the model. As the case for filters and wrappers, the optimality and efficiency of the technique depend on the training algorithm (learner) and/or model-performance metric used. Sobol sensitivity assumes normally distributed and statistically independent features. These two distributional assumptions are popular in many feature selection algorithms. It can, however, consider other feature probability distributions and can be implemented for statistically dependent features. Because it is variance-based measure, it can be applied for linear and nonlinear relationships between target variable and features. In terms of computational needs,

Sobol sensitivity approach can be considered one of the most tractable techniques because importance scores for all feature subsets are computed using the same Monte Carlo integral.

4 Application and Comparisons

In this section, we apply the proposed feature selection techniques to several data sets known in machine learning community and presents comparative evaluations of the results obtained with results obtained using: Random Forest (RF) and Support Vector Machine Recursive Feature Elimination (SVM-RFE).

Example 1: *Effect of noise on Sobol importance*.

In this example we demonstrate the behavior of Sobol importance approach under different levels of noise in the model starting from zero-level noise model (100 % accurate predictions). We consider a model function given by Friedman in [28]:

$$f(x_1, x_2, x_3, x_4, x_5) = 10\sin(\pi x_1 x_2) + 20\left(x_3 - \frac{1}{2}\right)^2 + 10x_4 + 5x_5 + \\ +\sigma h(x_6, x_7, ..., x_{15}). \tag{10}$$

We generate the dataset with 1000 training examples and 15 features (drawn from the normal distribution). Only first 5 features are important. Figure 4 shows the results of Algorithm 2 for different values of σ. Using 0.05 as the cutoff value to declare importance, it is easily seen that the first 5 features continue to be the important features for $\sigma \leq 1$. Once σ is inflated beyond 1, more features exhibit themselves as important. However, for all σ values (except $\sigma = 2.5$, the overall ranking of features continues to agree with the first five features being the subset of features demonstrating highest importance scores.

Example 2: *Comparisons using simulated data*.

Friedman model function used in example 1 is used here to generate a data set with 1000 training examples and 15 features drawn from the normal distribution, only first 5 features are used to calculate the function in (10). The values of the function are used as values of the target variable y.

We calculate Sobol importances based on four different models: Neural Network (NN), Support Vector Machine (SVM), Random Forest (RF) and Gradient Boosting Trees model (XGB). We compare the results with importances obtained using Random Forest (RF) and Support Vector Machine Recursive Feature Elimination (SVM-RFE). Figure 5 depicts the findings. All algorithms correctly identify the first five features as the important set of features. One exception is observed in SVM-RFE where features 12 and 14 are identified as equally important. This wrong feature identification can be due to nonlinearity in the relationship between target variable and features for which SVM-RFE can not capture.

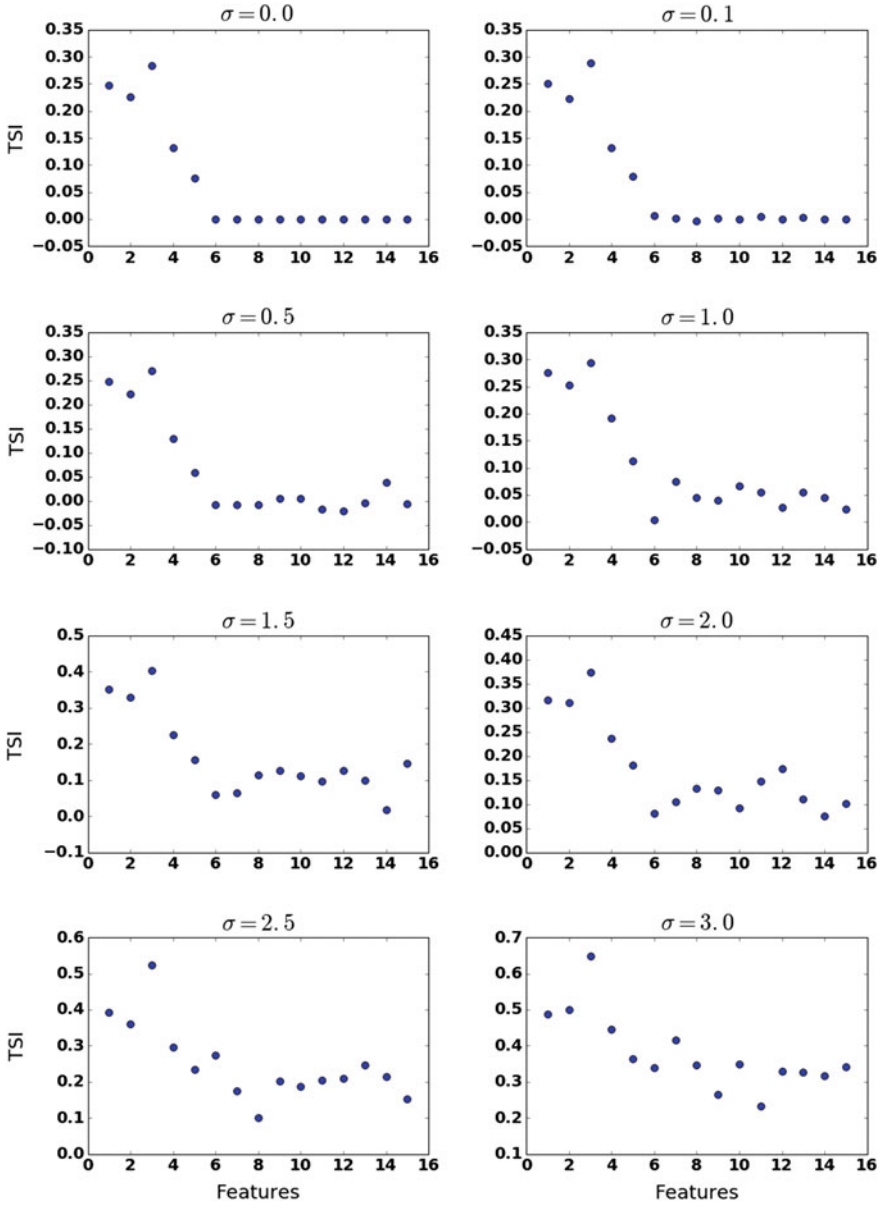


Fig. 4 Sobol importances for Friedman function with noise

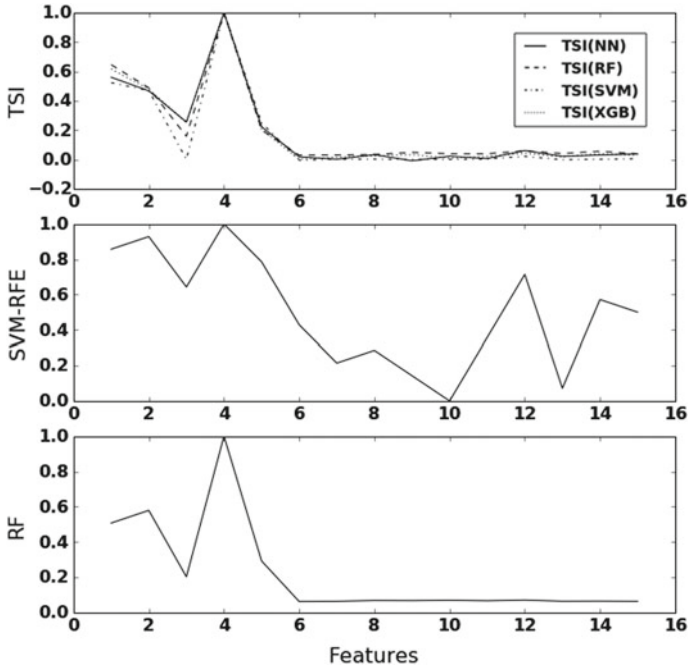


Fig. 5 Comparisons using simulated Friedman dataset

Example 3: *Comparisons using benchmark data sets*.

In this example, four benchmark data sets, described in Table 1, are used to evaluate the performance of Sobol sensitivity as compared to SVM-RFE and RF algorithms.

The model is trained using NN, RF and SVM methods and Sobol importance scores are calculated for all features in each data set using the three training methods. The results are compared to importance measures obtained using SVM-RFE and RF benchmark methods. For more meaningful comparisons, different threshold values for the number of important features selected from the benchmark importances. For example, threshold 10 means that we take top 10 important features from the benchmark algorithms. To compare the results we use Area Under the Curve (AUC) metric. The AUC calculates the overall differences in the feature rankings obtained by the benchmark method for a given threshold and those obtained from the Sobol importance scores. The general work flow is visualized in Fig. 6.

The resulting AUC values for the different models and the four datasets are plotted against different threshold values in Fig. 7. The Fig. 7 demonstrates that the different algorithms have succeeded in identifying comparable sets of important features. For example, the AUC between RF and Sobol RF feature rankings is higher than 0.9, which means that more than 90 % of features are common between the two algorithms.

Table 1 Datasets description

Set	Domain	Num. var.	Num. samples	Target	Data type	Ref.
SYLVA	Ecology	216	14394	Ponderosa pine	Continuous and discrete	WCCI 2006 Performance Prediction Challenge
HIVA	Drug discovery	1617	4229	Activity to AIDS HIV infection	Discrete (binary)	WCCI 2006 Performance Prediction Challenge
NOVA	Text	16969	1929	Separate politics from religion topics	Discrete (binary)	WCCI 2006 Performance Prediction Challenge
BANK	Financial	147	7063	Personal bankruptcy	Continuous and discrete	Foster and Stine [29]

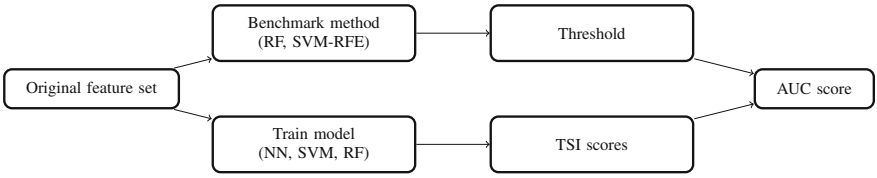


Fig. 6 Feature importances comparison framework

Furthermore, for each algorithm in Fig. 7, the top N% important features are selected and the model is trained by SVM algorithm on the selected subset of features. Table 2 presents the reduced model accuracy values expressed by the Root Mean Square Errors (RMSE). Reported in Table 2 also are the RMSE values for the trained model on all features in each data set.

Table 2 demonstrates that for all benchmark data sets, the reduced models give significantly more accurate predictions than that given by the full model. When top 10 % important features are used in the model, the Sobol approach gives better result than the benchmark algorithms SVM-RFE and RF. When including more than 10 % important features, SVM-RFE performs marginally better than Sobol approach in two of the data sets. In calculating variance contributions of features to the overall variability in the target variable, Sobol sensitivity approach can identify the small number of most important features more accurately than other methods. When larger number of features are desired in the model, the approach may fail to provide most accurate predictions due to the increased level of noise in the data. According to the analysis of Example 1, greater level of noise can distort the Sobol rankings of

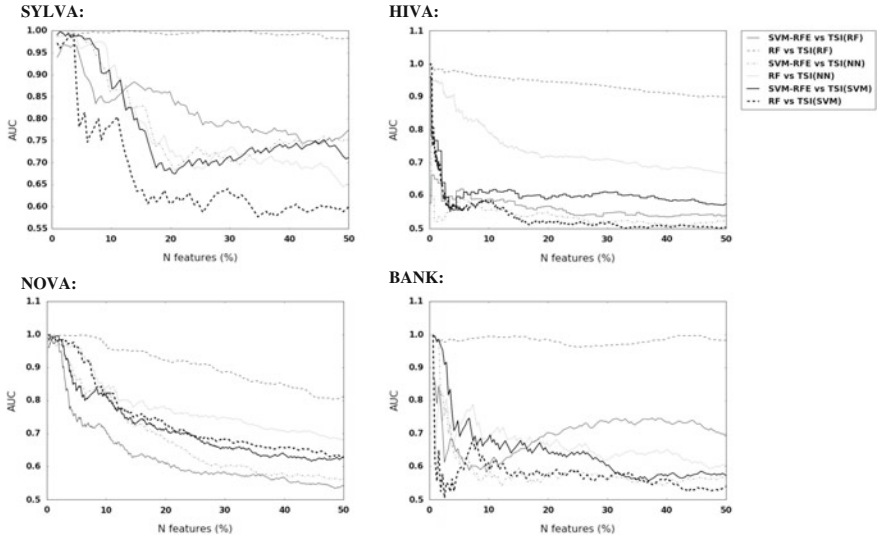


Fig. 7 Comparisons using benchmark data sets

important features. It may also imply that higher order interactions between features are needed for more accurate predictions in this situation.

The performance consistency of the three techniques measured by the standard deviation of model accuracy over N values was also examined. The standard deviation values, s, reported in Table 2 clearly exhibits comparable consistency between the three methods.

In summary, for the given data sets the total Sobol second order indices *TSI* produce the most accurate model predictions for the majority of cases. By varying the values of N, the *TSI* model accuracy shows small fluctuations (measured by s values) that are comparable to fluctuations observed by SVM-RFE or RF benchmark methods.

5 Conclusion

In this paper we implemented Sobol sensitivity analysis to select important features for the supervised data mining problem. We have proposed two algorithms for importance scoring: one algorithm to compute importance scores for the individual features and another one to compute importance scores for subsets of features. The main advantage of our proposed approach is lower computational cost and higher efficiency compared to many other existing algorithms. It can be applied for all types of relationships (linear or nonlinear) between target variable and features. A concern about the algorithm is that it estimates the importance of features with respect to the

Table 2 Reduced datasets model accuracy

N (top N % features selected)	SVM-RFE	RF	TSI(RF)	TSI(NN)	TSI(SVM)
SYLVA: RMSE (all features) = 0.144					
10 %	0.103	0.09	0.09	**0.085**	0.091
20 %	0.104	**0.088**	0.094	0.096	0.095
30 %	0.104	0.1	**0.094**	0.102	0.099
40 %	0.106	0.113	0.109	**0.102**	**0.102**
50 %	0.117	0.118	0.117	0.113	**0.11**
s	0.005	0.012	0.01	0.009	0.006
HIVA: RMSE (all features) = 0.179					
10 %	**0.163**	0.165	0.166	0.168	**0.163**
20 %	**0.162**	0.168	0.168	0.168	0.165
30 %	**0.161**	0.175	0.173	0.173	0.168
40 %	**0.165**	0.177	0.18	0.175	0.17
50 %	**0.168**	0.181	0.182	0.176	0.17
s	0.003	0.006	0.006	0.003	0.003
BANK: RMSE (all features) = 0.252					
10 %	0.227	0.235	0.24	0.225	**0.213**
20 %	**0.21**	0.236	0.239	0.224	0.217
30 %	**0.213**	0.229	0.238	0.223	0.216
40 %	**0.216**	0.229	0.23	0.225	0.222
50 %	**0.217**	0.226	0.221	0.23	0.226
s	0.006	0.004	0.007	0.002	0.005
NOVA: RMSE (all features) = 0.242					
10 %	0.256	0.275	0.27	**0.24**	0.28
20 %	0.24	0.241	0.259	**0.221**	0.26
30 %	0.237	0.24	0.253	**0.219**	0.248
40 %	0.234	0.24	0.246	**0.217**	0.238
50 %	0.236	0.244	0.246	**0.218**	0.233
s	0.008	0.01	0.009	0.009	0.017

model objective function which means that if the modeling algorithm is not accurate or overfitting, the Sobol approach may give misleading feature importances. The authors intend to further investigate the robustness of the approach to the training algorithm. As Example 3 has shown, the accuracy of the reduced model is higher than the full feature set model. It is then possible to train the model on a subset of features identified by a pre-processing algorithm and use the reduced model to compute Sobol feature importances. In addition to increasing Sobol importance reliability and efficiency, using reduced model to calculate importances reduces the computational

cost of the method. Another possible approach to reduce the computational cost is using Kolmogorov representation theorem [27] in which the model objective function can be expressed in an additive form of sub-functions, each as a single-variable function. If the hypothesis of additive model is true, then the Sobol algorithm can be simplified so as to require $2n$ model evaluations only.

Acknowledgments The authors wish to acknowledge the support of the American University of Sharjah, United Arab Emirates.

References

1. Hocking, R.R., Leslie, R.N.: Selection of the best subset in regression analysis. Technometrics **9**, 531–540 (1967)
2. Hoerl, A.E., Kennard, R.W.: Ridge regression: biased estimation for nonorthogonal problems. Technometrics **12**(1), 55–67 (1970)
3. Akaike, H.: Maximum likelihood identification of Gaussian autoregressive moving average models. Biometrika **60**(2), 255–265 (1973)
4. Akaike, H.: A new look at the statistical model identification. IEEE Trans. Autom. Control **AC-19**, 6, 716–723 (1974)
5. Schwarz, G.: Estimating the dimension of a model. Ann. Statis. **6**(2), 461–464 (1978)
6. Breiman, L.: Better subset regression using the nonnegative garrote. Technometrics **37**(4), 373–384 (1995)
7. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)
8. Tibshirani, R.J.: Regression shrinkage and selection via the lasso. J. Roy. Statis. Soc. **58**(1), 267–288 (1996)
9. Efron, B., Stein, C.: The Jackknife estimate of variance. Ann. Statis. **9**, 586–596 (1981)
10. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. Ann. Statist. **32**(2), 407–499 (2004)
11. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. Roy. Statis. Soc. **67**(2), 301–320 (2005)
12. Zou, H.: The adaptive lasso and its oracle properties. J. Am. Statis. Assoc. **101**, 476, 1418–1429 (2006)
13. Zhao, P., Yu, B.: Stagewise lasso. J. Mach. Learn. Res. **8**, 2701–2726 (2007)
14. Zhang, H.H., Ahn, J., Lin, X., Park, C.: Gene selection using support vector machines with non-convex penalty. Bioinformatics **22**(1), 88–95 (2006)
15. Yuan, M., Lin, Y.: On the non-negative garrotte estimator. J. Roy. Statis. Soc. **69**(2), 143–161 (2007)
16. Ishwaran, H.: Variable importance in binary regression trees and forests. Electron. J. Statis. **1**, 519–537 (2007)
17. Louppe, G., Wehenkel, L., Sutera, A., Geurts, P.: Understanding variable importances in forests of randomized trees. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (eds.). Advances in Neural Information Processing Systems vol. 26, pp. 431–439. Curran Associates, Inc. (2013)
18. Sobol, I.M.: On sensitivity estimation for nonlinear mathematical models (in Russian). Matematicheskoe Modelirovanie **2**, 112–118 (1990)
19. Sobol, I.M.: Sensitivity estimates for nonlinear mathematical models. Math. Model. Comput. Exper. **1**(4), 407–414 (1993)
20. Sobol, I.M.: Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Math. Comput. Simul. **55**, 271–280 (2001)

21. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**(1), 1157–1182 (2003)
22. Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L.A.: Feature Extraction: Foundations and Applications. Springer-Verlag, Berlin (2006)
23. Clarke, B., Fokoué, E., Zhang, H.H.: Principles and Theory for Data Mining and Machine Learning. Springer-Verlag, Berlin (2009)
24. Aliferis, C.F., Statnikov, A., Tsamardinos, I., Mani, S., Koutsoukos, X.D.: Local causal and Markov Blanket induction for causal discovery and feature selection for classification. J. Mach. Learn. Res. **11**, 171–234 (2010)
25. Janecek, A.G.K., Gansterer, W.N., Demel, M.A., Ecker, G.F.: On the relationship between feature selection and classification accuracy. In: JMLR: Workshop and Conference Proceedings, vol. 4, pp. 90–105 (2008)
26. Arwade, S.R., Moradi, M., Louhghalam, A.: Variance decomposition and global sensitivity for structural systems. Eng. Struct. **32**, 1–10 (2010)
27. Kolmogorov, A.N.: On the representation of continuous functions of several variables by superpositions of continuous functions of a smaller number of variables. Amer. Math. Soc. Transl. **17**, 369–373 (1961)
28. Friedman, J.: Multivariate adaptive regression splines. Ann. Statis. **19**(1), 1–67 (1991)
29. Foster, D.P., Stine, R.A.: Variable selection in data mining: building a predictive model for bankruptcy. J. Amer. Stat. Assoc. **99**, 303–313 (2004)

Basis Independence of Implicitly Defined Hamiltonian Circuit Dynamics

Jon Pierre Fortney

Abstract The Bloch-Crouch formulation of LC-circuit dynamics is seen to be an implicitly defined Hamiltonian system on a particular manifold. A particular basis independent Dirac structure is shown to be equivalent to the hybrid input-output representation of the Dirac structure used by Bloch and Crouch thereby allowing circuit dynamics to be written in a basis independent fashion.

Keywords Dirac structures · Hamiltonian dynamics · Implicitly defined Hamiltonian systems · Circuit theory · LCR circuits

Mathematics Subject Classification: 70H05 · 70H45 · 94C05 · 53D99 · 53Z05

1 Introduction

Bloch and Crouch [2] introduced a formulation of LC circuit dynamics which utilized a Dirac structure associated with the circuit. Their formulation turned out to be an implicity defined Hamiltonian system as introduced by van der Schaft, see for example [5, 10]. We note that utilizing the concept of implicitly defined port-Hamiltonian systems, also introduced by van der Schaft, one can extend the Bloch-Crouch formulation to LCR circuit dynamics. However, in writing down the Dirac structure Bloch and Crouch are implicitly specifying a basis.

In this paper we provide a basis independent form for this Dirac structure and then show that it is in fact equivalent to the Bloch-Crouch basis dependent form. Furthermore, nothing we do depends on the fact that the circuit is LC and so is also applicable to the implicitly defined port-Hamiltonian formulation for LCR circuits. Thus we show that the implicit (port-) Hamiltonian formulation of circuit dynamics is, as one would expect, basis independent. This is particularly interesting due to

J.P. Fortney (✉)
Department of Mathematics and Statistics,
Zayed University, P.O. Box 19282, Dubai, UAE
e-mail: jon.fortney@zu.ac.ae

T. Abualrub et al. (eds.), *Mathematics Across Contemporary Sciences*,
Springer Proceedings in Mathematics & Statistics 190,
DOI 10.1007/978-3-319-46310-0_5

recent work that illustrated the geometric relationship between implicitly defined port-Hamiltonian systems and pseudo-gradient systems [7], thereby extending the applicability of this work. This work is also interesting in a wider context where the use of Dirac structures has been extended to Lagrangian mechanics [11, 12]. These references also contains an extensive introduction to the history and background of interconnected and implicit systems for both the Hamiltonian and Lagrangian formalisms.

This paper is organized as follows. In part 2 we briefly review the definition of Dirac structures on vector spaces and on manifolds. We then define implicit Hamiltonian systems. In part 3 we review the Bloch-Crouch formulation of circuit dynamics. Our presentation is more geometrical than that of the original paper but makes it easier to see that this formulation is an implicitly defined Hamiltonian system. In part 4 we present a basis independent formulation of the Dirac structure utilized in part 3 and then show that it is in fact equivalent to the basis dependent formulation.

2 Basic Definitions

Dirac structures were originally introduced by Dorfman [6] and Courant [4] as simultaneous generalizations of both symplectic forms and Poisson structures. Given a vector space V and its dual V^* we can define the following symmetric bilinear form on $V \times V^*$.

$$\big\langle (v, \alpha), (w, \beta) \big\rangle_+ = \frac{\langle w, \alpha \rangle + \langle v, \beta \rangle}{2}$$

for all $(v, \alpha), (w, \beta) \in V \times V^*$, where $\langle \cdot, \cdot \rangle$ represents the canonical pairing between a vector space and its dual.

Definition 2.1 A general Dirac structure on an n-dimensional vector space V is a subspace $\mathcal{D} \subset V \times V^*$ such that $\mathcal{D}^\perp = \mathcal{D}$ where $\mathcal{D}^\perp = \{(v, \alpha) \in V \times V^* \mid \langle (v, \alpha), (w, \beta) \rangle_+ = 0, \ \forall (w, \beta) \in \mathcal{D}\}$.

We note here that a Dirac structure also satisfies an additional integrability condition, see [5]. In our context this integrability condition is automatically satisfied and plays no further role so we will not discuss it here.

Proposition 2.1 *A general Dirac structure on an n-dimensional vector space V is a subspace* $\mathcal{D} \subset V \times V^*$ *such that* ***(i)*** $\langle v, \alpha \rangle = 0, \ \forall (v, \alpha) \in \mathcal{D}$ *and* ***(ii)*** $\dim(\mathcal{D}) = n$.

This proposition is occasionally encountered as the definition of a general Dirac structure. While there are a number of different possible representations of Dirac structures we will only concern ourself with that introduced and utilized by Bloch and Crouch. The proof of this proposition is found in [2].

Proposition 2.2 (Hybrid Input-Output Representation) *Let V be an n-dimensional vector space with $W \subset V \times V^*$ an n-dimensional subspace of $V \times V^*$ and $\mathbb{J} : W \to W^*$ a skew symmetric linear map. Then $\mathcal{D} = graph(\mathbb{J})$ is a Dirac structure on V.*

The definition of Dirac structures on vector spaces can be readily extended to manifolds. Dirac structures on manifolds turn out to be a subbundle of $TM \oplus T^*M$ which are pointwise Dirac structures on the tangent vector space. The same symmetric bilinear pairing, evaluated pointwise on M, is used.

Definition 2.2 A general Dirac structure on a manifold M is a smooth vector subbundle $\mathcal{D} \subset TMx \oplus T^*M$ such that $\mathcal{D}^\perp = \mathcal{D}$ where $\mathcal{D}^\perp = \{(v, \alpha) \in TM \oplus T^*M | \langle (v, \alpha), (w, \beta) \rangle_+ = 0, \ \forall (w, \beta) \in \mathcal{D}\}$.

Again we present only the pertinent representation of a Dirac structure on a manifold.

Proposition 2.3 (Hybrid Input-Output Representation) *Let M be an n dimensional manifold with a general Dirac structure $\mathcal{D}$ and $W = V_1 \times V_2^* \subset TM \oplus T^*M$ an n dimensional vector bundle over M, i.e., for all $x \in M$, $W(x) = V_1(x) \times V_2^*(x) \subset T_xM \times T_x^*M$ is an n dimensional vector space. Then there exists a skew-symmetric linear map $\mathbb{J} : W \to W^*$ such that*

$$\mathcal{D} = \left\{ \big((v_1, \alpha_2), (\alpha_1, v_2)\big) \,\middle|\, (\alpha_1, v_2) = \mathbb{J}(v_1, \alpha_2), \ (v_1, \alpha_2) \in W \right\}.$$

Conversely, if $\mathcal{D}$ is the above set for some skew-symmetric linear map $\mathbb{J} : W \to W^*$ and some and $W = V_1 \times V_2^* \subset TM \oplus T^*M$ a smooth n dimensional vector bundle over M, then $\mathcal{D}$ is a general Dirac structure.

Finally we define implicit Hamiltonian systems as found in [1, 8].

Definition 2.3 Let $\mathcal{D}$ be a (general) Dirac structure on the manifold M and let $H \in C^\infty(M)$ be a smooth function on M called the Hamiltonian function. Then the implicit Hamiltonian system on M corresponding to $\mathcal{D}$ and H, denoted by $(M, \mathcal{D}, H)$, is defined by the specification that $\big(\dot{x}, dH(x)\big) \in \mathcal{D}(x)$.

Thus implicitly defined Hamiltonian systems can be utilized to define dynamics. Furthermore, implicit Hamiltonian systems are generalizations of classical Hamiltonian systems where the Dirac structure takes the place of the symplectic form.

3 Bloch-Crouch Formulation of LC Circuit Dynamics

In electrical engineering, Kirchhoff's current law states that the algebraic sum of the currents into or out of any node in an electrical circuit is zero. Kirchhoff's voltage law states that the algebraic sum of the voltages about any closed loop in an electrical circuit is zero. The word voltage in this statement of Kirchhoff's voltage law refers

to the voltage associated with a branch and is more accurately understood as the voltage drop across the branch. What is clear is that the network topology of the circuit determines the actual set of equations given by Kirchhoff's laws, thereby allowing graph-theoretic concepts to be used to write down Kirchhoff's laws.

A circuit can be represented as a connected graph $\mathcal{G}$ which consists of a set V of vertices, also called nodes, and a set E of edges, also called branches. A maximal tree $\mathcal{T}$ of G is a subgraph of $\mathcal{G}$ which is connected, contains all the nodes of $\mathcal{G}$, and has no loops. (In this paper the word loop is used to refer to what is called a circuit in graph-theoretical language; the word circuit is reserved for electrical circuits. We also generally use the word branch to refer to what is called an edge in graph-theoretical language.) The branches of $\mathcal{G}$ which are contained in the maximal tree $\mathcal{T}$ are called twigs and the branches contained in $\mathcal{L} = \mathcal{G} \setminus \mathcal{T}$ are called links. A fundamental cut set associated with a given twig is the set $\mathcal{C}$ of links, along with the given twig, that (a) when removed from $\mathcal{G}$ results in two disconnected graphs and (b) the removal of all but any one of the branches of $\mathcal{C}$ results in the graph remaining connected. A fundamental loop associated with a given link consists of the link along with the set of twigs connecting the link's nodes. The fundamental cut set associated with each twig and the fundamental loop associated with each link is unique.

The fundamental loops and fundamental cut sets associated with an arbitrarily chosen maximal tree can be used to give Kirchhoff's voltage and current laws. In a procedure explained fully in [3], one can find the fundamental cut-set matrix Q and the fundamental loop matrix B associated with a maximal tree $\mathcal{T}$. The entries of both Q and B are -1, 1, or 0, depending on whether specific branches are included in particular cut-sets or loops (zero or non-zero), and what their orientation with respect to the cut-set or loop is (positive or negative). In essence, the matrices B and Q encode the topology of the circuit. Furthermore, due to the topological relationship between cut-sets and loops, these matrices are related to each other by $B = -Q^T$, also proved in [3].

Let $v_{\mathcal{T}}$ and $v_{\mathcal{L}}$ denote the twig and link voltages respectively and $i_{\mathcal{T}}$ and $i_{\mathcal{L}}$ denote the twig and link currents respectively. It is then possible to write Kirchhoff's laws in terms of the graph-theoretic matrices Q and B as

$$\begin{bmatrix} v_{\mathcal{T}} \\ v_{\mathcal{L}} \end{bmatrix} = \begin{bmatrix} I \\ Q \end{bmatrix} v_{\mathcal{T}} \quad \text{and} \quad \begin{bmatrix} i_{\mathcal{T}} \\ i_{\mathcal{L}} \end{bmatrix} = \begin{bmatrix} B \\ I \end{bmatrix} i_{\mathcal{L}}.$$

The above equations can be combined into

$$\begin{bmatrix} i_{\mathcal{T}} \\ v_{\mathcal{L}} \end{bmatrix} = \begin{bmatrix} 0 & -Q^T \\ Q & 0 \end{bmatrix} \begin{bmatrix} v_{\mathcal{T}} \\ i_{\mathcal{L}} \end{bmatrix} \tag{1}$$

where the matrix is clearly a skew symmetric linear map. To simplify our exposition we will make the reasonable assumption that Faraday's laws are linear and time invariant. Thus we have

$$i(t) = C\frac{dv(t)}{dt} \quad \text{and} \quad v(t) = L\frac{di(t)}{dt}$$

where C and L are constants associated with the capacitors and inductors. Integrating both sides we get $q(t) = Cv(t)$ and $\phi(t) = Li(t)$ where q is charge and ϕ is flux. Therefore we have $\dot{q} = i$ and $\dot{\phi} = v$.

The manifold on which the Bloch-Crouch formulation of circuit dynamics resides is given by the set of all capacitor charges and inductor fluxes for the circuit which we will write as $M = \{(q_{\mathcal{T}}^C, q_{\mathcal{L}}^C, \phi_{\mathcal{T}}^L, \phi_{\mathcal{L}}^L)\}$. For $m \in M$, $m = (q_{\mathcal{T}}^C, q_{\mathcal{L}}^C, \phi_{\mathcal{T}}^L, \phi_{\mathcal{L}}^L)$, we have the tangent space of M at the point m to be given by

$$\begin{aligned} T_m M &= \left\{(\dot{q}_{\mathcal{T}}^C(m), \dot{q}_{\mathcal{L}}^C(m), \dot{\phi}_{\mathcal{T}}^L(m), \dot{\phi}_{\mathcal{L}}^L(m))\right\} \\ &= \left\{(i_{\mathcal{T}}^C(m), i_{\mathcal{L}}^C(m), v_{\mathcal{T}}^L(m), v_{\mathcal{L}}^L(m))\right\}. \end{aligned}$$

Recalling that currents and voltages are considered dual to each other (which is made mathematically precise in the next section) we have

$$T_m^* M = \left\{(v_{\mathcal{T}}^C(m), v_{\mathcal{L}}^C(m), i_{\mathcal{T}}^L(m), i_{\mathcal{L}}^L(m))\right\}.$$

The Hamiltonian function $H : M \to \mathbb{R}$ associated with a circuit is given by

$$\begin{aligned} H(q^C, \phi^L) = &\sum_{\mathcal{T}} \frac{(q_{\mathcal{T}}^C)^2}{2C_{\mathcal{T}}} + \sum_{\mathcal{L}} \frac{(q_{\mathcal{L}}^C)^2}{2C_{\mathcal{L}}} \\ &+ \sum_{\mathcal{T}} \frac{(\phi_{\mathcal{T}}^L)^2}{2L_{\mathcal{T}}} + \sum_{\mathcal{L}} \frac{(\phi_{\mathcal{L}}^L)^2}{2L_{\mathcal{L}}} \end{aligned}$$

where on the far right we have partitioned our charges and fluxes according to whether the associated capacitors or inductors are on twigs or links. This in turn gives us

$$\begin{aligned} \left.\frac{\partial H}{\partial \phi_{\mathcal{T}}^L}\right|_m &= i_{\mathcal{T}}^L(m), \quad & \left.\frac{\partial H}{\partial \phi_{\mathcal{L}}^L}\right|_m &= i_{\mathcal{L}}^L(m), \\ \left.\frac{\partial H}{\partial q_{\mathcal{T}}^C}\right|_m &= v_{\mathcal{T}}^C(m), \quad & \left.\frac{\partial H}{\partial q_{\mathcal{L}}^C}\right|_m &= v_{\mathcal{L}}^C(m) \end{aligned}$$

resulting in $dH(m) \in T_m^* M$ as expected. Notice that these equations are in fact the integrated form of Faraday's laws mentioned above.

We now show that for each $m \in M$ Eq. (1) is the hybrid-input-output representation of a Dirac structure on $T_m M$. By making a finer partition of the variables in (1) we obtain

$$\begin{bmatrix} i_{\mathcal{T}}^{C} \\ i_{\mathcal{T}}^{L} \\ v_{\mathcal{L}}^{C} \\ v_{\mathcal{L}}^{L} \end{bmatrix} = \begin{bmatrix} 0 & 0 & -Q_{11}^{T} & -Q_{21}^{T} \\ 0 & 0 & -Q_{12}^{T} & -Q_{22}^{T} \\ Q_{11} & Q_{12} & 0 & 0 \\ Q_{21} & Q_{22} & 0 & 0 \end{bmatrix} \begin{bmatrix} v_{\mathcal{T}}^{C} \\ v_{\mathcal{T}}^{L} \\ i_{\mathcal{L}}^{C} \\ i_{\mathcal{L}}^{L} \end{bmatrix}$$

which becomes, upon rearrangement,

$$\begin{bmatrix} v_{\mathcal{L}}^{C} \\ i_{\mathcal{T}}^{L} \\ i_{\mathcal{T}}^{C} \\ v_{\mathcal{L}}^{L} \end{bmatrix} = \begin{bmatrix} 0 & Q_{12} & Q_{11} & 0 \\ -Q_{12}^{T} & 0 & 0 & -Q_{22}^{T} \\ -Q_{11}^{T} & 0 & 0 & -Q_{21}^{T} \\ 0 & Q_{22} & Q_{21} & 0 \end{bmatrix} \begin{bmatrix} i_{\mathcal{L}}^{C} \\ v_{\mathcal{T}}^{L} \\ v_{\mathcal{T}}^{C} \\ i_{\mathcal{L}}^{L} \end{bmatrix}.$$

We shall name the skew-symmetric matrix above $\mathbb{J}$. For each $m \in M$ we define the vector spaces

$$V_1(m) = \left\{ \begin{bmatrix} i_{\mathcal{L}}^{C} \\ v_{\mathcal{T}}^{L} \end{bmatrix} \right\}, \quad V_2(m) = \left\{ \begin{bmatrix} i_{\mathcal{T}}^{C} \\ v_{\mathcal{L}}^{L} \end{bmatrix} \right\},$$
$$V_1^*(m) = \left\{ \begin{bmatrix} v_{\mathcal{L}}^{C} \\ i_{\mathcal{T}}^{L} \end{bmatrix} \right\}, \quad V_2^*(m) = \left\{ \begin{bmatrix} v_{\mathcal{T}}^{C} \\ i_{\mathcal{L}}^{L} \end{bmatrix} \right\}.$$

It is clear that $T_m M = V_1(m) \times V_2(m)$ and $T_m^* M = V_1^*(m) \times V_2^*(m)$ and $W(m) = V_1(m) \times V_2^*(m) \subset T_m M \times T_m^* M$. Thus $\mathbb{J}$ defines a skew-symmetric linear mapping $\mathbb{J}(m) : W(m) \to W^*(m)$. The matrix $\mathbb{J}$ and the vector spaces $W(m)$ and $W^*(m)$ are independent of $m \in M$ so by the above proposition we have the hybrid input-output representation of a Dirac structure $\mathcal{D}$ on M.

By letting $x_1 = \{(q_{\mathcal{L}}^{C}, \phi_{\mathcal{T}}^{L})\}$ and $x_2 = \{(q_{\mathcal{T}}^{C}, \phi_{\mathcal{L}}^{L})\}$ we have

$$\dot{x}_1 = \begin{bmatrix} \dot{q}_{\mathcal{L}}^{C} \\ \dot{\phi}_{\mathcal{T}}^{L} \end{bmatrix} \in V_1, \quad \dot{x}_2 = \begin{bmatrix} \dot{q}_{\mathcal{T}}^{C} \\ \dot{\phi}_{\mathcal{L}}^{L} \end{bmatrix} \in V_2,$$
$$\frac{\partial H}{\partial x_1} = \begin{bmatrix} \frac{\partial H}{\partial q_{\mathcal{L}}^{C}}^{T} \\ \frac{\partial H}{\partial \phi_{\mathcal{T}}^{L}}^{T} \end{bmatrix} \in V_1^*, \quad \frac{\partial H}{\partial x_2} = \begin{bmatrix} \frac{\partial H}{\partial q_{\mathcal{T}}^{C}}^{T} \\ \frac{\partial H}{\partial \phi_{\mathcal{L}}^{L}}^{T} \end{bmatrix} \in V_2^*.$$

The specification $(\dot{x}, dH) \in \mathcal{D}$ then leads to

$$\begin{bmatrix} \frac{\partial H}{\partial q_{\mathcal{L}}^{C}}^{T} \\ \frac{\partial H}{\partial \phi_{\mathcal{T}}^{L}}^{T} \\ \dot{q}_{\mathcal{T}}^{C} \\ \dot{\phi}_{\mathcal{L}}^{L} \end{bmatrix} = \begin{bmatrix} 0 & Q_{\mathcal{L}_{12}} & Q_{\mathcal{L}_{11}} & 0 \\ -Q_{\mathcal{L}_{12}}^{T} & 0 & 0 & -Q_{\mathcal{L}_{22}}^{T} \\ -Q_{\mathcal{L}_{11}}^{T} & 0 & 0 & -Q_{\mathcal{L}_{21}}^{T} \\ 0 & Q_{\mathcal{L}_{22}} & Q_{\mathcal{L}_{21}} & 0 \end{bmatrix} \begin{bmatrix} \dot{q}_{\mathcal{L}}^{C} \\ \dot{\phi}_{\mathcal{T}}^{L} \\ \frac{\partial H}{\partial q_{\mathcal{T}}^{C}}^{T} \\ \frac{\partial H}{\partial \phi_{\mathcal{L}}^{L}}^{T} \end{bmatrix}.$$

But these are exactly the system of equations one obtains from Kirchhoff's and Faraday's laws and so are the dynamical equations for LC circuits. Hence LC circuit dynamics are described by an implicitly defined Hamiltonian system. This

geometrical formulation can be readily extended to LCR circuits by utilizing implicitly defined port-Hamiltonian systems. The details of this will not be provided here.

4 Basis Independent Formulation of $\mathcal{D}$

Using concepts from algebraic topology we present a basis free formulation of Kirchhoff's laws. This formulation of Kirchhoff's laws has been previously mentioned in the literature, see for example Smale [9]. We show that this formulation is a Dirac structure, an observation which we believe is new. Then we show a maximal tree of the graph associated with the circuit induces a basis for this Dirac structure. With this basis the hybrid input-output representation of the Dirac structure is obtained.

An electrical circuit has an associated graph $\mathcal{G} = (V, E)$ consisting of a collection of vertices, also called nodes, or 0-simplices, $V = \{x_1, \ldots, x_n\}$ and edges, also called branches, or 1-simplices, $E = \big\{\{x_i, x_j\}\big\}$. Graphs of electrical circuit have the property that every branch is part of a loop. The state a circuit is in at any given time is determined by the currents i_k, along each branch and the voltage drops, v_k, across each branch, where k indexes the branches. The voltage potential at each node is measured relative to ground with the voltage drop being the difference between the node voltages potentials. Implicit in measuring the currents along a branch and the voltage drop across a branch is the fact that the branch must have an (arbitrary) orientation assigned to them. We use $[x_i, x_j]$ to denote the directed branch $\{x_i, x_j\}$. Clearly $[x_i, x_j] = -[x_j, x_i]$. No direction can be assigned to a point, but for consistency we use $[x_i]$ to denote the node $\{x_i\}$.

We then define the vector spaces $C_1(\mathcal{G}) = \text{span}\big\{[x_i, x_j]|\{x_i, x_j\} \in E\big\}$ and $C_0(\mathcal{G}) = \text{span}\big\{[x_i]|x_i \in V\big\}$ as formal sums over the reals. Since $C_i(\mathcal{G}), i = 0, 1$, are vector spaces we can define their dual spaces, $C^i(\mathcal{G}) = (C_i(\mathcal{G}))^*$, for $i = 0, 1$. Elements of $C_0(\mathcal{G})$, $C_1(\mathcal{G})$, $C^0(\mathcal{G})$, and $C^1(\mathcal{G})$ are called 0-chains, 1-chains, 0-cochains, and 1-cochains of $\mathcal{G}$ respectively. Also, $\{[x_i]\}$ $\{[x_i, x_j]\}$, $\{[x_i]^*\}$, and $\{[x_i, x_j]^*\}$ are called the standard bases of $C_0(\mathcal{G})$, $C_1(\mathcal{G})$, $C^0(\mathcal{G})$, and $C^1(\mathcal{G})$ respectively. A linear mapping called the boundary operator $\partial_1 : C_1(\mathcal{G}) \to C_0(\mathcal{G})$ is defined on basis elements of $C_1(\mathcal{G})$ by $\partial_1([x_i, x_j]) = [x_i] - [x_j]$. (This is simply the pertinent special case of the boundary operator from algebraic topology.) Its adjoint is a mapping $\partial_1^* : C^0(\mathcal{G}) \to C^1(\mathcal{G})$. Since ∂_1 and ∂_1^* are the boundary operators most relevant here we will in general simply write them as ∂ and ∂^*,

$$C^1(\mathcal{G}) \xleftarrow{\partial^*} C^0(\mathcal{G})$$
$$C_1(\mathcal{G}) \xrightarrow{\partial} C_0(\mathcal{G}).$$

A circuit's current is considered to be a vector $i = [i_1, \ldots, i_b]^T$ in $C_1(\mathcal{G})$ with respect to the standard basis. Thus currents are 1-chains. Note that as required an orientation for each edge is provided in this formulation. Likewise, an electrical circuit with n nodes has a voltage associated with each node. (This would be the

voltage potential at the node with respect to ground.) Thus a circuit's voltage can, in a sense, be considered to be a vector $v = [v_1, \ldots, v_n]^T$ in $C^0(\mathcal{G})$ with respect to the standard basis. Note, this is not what is most usually called the voltage in engineering. A straightforward calculation shows that in $\partial_1^*(v) \in C^1(\mathcal{G})$ the coefficient of each basis element $[x_i, x_j]^*$ in the terms of $\partial_1^*(v)$ is the voltage drop across the oriented branch $[x_i, x_j]$. Hence the voltage drops across branches are elements of $C^1(\mathcal{G})$ and are thus 1-cochains. Thus currents and voltage drops are dual variables in a mathematically precise way. It is these voltage drops which are referred to as simply the voltage in engineering. In keeping with the traditional engineering terminology and notation we will continue to abuse notation and denote the 1-cochains $\partial_1^*(v)$ by v as well.

Kirchhoff's current law (KCL) states that the algebraic sum of the currents into and out of any node is identically zero which can be written as $i \in \mathrm{Ker}(\partial_1)$. Kirchhoff's voltage law (KVL) states that the algebraic sum of the voltages (really voltage drops) around any closed path in a circuit is identically zero. One-cochains v for which this is true are clearly in $\mathrm{Ker}(\partial_2^*) \subset C^1(\mathcal{G})$, where $\partial_2^* : C^1(\mathcal{G}) \to C^2(\mathcal{G})$ is the adjoint of the boundary operator on 2-simplices. However, by choosing one node as "ground" it can be seen that in the current situation $\mathrm{Ker}(\partial_2^*) = \mathrm{Im}(\partial_1^*)$. This gives $\mathrm{Ker}(\partial_1) \oplus \mathrm{Im}(\partial_1^*) \equiv \mathrm{Ker}(\partial) \oplus \mathrm{Im}(\partial^*)$ as a basis independent formulation of Kirchhoff's laws. Next we show this space is a Dirac structure.

Theorem 4.1 *$\mathcal{D} = Ker(\partial) \oplus Im(\partial^*)$ is a Dirac structure.*

Proof We first show $\langle v, i \rangle = 0$ for all $(i, v) \in \mathrm{Ker}(\partial) \oplus \mathrm{Im}(\partial^*)$. There clearly exists a $c \in C^0(\mathcal{G})$ such that $v = \partial^* c$. Hence we have $\langle v, i \rangle = \langle \partial^* c, i \rangle = \langle c, \partial i \rangle = \langle c, 0 \rangle = 0$. We note that this is actually Tellegen's theorem from electrical engineering. Next we show condition $\dim(\mathcal{D}) = n$. By the fundamental theorem of linear algebra we have that $\dim(\mathrm{Ker}(\partial)) + \dim(\mathrm{Im}(\partial)) = \dim(C_1(\mathcal{G}))$. We also have that $\dim(\mathrm{Im}(\partial)) = \mathrm{rank}(\partial) = \mathrm{rank}(\partial^*) = \dim(\mathrm{Im}(\partial^*))$. Combining we get that $\dim(\mathrm{Ker}(\partial) \oplus \mathrm{Im}(\partial^*)) = \dim(C_1(\mathcal{G}))$. Thus both conditions are satisfied and $\mathrm{Ker}(\partial) \oplus \mathrm{Im}(\partial^*)$ is a Dirac structure. We note that in fact this proof holds for any ∂_n. Thus we have $\mathrm{Ker}(\partial_n) \oplus \mathrm{Im}(\partial_n^*)$ is a Dirac structure for all n. □

Theorem 4.2 *Let $\mathcal{G}$ be the graph associated with an electrical circuit. Then given any maximal tree $\mathcal{T}$ of $\mathcal{G}$, the set of links of $\mathcal{T}$ induce a basis on $Ker(\partial)$ and the set of twigs of $\mathcal{T}$ induce a basis on $Im(\partial^*)$.*

Proof Suppose that a circuit has graph $\mathcal{G}$ with n nodes and b oriented branches. Any maximal tree on $\mathcal{G}$ has $n - 1$ twigs and $l = b - (n - 1)$ links. Let $[x_{i_k}, x_{j_k}]$, $1 \le k \le l$, be the links and $[x_{i_k}, x_{j_k}]$, $l + 1 \le k \le b$, be the twigs.
(a) links $\longrightarrow$ linearly independent elements of $\ker(\partial)$:

Between any two vertices of $\mathcal{G}$ there exists a unique path which lies in $\mathcal{T}$. Existence of such a path follows from the maximal tree being connected and containing all vertices of $\mathcal{G}$. Suppose the path is not unique, then there are two paths on $\mathcal{T}$ connecting the two vertices. This would give rise to a loop on $\mathcal{T}$ which contradicts the fact that $\mathcal{T}$ contains no loops. Thus each link $[x_{i_k}, x_{j_k}]$, together with the unique path on $\mathcal{T}$

between the vertices $[x_{i_k}]$ and $[x_{j_k}]$ which constitute the boundary of the link, gives a unique loop. In other words, for each k, $1 \leq k \leq l$, we define the unique loop set as the set of branches

$$\mathcal{C}_k = \Big\{ \underbrace{[x_{i_k}, x_{j_k}]}_{\text{link}}, \underbrace{[x_{i_{l_1}}, x_{j_{l_1}}], \ldots, [x_{i_{l_{n(k)}}}, x_{j_{l_{n(k)}}}]}_{\text{twigs on path between } [x_{i_k}], [x_{j_k}]} \Big\}.$$

By an abuse of notation define $\mathcal{C}_k \in C_1(\mathcal{G})$, $1 \leq k \leq l$, by

$$\mathcal{C}_k = [x_{i_k}, x_{j_k}] + \sum_{m=l+1}^{b} q_{km}[x_{i_m}, x_{j_m}]$$

where $q_{km} = 0$ if the twig $[x_{i_m}, x_{j_m}]$ is not one of the twigs on the path between $[x_{i_k}]$ and $[x_{j_k}]$. If $[x_{i_m}, x_{j_m}]$ is one of the twigs on the path between $[x_{i_k}]$ and $[x_{j_k}]$ then $q_{km} = +1$ if, when transversing the loop, the orientation of the twig is the same as that of the kth link and $q_{km} = -1$ if the orientation is opposite that of the kth link. Furthermore, since each $\mathcal{C}_k$, $1 \leq k \leq l$, contains a term associated with a different link then these $\mathcal{C}_k$ are linearly independent elements in $C_1(\mathcal{G})$. Since each of these elements is a loop in $\mathcal{G}$ we have

$$\partial\Big(\mathcal{C}_k\Big) = \partial\Big([x_{i_k}, x_{j_k}] + \sum_{t=l+1}^{b} q_{kt}[x_{i_t}, x_{j_t}]\Big) = 0$$

for each $1 \leq k \leq l$. In other words, these $l = b - (n-1)$ elements are in $\mathrm{Ker}(\partial)$.
(b) twigs $\longrightarrow$ linearly independent elements of $\mathrm{Im}(\partial^*)$:

Here we first note that for each twig of $\mathcal{T}$ there exists a unique cut set, denoted by $\mathcal{C}^k$, $l+1 \leq k \leq b$, of the graph $\mathcal{G}$ which consists of the one twig and some links. To see this we note that by removing twig k from $\mathcal{T}$ we separate $\mathcal{T}$ into two disjoint components $\mathcal{G}_I^k$ and $\mathcal{G}_T^k$ where the initial vertex of twig k is in $\mathcal{G}_I^k$ and the terminal vertex of twig k is in $\mathcal{G}_T^k$. We then remove all links that have an initial vertex in $\mathcal{G}_I^k$ and a terminal vertex in $\mathcal{G}_T^k$ or that have an initial vertex in $\mathcal{G}_T^k$ and a terminal vertex in $\mathcal{G}_I^k$. The set of removed branches constitutes $\mathcal{C}^k$. Since we constructed $\mathcal{C}^k$ its existence is clear. Uniqueness is also obvious, for suppose there existed two different cut sets $\mathcal{C}^k$ and $\tilde{\mathcal{C}}^k$ associated with twig k. Since the two cut sets are different then, without loss of generality, there must exist some links in $\mathcal{C}^k$ which are not in $\tilde{\mathcal{C}}^k$. But these links connect the components $\mathcal{G}_I^k$ and $\mathcal{G}_T^k$ and hence $\tilde{\mathcal{C}}^k$ can not be a cut set, a contradiction. Therefore, for each k, $l+1 \leq k \leq b$ we have an associated cut set

$$\mathcal{C}^k = \Big\{ \underbrace{[x_{i_k}, x_{j_k}]}_{\text{twig}}, \underbrace{[x_{i_{l_1}}, x_{j_{l_1}}], \ldots, [x_{i_{l_{n(k)}}}, x_{j_{l_{n(k)}}}]}_{\text{links in cut set}} \Big\}.$$

As above define $\mathcal{C}^k \in C^1(\mathcal{G})$, $l+1 \leq k \leq b$, by

$$\mathcal{C}^k = [x_{i_k}, x_{j_k}]^* + \sum_{m=1}^{l} r_{km}[x_{i_m}, x_{j_m}]^*,$$

where $r_{km} = 0$ if the link $[x_{i_m}, x_{j_m}]$ is not in the kth cut set. If $[x_{i_m}, x_{j_m}]$ is in the cut set, then $r_{km} = +1$ if $[x_{i_m}] \in \mathcal{G}_I^k$ and $[x_{j_m}] \in \mathcal{G}_T^k$ and $r_{km} = -1$ if $[x_{i_m}] \in \mathcal{G}_T^k$ and $[x_{j_m}] \in \mathcal{G}_I^k$. Since each of these $\mathcal{C}^k$ contains a term for a different twig then it is clear that they are linearly independent.

Next we show, for some fixed k, $l+1 \leq k \leq b$ that $\mathcal{C}^k \in \mathrm{Im}(\partial^*)$. To do this we must find a 0-cochain $c^k = \sum_{m=1}^{n} c_m^k [x_m]^* \in C^0(\mathcal{G})$ such that $\partial^*(c^k) = \mathcal{C}^k$. If this is true, then for each $i \in C_1(\mathcal{G})$ we have $\langle \mathcal{C}^k, i\rangle = \langle \partial^*(c^k), i\rangle = \langle c^k, \partial(i)\rangle$. Suppose that $[x_{i_{\tilde{k}}}, x_{j_{\tilde{k}}}]$ is a basis element of $C_1(\mathcal{G})$ associated with one of the twigs of $\mathcal{T}$. Then

$$\left\langle \mathcal{C}^k, [x_{i_{\tilde{k}}}, x_{j_{\tilde{k}}}] \right\rangle = [x_{i_k}, x_{j_k}]^* [x_{i_{\tilde{k}}}, x_{j_{\tilde{k}}}] = \delta_{\tilde{k}}^k,$$

where $\delta_{\tilde{k}}^k = 1$ if $\tilde{k} = k$ and 0 otherwise, and

$$\left\langle \mathcal{C}^k, [x_{i_{\tilde{k}}}, x_{j_{\tilde{k}}}] \right\rangle = \Big(\sum_{m=1}^{n} c_m^k [x_m]^* \Big) \Big([x_{i_{\tilde{k}}}] - [x_{j_{\tilde{k}}}] \Big) = c_{i_{\tilde{k}}}^k - c_{j_{\tilde{k}}}^k.$$

Thus we have $c_{i_{\tilde{k}}}^k - c_{j_{\tilde{k}}}^k = \delta_{\tilde{k}}^k$. Next, for links $[x_{i_{\tilde{l}}}, x_{j_{\tilde{l}}}] \in \mathcal{C}^k$, where here $\mathcal{C}^k$ denotes the cut set, we have

$$\begin{aligned} r_{k\tilde{l}} &= \left\langle \mathcal{C}^k, [x_{i_{\tilde{l}}}, x_{j_{\tilde{l}}}] \right\rangle \\ &= \Big(\sum_{m=1}^{n} c_m^k [x_m]^* \Big) \Big([x_{i_{\tilde{l}}}] - [x_{j_{\tilde{l}}}] \Big) \\ &= c_{i_{\tilde{l}}}^k - c_{j_{\tilde{l}}}^k. \end{aligned}$$

For links $[x_{i_{\tilde{l}}}, x_{j_{\tilde{l}}}] \notin \mathcal{C}^k$, $0 = \left\langle \mathcal{C}^k, [x_{i_{\tilde{l}}}, x_{j_{\tilde{l}}}] \right\rangle = c_{i_{\tilde{l}}}^k - c_{j_{\tilde{l}}}^k$. For a fixed k, $l+1 \leq k \leq b$, we have the following system of equations:

$$\begin{aligned} c_{i_k}^k - c_{j_k}^k &= 1 : \text{One equation, cut set } \mathcal{C}^k \text{ twig} \\ c_{i_{\tilde{k}}}^k - c_{j_{\tilde{k}}}^k &= 0 : n-2 \text{ equations, non-cut set } \mathcal{C}^k \text{ twigs} \\ c_{i_{\tilde{l}}}^k - c_{j_{\tilde{l}}}^k &= \pm 1 : |\mathcal{C}^k| - 1 \text{ equations, cut set } \mathcal{C}^k \text{ links} \\ c_{i_{\tilde{l}}}^k - c_{j_{\tilde{l}}}^k &= 0 : b - (n-1) - (|\mathcal{C}^k| - 1) \\ &\qquad \text{equations, non-cut set } \mathcal{C}^k \text{ links.} \end{aligned}$$

Thus for our fixed k we have b equations in the variables c_m^k where $1 \leq m \leq n$ and $b \geq n$ since the number of branches in an electrical network is greater than or equal to the number of nodes. Thus we have either an equal number of equations and variables or more equations than variables. It is easy to see that if we let $c_m^k = 1$ for all $[x_m] \in \mathcal{G}_I^k$ and $c_m^k = 0$ for all $[x_m] \in \mathcal{G}_T^k$ this constitutes a non-trivial solution set for this system of equations.

Consider first the one equation $c_{i_k}^k - c_{j_k}^k = 1$ associated with the twig $[x_{i_k}, x_{j_k}]^*$ in $\mathcal{C}^k$. This twig was in fact used to determine the sets $\mathcal{G}_I^k$ and $\mathcal{G}_T^k$ and clearly $[x_{i_k}] \in \mathcal{G}_I^k$ and $[x_{j_k}] \in \mathcal{G}_T^k$. Thus, with the choice made above we have $c_{i_k}^k - c_{j_k}^k = 1 - 0 = 1$, satisfying the first equation. For the next $n-2$ equations, associated with the non-cut set $\mathcal{C}^k$ twigs, then it is clear that these twigs either have both of their nodes in $\mathcal{G}_I^k$ or have both of their nodes in $\mathcal{G}_T^k$. In the fist case $c_{i_{\bar{k}}}^k - c_{j_{\bar{k}}}^k = 1 - 1 = 0$ and in the second case $c_{i_{\bar{k}}}^k - c_{j_{\bar{k}}}^k = 0 - 0 = 0$.

For the following $|\mathcal{C}^k| - 1$ equations consider how the sign of the $r_{k\bar{l}}$ were chosen. The sign was positive if the initial and terminal nodes of the link are in the same sets as the initial and terminal nodes of the kth twig respectively, and negative otherwise. In the first case we have $c_{i_{\bar{l}}}^k = 1$ and $c_{j_{\bar{l}}}^k = 0$ giving us $c_{i_{\bar{l}}}^k - c_{j_{\bar{l}}}^k = 1 - 0 = 1$ which is what we wanted. In the second case we have $c_{i_{\bar{l}}}^k = 0$ and $c_{j_{\bar{l}}}^k = 1$ giving us $c_{i_{\bar{l}}}^k - c_{j_{\bar{l}}}^k = 0 - 1 = -1$ which is again what we wanted. Finally, it is clear that for the $b - (n-1) - (|\mathcal{C}^k| - 1)$ equations associated with the non-cut set $\mathcal{C}^k$ links that those links either have both of their nodes in $\mathcal{G}_I^k$ or have both of their nodes in $\mathcal{G}_T^k$. In the fist case we have $c_{i_{\bar{l}}}^k - c_{j_{\bar{l}}}^k = 1 - 1 = 0$ and in the second case we have $c_{i_{\bar{l}}}^k - c_{j_{\bar{l}}}^k = 0 - 0 = 0$. Thus we have found a c^k such that $\partial^*(c^k) = \mathcal{C}^k$. This process can be carried out for each $l+1 \leq k \leq b$, thus giving a 0-cochain in $C^0(\mathcal{G})$ which is the pre-image the 1-cochain $\mathcal{C}^k$. Therefore we have that the set of $\mathcal{C}^k$, $l+1 \leq k \leq b$, is a linearly independent set of $\mathrm{Im}(\partial^*)$.

Therefore, from part (a) we have $b - (n-1)$ linearly independent elements of $\mathrm{Ker}(\partial)$ and from part (b) we have $n-1$ linearly independent elements of $\mathrm{Im}(\partial^*)$. Together they are b linearly independent elements of $\mathrm{Ker}(\partial) \oplus \mathrm{Im}(\partial^*)$. But by above lemma we have that $\dim(\mathrm{Ker}(\partial) \oplus \mathrm{Im}(\partial^*)) = \dim(C_1(\mathcal{G})) = b$. Thus we have that these b elements are a basis for $\mathrm{Ker}(\partial) \oplus \mathrm{Im}(\partial^*)$ and hence the elements from part (a) are a basis for $\mathrm{Ker}(\partial)$ and the elements from part (b) are a basis of $\mathrm{Im}(\partial^*)$. So we actually have links induce a basis of $\{\mathcal{C}_k | 1 \leq k \leq l\}$ of $\mathrm{Ker}(\partial)$ and twigs induce a basis of $\{\mathcal{C}^k | l+1 \leq k \leq b\}$ of $\mathrm{Im}(\partial^*)$. □

Corollary 4.1 *A hybrid input-output representation of the Dirac structure $Ker(\partial) \oplus Im(\partial^*)$ is induced by the above basis.*

Proof By property (i) of the above proposition on Dirac structures we have that the basis elements of $\mathrm{Im}(\partial^*)$ are constraint one-forms on $C_1(\mathcal{G})$. That is, $\langle \mathcal{C}^k, i \rangle = 0$, $l+1 \leq k \leq b$ give $n-1$ constraint equations on $i \in C_1(\mathcal{G})$ that indicate when $i \in \mathrm{Ker}(\partial)$. Writing $\mathcal{C}^k$ as a row vector with respect to the standard basis of $C^1(\mathcal{G})$ and i as a column vector with respect to the standard basis of $C_1(\mathcal{G})$ we have $\langle \mathcal{C}_k, i \rangle = \mathcal{C}_k \cdot i$ for each k, $l+1 \leq k \leq b$. This gives

$$i \in \mathrm{Ker}(\partial) \Leftrightarrow \begin{bmatrix} \longleftarrow \mathcal{C}^{l+1} \longrightarrow \\ \vdots \\ \longleftarrow \mathcal{C}^{b} \longrightarrow \end{bmatrix} \begin{bmatrix} i_1 \\ \vdots \\ i_b \end{bmatrix} = 0$$

$$\Leftrightarrow \quad \left[I \vdots R \right] \begin{bmatrix} i_{\mathcal{T}} \\ i_{\mathcal{L}} \end{bmatrix} = 0 \quad \Leftrightarrow \quad \begin{bmatrix} i_{\mathcal{T}} \\ i_{\mathcal{L}} \end{bmatrix} = \begin{bmatrix} -R \\ I \end{bmatrix} i_{\mathcal{L}}$$

where the identity matrix I and the matrix R have the appropriate size. (The matrix $-R$ here is B in section three.) We also note that this is equivalent to writing KCL with respect to the standard bases of $C_1(\mathcal{G})$.

Similarly, the basis elements of $\mathrm{Ker}\partial$ are constraint vectors on $C^1(\mathcal{G})$. That is, $\langle v, \mathcal{C}_k \rangle = 0$, $1 \leq k \leq l$, gives l constraint equations on $v \in C^1(\mathcal{G})$ that indicate when $v \in \mathrm{Im}(\partial^*)$. Writing $\mathcal{C}_k$ and v with respect to the standard basis as above we have $\langle v, \mathcal{C}_k \rangle = v \cdot \mathcal{C}_k$ for each k, $1 \leq k \leq l$. This gives

$$v \in \mathrm{Im}(\partial^*) \Leftrightarrow \begin{bmatrix} \longleftarrow \mathcal{C}_1 \longrightarrow \\ \vdots \\ \longleftarrow \mathcal{C}_l \longrightarrow \end{bmatrix} \begin{bmatrix} v_1 \\ \vdots \\ v_b \end{bmatrix} = 0$$

$$\Leftrightarrow \quad \left[Q \vdots I \right] \begin{bmatrix} v_{\mathcal{T}} \\ v_{\mathcal{L}} \end{bmatrix} = 0 \quad \Leftrightarrow \quad \begin{bmatrix} v_{\mathcal{T}} \\ v_{\mathcal{L}} \end{bmatrix} = \begin{bmatrix} I \\ -Q \end{bmatrix} v_{\mathcal{T}}$$

where the identity matrix I and the matrix Q have the appropriate sizes. (The matrix Q here is negative of the one in section three.) We also note that this is equivalent to writing KVL with respect to the standard bases of $C^1(\mathcal{G})$.

Recalling that the sum of the voltages drops around any closed loop is zero, we consider the same closed loops that were constructed in (a) of the proof of the above theorem we have that

$$\begin{aligned} & 0 = v_1 + \sum_{t=l+1}^{b} q_{1t} v_t, \quad \ldots \quad 0 = v_l + \sum_{t=l+1}^{b} q_{lt} v_t \\ & \Rightarrow [v_1, \ldots, v_l]^T = -Q[v_{l+1}, \ldots, v_b]^T \\ & \Rightarrow v_{\mathcal{L}} = -Q v_{\mathcal{T}}. \end{aligned}$$

Now suppose that $[v_1, \ldots, v_b]^T = \sum_{k=1}^{b} v_k [x_{i_k}, x_{j_k}]^*$ and $[\tilde{v}_{l+1}, \ldots, \tilde{v}_b]^T = \sum_{k=l+1}^{b} \tilde{v}_k \mathcal{C}^k$ represent the same current in $\mathrm{Im}(\partial^*)$. Setting them equal we have

$$\begin{aligned}
&\sum_{k=1}^{b} v_k [x_{i_k}, x_{j_k}]^* = \sum_{k=l+1}^{b} \tilde{v}_k \mathcal{C}^k \\
\Rightarrow &\sum_{m=1}^{l} v_m [x_{i_m}, x_{j_m}]^* + \sum_{k=l+1}^{b} v_k [x_{i_k}, x_{j_k}]^* \\
&= \sum_{k=l+1}^{b} \tilde{v}_k \left([x_{i_k}, x_{j_k}]^* + \sum_{t=1}^{l} r_{kt} [x_{i_t}, x_{j_t}]^* \right) \\
\Rightarrow &\; v_k = \tilde{v}_k, \quad l+1 \le k \le b, \quad \text{and} \\
&\; v_m = \sum_{k=l+1}^{b} \tilde{v}_k r_{km}, \quad 1 \le m \le l.
\end{aligned}$$

The last equation gives us that $[v_1, \ldots, v_l]^T = R^T [v_{l+1}, \ldots, v_b]^T$, or $v_{\mathcal{L}} = R^T v_{\mathcal{T}}$. Thus we have $R^T = -Q$. By combining this with the above equations, and by relabeling Q as $-Q$, we obtain Eq. (1), the hybrid input-output representation of the Dirac structure on the vector space $C_1(\mathcal{G})$. □

The Dirac structure $\mathrm{Ker}(\partial) \oplus \mathrm{Im}(\partial^*)$ obtained from Kirchhoff's laws is the basis independent formulation of the Dirac structure originally used by Bloch and Crouch. We define the Dirac structure $\mathcal{D}$ on M by $\mathcal{D}(m) = \mathrm{Ker}(\partial) \oplus \mathrm{Im}(\partial^*), \ \forall m \in M$ which allows the implicitly defined Hamiltonian formulation of circuit dynamics, as specified by $(\dot{x}, dH) \in \mathcal{D}$, to be expressed independent of basis.

References

1. Blankenstein, G., van der Schaft, A.J.: Symmetry and reduction in implicit generalized Hamiltonian systems. Rep. Math. Phys. **47**, 57–100 (2001)
2. Bloch, A., Crouch, P.: Representations of Dirac structures on vector spaces and nonlinear L-C circuits. Proc. Sympos. Pure Math. **64**, 103–117 (1999)
3. Chua, L., Desoer, C., Kuh, E.: Linear and Nonlinear Circuits. McGraw-Hill, Boston (1987)
4. Courant, T.J.: Dirac manifolds. Trans. Amer. Math. Soc. **319**, 631–661 (1990)
5. Dalsmo, M., van der Schaft, A.J.: On representations and integrability of mathematical structures in energy conserving physical systems. SIAM J. Control Optim. **37**, 54–91 (1998)
6. Dorfman, I.: Dirac structures of integrable evolution equations. Phys. Lett. A **125**, 240–246 (1987)
7. Fortney, J.P.: Dirac structures in pseudo-gradient systems with an emphasis on electrical networks. IEEE Trans. Circuits Syst. I. Regul Pap. 57 (2010)
8. Maschke, B.M., van der Schaft, A.J.: An intrinsic Hamiltonian formulation of the dynamics of L-C circuits. IEEE Trans. Circuits Syst. I. Regul Pap. **42**, 73–82 (1995)
9. Smale, S.: On the mathematical foundations of electrical circuit theory. J. Differ. Geom. **7**, 193–210 (1972)

10. van der Schaft, A.J.: Implicit Hamiltonian systems with symmetry. Rep. Math. Phys. **41**, 203–221 (1998)
11. Yoshimura, H., Marsden, J.E.: Dirac structures in Lagrangian mechanics Part I: Implicit Lagrangian systems. J. Geom. Phys. **57**, 133–156 (2006)
12. Yoshimura, H., Marsden, J.E.: Dirac structures in Lagrangian mechanics Part II: Variational structures. J. Geom. Phys. **57**, 209–250 (2006)

On a New Class of Variational Problems

Hichem Hajaiej

Abstract In this paper, we study a class of discrete fractional variational problems modeling some phenomena arising in electron transports in bipolymers like organic semi-conductors, molecular crystals and DNA. For some non-linearities covered by our class of functionals, the underlying PDEs model Ferro-Magnets and spin glasses. They also appear in the approximation of the Bose-Einstein condensation.

Keywords Fractional · Laplacian · Discrete · Minimization · Constraints · Inequalities

1 Introduction

Fractional differential equations involving the fractional Laplacian $(-\Delta)^s$, $0 < s < 1$, arise in many fields; medicine, geology, hydrology, mathematical physics and mathematical biology; [1, 3, 4, 6–11] and references therein. The model case of nonlinear fractional Schrödinger equations describing the above problems is:

$$\left.\begin{array}{l} i\partial_t \Phi(t,x) + (-\Delta)^s \Phi(t,x) + f(|x|, |\Phi(t,x)|) = 0 \\ \Phi(0,x) = \Phi_0(x). \end{array}\right\} \tag{1.1}$$

There are many interesting underlying problems related to (1.1), especially the study of existence and uniqueness of the solutions. The ones which are of particular interest are the, so called, standing waves, i.e., $\Phi(t,x) = e^{-i\lambda t}u(x)$. Such Φ solves (1.1) if and only if u is a solution of the following fractional elliptic equation:

$$(-\Delta)^s u + f(|x|, |u|) + \lambda u = 0, \tag{1.2}$$

where λ is a Lagrange multiplier.

H. Hajaiej (✉)
Institute of Mathematical Sciences, New York University, 1555 Century Avenue, Pudong New District, Shanghai 200122, China
e-mail: hh62@nyu.edu; hichem.hajaiej@gmail.com

T. Abualrub et al. (eds.), *Mathematics Across Contemporary Sciences*,
Springer Proceedings in Mathematics & Statistics 190,
DOI 10.1007/978-3-319-46310-0_6

Ground state solutions of (1.2) are obtained by minimizing the following fractional constrained variational problem:

$$I_c = \inf\{E(u) : u \in S_c\} \tag{1.3}$$

$$E(u) = \frac{1}{2}|\nabla_s u|_2^2 - \int_{\mathbb{R}^N} F(|x|, u)dx,$$

$$|\nabla_s u|_2^2 = C_{N,s} \int_{\mathbb{R}^N} \int_{\mathbb{R}^N} \frac{|u(x) - u(y)|^2}{|x - y|^{N+2s}}\, dxdy,\ C_{N,s} = (2^{s-1} s/\pi^{N/2}\, \Gamma(\frac{N+s}{2})/\Gamma(1 - \frac{s}{2}))$$

$$F(r, t) = \int_0^t f(r, p)dp \quad \text{and} \quad S_c = \left\{ u \in H^s(\mathbb{R}^N)) \int_{\mathbb{R}^N} u^2 = c^2 \right\}.$$

Solutions of (1.3) are the best candidates to guarantee the orbital stability of the corresponding standing waves, which is one of the most important properties that one has to investigate for (1.1) due to its tight connections to applications of this fractional nonlinear Schrödinger equation.

(1.3) constitutes in itself a branch of nonlinear analysis as shown by the numerous articles dedicated to this topic during the last decades. In many relevant cases, it is crucial to derive some qualitative and quantitative properties of the minimizers of (1.3) before studying their orbital stability. However in some situations, one has to model (1.3) and then to numerically get the desired informations: This is the general scheme to achieve this goal.

Step 1: First functions and functionals are discretized by replacing $\mathbb{R}^N$ by centered balls in zero, by replacing derivatives by finite differences and by restricting the function spaces to finite dimensional spaces.

Step 2: The next step consists of analyzing the solutions of (1.3) of finite difference/finite element problems; apriori bounds, asymptotics, symmetry, radiality...

Step 3: In the last step, algorithms are designed to compute numerical approximations of solutions of the discretized problems.

In this paper, we consider a one-dimensional lattice $h\mathbb{Z}$ with a mesh size $h > 0$. We denote $x_m = hm$ with $m \in \mathbb{Z}$ and $\Phi_h : \mathbb{R} \times h\mathbb{Z} \to \mathbb{C}$. Then (1.1) becomes in the discrete setting:

$$\begin{cases} i\, \frac{d}{dt} \Phi_h(t, x_m) = h \sum_{n \neq m} \frac{\Phi_h(t, x_m) - \Phi_h(t, x_n)}{|x_m - x_n|^{1+2s}} + \\ f(|x_m|, |\Phi_h(x_m)|) \\ \Phi_h(0, x_m) = \Phi_h^0(x_m). \end{cases} \tag{1.4}$$

Here the fractional power s can also be interpreted as a fixed parameter controlling the decay behavior of the lattice interaction; [3, 11, 12].

In [3], the authors have considered the cubic nonlinear Schrödinger equation (1.4) in the special case $f(r,t) = |t|^2 t$; $N = 1$. More precisely, they have studied (see 2.1):

$$\begin{cases} i\,\frac{d}{dt}u_h(t,x_m) = \frac{1}{\beta(h)}\sum_{n\neq m} J_{|n-m|}\,[u_h(t,x_m) - u_h(t,x_n)] \\ \pm|u_h(t,x_m)|^2\,u_h(t,x_m) \\ u_h(0,x_m) = u_h(x_m), \end{cases}$$

where $x_m = mh$, $J_n = |n|^{1-2s}$ and $\beta(h) = h^{2s}$.

(1.4) can also be viewed as the model of quantum particles on a lattice with repulsive or self-interactions (depending on the sign of f), [3]. When $h \to 0^+$, one does expect that Φ_h tends in some sense to the solution $\Phi(t,x)$ of (1.1).

When this happens, the justification that (1.4) models perfectly (1.1) is excellent. These kind of simulations are very complicated because of the nonlocal properties of the operator involved. However, some promising progress has been recently made in this direction [4].

Now as indicated in Step 1 above, let Ω be a centered interval in zero and consider a regular step size ($x_m = hm$, $x_{m+1} - x_m = h$).

Then discrete standing waves of (1.4) solve the following problem:

$$\sum_{x\in\Omega_h} \frac{u(x+h) + u(x-h) - 2u(x)}{h^{1+2s}} + f(|x|,|u|) + \lambda u = 0. \tag{1.5}$$

The corresponding discretized energy functional is:

$$J_h(u) = \sum_{x\in\Omega_h} \{\frac{1}{2}|\nabla_h^s u|_2 - F(|x|,u)\}h,$$

where Ω_h consists of the points of regular mesh of steps h that belong to the centered interval Ω and:

$$|\nabla_h^s u|_2^2 = \sum_{k,\ell} \frac{|u(x_k) - u(x_\ell)|^2}{|k-\ell|^{1+2s}h^{2s}}. \tag{1.6}$$

Then an equivalent formulation of the discretized constrained variational problem is:

$$I_h^c = \inf\left\{E_h(u_h) : \sum_{m\in\mathbb{Z}} u_h^2(x_m)h = c^2\right\} \tag{1.7}$$

$$E_h(u_h) = \frac{1}{2}|\nabla_h^s u_h|_2^2 h - \sum_{m\in\mathbb{Z}} F(|x_m|, u_h(x_m))h$$

for u_h a lattice function defined on $h\mathbb{Z}$. The more one knows about the qualitative properties of solution of (1.7), the more efficient and less difficult is the design of algorithms.

Very recently, the author has established optimal assumptions under which all the minimizers of the continuous constrained variational problem (1.3) are Schwarz symmetric (i.e. radial and radially decreasing); [3.4]. His method hinges on the following rearrangement inequalities:

$$|\nabla_s u^*|_2 \leq |\nabla_s u|_2 \tag{1.8}$$

$$|u|_2 = |u^*|_2 \tag{1.9}$$

$$\int_{\mathbb{R}^N} F(|x|, u)dx \leq \int_{\mathbb{R}^N} F(|x|, u^*)dx \tag{1.10}$$

where u^* is the Schwarz rearrangement of u, [10].

The main goal of the present paper is to extend the results of [4] to the discrete case. The key step to reach this objective is to prove (1.8) to (1.10) in this setting:

$$\sum_{k,\ell\in\mathbb{Z}} \frac{|u_h^*(x_k) - u_h^*(x_\ell)|^2}{|k-\ell|^{1+2s}} \leq \sum_{k,\ell\in\mathbb{Z}} \frac{|u_h(x_k) - u_h(x_\ell)|^2}{|k-\ell|^{1+2s}} \tag{1.11}$$

$$\sum_{k\in\mathbb{Z}} u_h^2(x_k) = \sum_{k\in\mathbb{Z}} (u_h^*)^2(x_k) \tag{1.12}$$

$$\sum_{k\in\mathbb{Z}} F(|x_k|, u_h(x_k)) \leq \sum_{k\in\mathbb{Z}} F(|x_k|, u_h^*(x_k)) \tag{1.13}$$

(u_h^*) denotes the discrete Schwarz symmetrization of u_h, [Definition 2.3]. Amazingly the situation in the discrete setting is very intricate and Challenging. This is due to the appearance of some unexpected phenomena in this kind of problems. In fact Mc Kenna and Reichel have proved in [5] that critical points of a class of discretized variational problems do not generally inherit the same symmetry properties as the critical points of the corresponding continuous problems. More precisely, they were able to show that unlike the continuous case, there are spurious situations in the discrete one, i.e., solutions with no relation to the ones in the continuous setting.

In this work, we will show that such situations cannot occur in our context thanks to the rearrangement inequalities (1.11)–(1.13). Moreover we will establish cases of equality in (1.11) and (1.13). Therefore, we are able to determine hypotheses on F and s for which all the minimizers of (1.7) are Schwarz symmetric (Theorem 4.1). Proving that solutions of the discretized constrained variational problem (1.7) inherit symmetry and monotonocity properties is extremely important for the design of numerical Scheme; [2]. It also implies that we need only to solve numerically

these problems on a quarter region instead of the full one. this considerably cuts down the computational cost.

Our paper is organized as follows. In the next section, we give some definitions and preliminary results. In Sect. 3, we will prove (1.11)–(1.13). In the last section, we will show that these inequalities are extremely helpful to prove that all the minimizers of (1.7) are Schwarz symmetric.

From now on h is a fixed stepsize and for $m \in \mathbb{Z}$ $x_m = hm$.

2 Notations and Preliminaries

2.1 Discrete Function Spaces

Definition 2.1 For sequences $u_h, v_h : h\mathbb{Z} \to \mathbb{R}$, we define:

$$(u_h, v_h)_{L_h^2} = h \sum_{m\in\mathbb{Z}} u_h(x_m)v_h(x_m),$$
$$\|u_h\|_{L_h^2}^2 = h \sum_{m\in\mathbb{Z}} u_h^2(x_m).$$

And more generally for $1 \le p < \infty$, we define:

$$\|u_h\|_{L_h^p} = (h \sum_{m\in\mathbb{Z}} |u_h(x_m)|^p)^{1/p} \tag{2.1}$$

$$L_h^p = \{u_h \in \mathbb{R}^{h\mathbb{Z}}, \|u\|_{L_h^p} < +\infty\} \text{ is a complete Banach space}$$
$$\mathbb{R}^{h\mathbb{Z}} = \{f : h\mathbb{Z} \to \mathbb{R}\}.$$

For $u_h \in L_h^2$, we define its Fourier transform $\hat{u}_h : [-\pi, \pi] \to \mathbb{C}$ by

$$\hat{u}_h(k) = \frac{1}{\sqrt{2\pi}} \sum_{m\in\mathbb{Z}} u_h(x_m)e^{-imk}.$$

Since $u_h \in L_h^2$, it follows that $\hat{u}_h \in L^2([-\pi, \pi])$ and that

$$u_h(x_m) = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} \hat{u}_h(k)e^{imk}dk.$$

Using Parseval's identity, we obtain:

$$(u_h, v_h)_{L_h^2} = h \int_{-\pi}^{\pi} \hat{u}_h(k)\hat{v}_h(k)dk.$$

Thanks to this observation, we can introduce the following fractional Sobolev norm for lattice functions $u_h \in L_h^2$: Let $0 \leq s \leq 1$ be given, we define $\|u_h\|_{H_h^s}$ for $u_h \in L_h^2$ by setting

$$\|u_h\|_{H_h^s}^2 = h \int_{-\pi}^{\pi} (1 + h^{-2s}|k|^{2s})|\hat{u}_h(k)|dk. \tag{2.2}$$

Obviously $\|u_h\|_{H_h^0} = \|u_h\|_{L_h^2}$ and $\|u\|_{H_h^s} < \infty$ for any $u_h \in L_h^2$.

Definition 2.2 The L_h^2 norm of the fractional gradient of a lattice function u_h is defined by:

$$\|\nabla_s u_h\|_{L_h^2}^2 = \sum_{k\in\mathbb{Z}} \sum_{\ell\in\mathbb{Z}} \frac{|u_h(x_k) - u_h(x_\ell)|^2}{|x_k - x_\ell|^{1+2s}} \frac{1}{h^{2s+1}}. \tag{2.3}$$

2.2 Discrete Functional Inequalities

First let us recall, for the convenience of the reader, that Sobolev embeddings are still valid in the discrete setting. In our context, we will need the following ones: For $s < \frac{1}{2}$, we have:

$$H_h^s \text{ is continuously embedded in } L_h^{p+2} \text{ for } p < 4s \tag{2.4}$$

$H_h^s(\widetilde{\mathbb{Z}})$ is compactly embedded in $L_h^{\ell+2}(\widetilde{\mathbb{Z}})$ for $p < 4s$ and $\widetilde{\mathbb{Z}}$ is a bounded lattice of $\mathbb{Z}$. (2.5)

Discrete fractional Gagliardo-Nirenberg Inequality:

Following the proof of [3, Lemma 3.2], we can easily prove that:

$$\|u_h\|_{L_h^{\ell+2}} \leq K \|\nabla_s u_h\|_{L_h^2}^{\theta} \|u\|_{L_h^2}^{1-\theta} \tag{2.6}$$

for any $\ell < 4s$, $\theta = \dfrac{\ell}{2s(\ell+2)}$.

2.3 Schwarz Symmetrization in $h\mathbb{Z}$

Definition 2.3 If $u_h : h\mathbb{Z} \to \mathbb{R}_+$ is bounded from above, the discrete Schwarz symmetrization of u_h is the unique function $u_h^* : h\mathbb{Z} \to \mathbb{R}_+$ such that:

1. For all $k \geq 0 : u_h^*(x_k) \geq u_h^*(x_{-k}) \geq u_h^*(x_{k+1})$.
2. For all $t \in \mathbb{R} : \#\{k \in \mathbb{Z} : u_h^*(x_k) > t\} = \#\{k \in \mathbb{Z} : u_h(x_k) > t\}$.

We can define explicity u_h^* by the following formula:

$$u_h^*(x_k) = \begin{cases} \sup\{t \in \mathbb{R} : \#\{\ell \in \mathbb{Z} : u_h(x_\ell) > t\} \leq 2|k| + 1 & \text{if } k \leq 0 \\ \sup\{t \in \mathbb{R} : \#\{\ell \in \mathbb{Z} : u_h(x_\ell) > t\} \leq 2k & \text{if } k \geq 0 \end{cases}.$$

The construction of u_h^* goes thus by taking for $u_h^*(0)$ the maximum value of u_h, for $u_h^*(h)$ the second largest value of u_h, $u_h^*(-h)$ the third one and so on.

Definition 2.4 A function $u_h : h\mathbb{Z} \to \mathbb{R}_+$ is admissible if $\#\{\ell \in \mathbb{Z} : u_h(x_\ell) \geq t\} < \infty$ for all $t > 0$.

Remark If $u_h \in L_h^p$ for $1 < p < \infty$ and u is non−negative, then u is admissible.

Lemma 2.5 *For every $t > 0$ then:*

$$\#\{k \in \mathbb{Z}, u_h(x_k) = t\} = \#\{k \in \mathbb{Z} : u_h^*(x_k) = t\}.$$

Proposition 2.6 (Cavalieri's principle in the discrete setting
Let $f : \mathbb{R}_+ \to \mathbb{R}_+$, $u_h : h\mathbb{Z} \to \mathbb{R}_+$ be admissible.
If $f(0) = 0$, then:

$$\sum_{k\in\mathbb{Z}} f(u_h(x_k)) = \sum_{k\in\mathbb{Z}} f(u_h^*(x_k)). \tag{2.7}$$

Proof We have:

$$\sum_{k\in\mathbb{Z}} f(u_h(x_k)) = \sum_t f(t)\#\{k \in \mathbb{Z} : u_h(x_k) = t\} + f(0)\,\#\{k \in \mathbb{Z} : u_h(x_k) = 0\}.$$

The analogous happens for u_h^* and we can conclude using Lemma 2.5 and the fact that $f(0) = 0$.

Corollary 2.7 *If $u_h \in L_h^p$ and u_h is non-negative, then $u_h^* \in L_h^p$ and*

$$\|u_h\|_{L_h^p} = \|u_h^*\|_{L_h^p}. \tag{2.8}$$

Now we need to prove some preliminary results about approximation of a Schwarz rearrangement u_h^* of u_h by repeated polarizations. This will be crucial to establish the discrete symmetrization inequalities (1.11) and (1.13). We will use some ideas and techniques developed by the author in [10] in the continuous setting. Let us first define the polarization in the discrete setting.

Definition 2.8 The set of semi finite open intervals whose boundary is contained in $h\mathbb{Z}/2$ is denoted by $\mathcal{H}^h = [ah/2, +\infty[a \in \mathbb{Z}$. For $H \in \mathcal{H}^h$, the reflexion with respect to ∂H is denoted by σ_H. Note that if $H \in \mathcal{H}^h$, $\sigma_H(h\mathbb{Z}) = h\mathbb{Z}$.

Definition 2.9 The polarization of $u_h : h\mathbb{Z} \to \mathbb{R}_+$ with respect to $H \in \mathcal{H}^h$ is the function $u_h^H : h\mathbb{Z} \to \mathbb{R}_+$ defined by:

$$u_h^H(x_k) = \begin{cases} \max\{u_h(x_k), u_h(\sigma_H(x_k))\} & \text{if } x_k \in h\mathbb{Z} \cap H \\ \min\{u_h(x_k), u_h(\sigma_H(x_k))\} & \text{if } x_k \in h\mathbb{Z}\backslash H. \end{cases}$$

Proposition 2.10 *Let $u_h : h\mathbb{Z} \to \mathbb{R}_+$ and $v_h : h\mathbb{Z} \to \mathbb{R}_+$ be admissible. If $u_h v_h \in L_h^1$ and $u_h^H v_h^H \in L_h^1$, then*

$$\sum_{k\in\mathbb{Z}} u_h(x_k)v_h(x_k) \leq \sum_{k\in\mathbb{Z}} u_h^H(x_k)v_h^H(x_k). \tag{2.9}$$

Moreover if $v_h = v_h^H$ and there is equality in (2.9), then

$$u_h^H(x_k) = u_h(x_k) \text{ and } u_h^H(\sigma_H(x_k)) = u_h(\sigma_H(x_k))$$

for any $x_k \in h\mathbb{Z} \cap H$ such that $v_h(x_k) > v_h(\sigma_H(x_k))$.

Proof For any $x_k \in h\mathbb{Z} \cap H$:

$$\begin{array}{l} u_h(x_k)v_h(x_k) + u_h(\sigma_H(x_k))v_h(\sigma_H(x_k)) \leq \\ u_h^H(x_k)v_h^H(x_k) + u_h^H(\sigma_H(x_k))v_h^H(\sigma_H(x_k)). \end{array} \tag{2.10}$$

Summing these inequalities and noticing that $u_h^H(x_k)v_h^H(x_k) = u_h(x_k)v_h(x_k)$ for $x_k \in h\mathbb{Z} \cap /H$, we obtain (2.9).

In case, we have equality in (2.9), we have also equality in (2.10) for $x_k \in h\mathbb{Z} \cap H$. But by our assumption on v, this means that $u_h(\sigma_H(x_k)) \leq u_h(x_k)$ for every $x_k \in h\mathbb{Z} \cap H \Rightarrow u_h = u_h^H$.

For the latter, we will need two particular type of polarizations.

Definition 2.11 $H_+ =]0, +\infty[$, $H_- =]-\infty, \frac{h}{2}[$ so that:

$$u_h^{H_+}(x_k) = \begin{cases} \max(u_h(x_k), u_h(x_k)) & \text{if } k \geq 0 \\ \min(u_h(x_k), u_h(x_k)) & \text{if } k \leq 0 \end{cases} \tag{2.11}$$

$$u_h^{H_-}(x_k) = \begin{cases} \max(u_h(x_k), u_h(x_{1-k})) & \text{if } k \leq 0 \\ \min(u_h(x_k), u_h(x_{1-k})) & \text{if } k \geq 1 \end{cases}. \tag{2.12}$$

The aim of the following paragraph is to show that u_h^* is a limit of iterated polarization. For $u_h : h\mathbb{Z} \to \mathbb{R}_+$, define $T_h u_h = (u_h^{H_-})^{H_+}$. Iterating T_h, one gets $u, u^{H_-H_+}, u^{H_-H_+H_-H_+}$ we shall prove that $T_h^n u_h$ goes to u_h^* as $n \to \infty$.

Proposition 2.12 *The sequence $(T_h^n u_h)_{n\geq 0}$ is precompact in (X_h, d), where the metric d is defined by*

$$d(u_h, v_h) = \sum_{k\in\mathbb{Z}} \frac{|u_h(x_k) - v_h(x_k)|}{1 + 2^{|k|}|u_h(x_k) - v_h(x_k)|}.$$

Moreover, for any cluster point v_h:

$$\#\{k \in \mathbb{Z} : v_h(x_k) > t\} = \#\{k \in \mathbb{Z} : u_h(x_k) > t\}$$

for any u_h admissible and any $t > 0$.

Proof $X_h = \{u_h : h\mathbb{Z} \to \mathbb{R}_+\}$ endowed with the metric d which is a complete metric space:

$$u_{h,n}(x_k) \to x_h(x_k)\ \forall\ k \in \mathbb{Z} \Leftrightarrow d(u_{h,n}, u_h) \to 0.$$

Now first observe that by induction on $n \geq 0$, we certainly have that

$$\inf_{|\ell| \leq |k|} u_h(x_\ell) \leq (T_h^n u_h)(x_k) \leq \sup_{|\ell| \geq |k|} u_h(x_k).$$

The precompactness then follows by a standard diagonal argument.

Let v_h be a cluster point of the sequence $(T_h^n u_h)$. Assume that $T_h^{n_j} u_h(x_k) \to v_h(x_k)$ $\forall\ k \in \mathbb{Z}$. Then:

$$\#\{k \in \mathbb{Z} : u_h^H(x_k) > t\} = \#\{k \in \mathbb{Z} : u_h(x_k) > t\};\ \forall t > 0.$$

Therefore for any $n \geq 0$:

$$\{k \in \mathbb{Z} : (T_h^n u_h)(x_k) > t\} = \#\{k \in \mathbb{Z} : u_h(x_k) > t\};\ \forall t > 0.$$

This implies that:

$$\#\{k \in \mathbb{Z} : v_h(x_k) > t\} \leq$$

$$\leq \underline{\lim}_{\ell \to \infty} \#\{k \in \mathbb{Z} : u_h(x_k) > t\};\ \forall t > 0.$$

For the converse inequality, let $A > 0$ and note that:

$$\#\{k \in \mathbb{Z} : |k| \leq A \text{ and } u_h^H(x_k) \geq t + \frac{1}{A}\} \geq$$

$$\geq \#\{k \in \mathbb{Z} : |k| \leq A \text{ and } u_h(x_k) \geq t + \frac{1}{A}\},$$

hence

$$\#\{k \in \mathbb{Z} : |k| \leq A \text{ and } (T_h^n u_h)(x_k) \geq t + \frac{1}{A}\} \geq$$

$$\#\{k \in \mathbb{Z} : |k| \leq A \text{ and } u_h(x_k) \geq t + \frac{1}{A}\};\ \forall t > 0.$$

And we can conclude by letting A tend to infinity that

$$\#\{k \in \mathbb{Z} : v_h(x_k) \geq t\} = \#\{k \in \mathbb{Z} : u_h(x_k) \geq t\}; \ \forall t > 0.$$

Proposition 2.13 *If $u_h : h\mathbb{Z} \to \mathbb{R}_+$ is admissible, then:*

$$T_h^n u_h(x_k) \to u_h^*(x_k) \text{ for every } k \in \mathbb{Z}.$$

Proof We know by Proposition 2.12 that $(T_h^n u_h)_{n\geq 0}$ is precompact in (X_h, d). Assume that $T_h^{n_j} u_h(x_k) \to v_h(x_k)$ for every $k \in \mathbb{Z}$. We need to prove that $v_h = u_h^*$.
First not that for any $k \geq 0 : T_h^n u_h(x_k) \geq T_h^n u_h(x_{-k})$ so $v(x_k) \geq v(x_{-k})$.
For $k \geq 0$, $\ell \in \mathbb{Z}$ set $w_\ell(x_k) = \begin{cases} 1 & \text{if } \ell \leq k \\ 0 & \text{if } \ell > k \end{cases}$
$w^{H_-} = w$, and by the previous proposition

$$\sum_{\ell\in\mathbb{Z}} (T_h^{n_j} u_h)^{H_-}(x_\ell) w_k(x_\ell) \leq \sum_{\ell\in\mathbb{Z}} v_h(x_\ell) w_k(x_\ell).$$

Letting $j \to \infty$, one gets

$$\sum_{\ell\in\mathbb{Z}} v_h^{H_-}(x_\ell) w_k(x_\ell) \leq \sum_{\ell\in\mathbb{Z}} v_h(x_\ell) w_h(x_\ell)$$

since $\sigma_H(x_{-k}) = x_{k+1}$ one has $w_k(x_{-\ell}) = w_k(\sigma_H(x_{-\ell}))$.

Thus using cases of equality established in Proposition 2.10, we can conclude.

3 Discrete Symmetrization Inequalities

Definition 3.1

- A function $G : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is supermodular if:

$$G(x + x_0, y + y_0) + G(x, y) \geq G(x, y + y_0) + G(x + x_0, y) \tag{3.1}$$

  for any $x, y \in \mathbb{R}, x_0, y_0 > 0$.
- A function $K : |h\mathbb{Z}| \times \mathbb{R}_+ \to \mathbb{R}$ is $|h\mathbb{Z}|$ supermodular ($|h\mathbb{Z}| = h\mathbb{N}$) if

$$K(|(m + m_0)h|, y + y_0) + K(|mh|, y) \geq K(|mh|, y + y_0) + K(|(m + m_0)h|, y) \tag{3.2}$$

  for any $m \in \mathbb{Z}, y \in \mathbb{R}_+, m_0 \in \mathbb{N}, y_0 > 0$.

We say that G is strictly supermodular if (3.1) holds true with a strict sign. The function K is $h\mathbb{Z}$ strictly supermodular when (3.2) holds with a strict sign.

In the sequel, we will make a frequent use of the following property : If u_h is admissible:

$$u_h = u_h^* \Leftrightarrow u_h = u_h^H \ \ \forall\, \sigma_H \in H. \tag{3.3}$$

Theorem 3.1

(i) Let $G : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ be a supermodular function then:

$$\sum_{k\in\mathbb{Z}} G(u_h(x_k), v_h(x_k)) \le \sum_{k\in\mathbb{Z}} G(u_h^H(x_k), v_h^H(x_k)) \tag{3.4}$$

for any admissible functions u_h and v_h.
If $v_h = v_h^H$ and G is strictly supermodular, then equality holds true in (3.4) if and only if $u_h = u_h^H$.

(ii) If in addition $G(.,.)$ is continuous, non-decreasing with respect to each variable and $\sum_{k\in\mathbb{Z}} G(u_h^*(x_k), v_h^*(x_k)) < \infty$ then we have

$$\sum_{k\in\mathbb{Z}} G(u_h(x_k), v_h(x_k)) \le \sum_{k\in\mathbb{Z}} G(u_h^*(x_k), v_h^*(x_k)) \tag{3.5}$$

for any admissible function u_h.
If $v_h = v_h^*$ and G is strictly supermodular, then equality holds in (3.5) if and only if $u_h = u_h^*$.

Proof

(i) By the supermodularity of G, we certainly have for any $x_k \in h\mathbb{Z} \cap H$ that:

$$\begin{aligned} G(u_h(x_k), v_h(x_k)) \ &+G(u_h(\sigma_H(x_k)), v_h(\sigma_H(x_k)) \\ &\le G(u_h^H(x_k), v_h^H(x_k)) + G(u_h^H(\sigma_H(x_k)), v_h^H(\sigma_H(x_k)). \end{aligned} \tag{3.6}$$

Summing up this inequality and noticing that $u_h^H(x_k)v_h^H(x_k) = u_h(x_k)v_h(x_k)$ for any $x_k \in h\mathbb{Z} \cap \partial H$, we obtain (3.4).
Now in case we have equality in (3.3), we will also have equality in (3.6) for any $x_k \in h\mathbb{Z} \cap \partial H$ by the strict supermodularity of G. Now since we are also assuming that $v_h = v_h^H$, it follows that

$$u_h(\sigma_H(x_k)) \le u_h(x_k) \ \ \forall\, x_k \in h\mathbb{Z} \cap H; \ \text{i.e.}\ , u_h = u_h^H. \tag{3.7}$$

(ii) By the continuity and the monotonicity of G, (3.5) follows immediately from (3.4) by applying the Theorem of monotone convergence. More precisely if $(T_{u_h}^n)$ is the sequence of iterated polarizations constructed in Sect. 2, we obviously have:

$$\sum_{k\in\mathbb{Z}} G(u_h(x_k), v_h(x_k)) \leq \sum_{k\in\mathbb{Z}} G(T_{u_h}(x_k), T_{v_h}(x_k)) \leq ... \leq \sum_{k\in\mathbb{Z}} G(T^n_{u_h}(x_k), T^n_{v_h}(x_k)) \tag{3.8}$$

Thus letting n go to infinity, the result follows and we certainly have

$$\sum_{k\in\mathbb{Z}} G(u_h(x_k), v_h(x_k)) \leq \sum_{k\in\mathbb{Z}} G(T_{u_h}(x_k), T_{v_h}(x_k)) \leq ... \leq \sum_{h\in\mathbb{Z}} G(u^*_h(x_k), v_h(x_k)). \tag{3.9}$$

Now if we have equality in (3.5), we will certainly have equality in (3.9):

$$\sum_{k\in\mathbb{Z}} G(u_h(x_k), v_h(x_k)) = \sum_{k\in\mathbb{Z}} G(T_{u_h}(x_k), T_{v_h}(x_k)) = ... = \sum_{k\in\mathbb{Z}} G(u^*_h(x_k), v^*_h(x_k)).$$

But we are supposing that $v_h = v^*_h \Rightarrow v^H_h = v_h \; \forall \, H$. Thus using cases of equality of part (i), it follows that $u_h = u^H_h \; \forall \, H$, which is equivalent to say that $u_h = u^*_h$ by (3.3).

Remark Hypotheses on G used in part (ii) can be relaxed.

In fact, it is sufficient to suppose that G is supermodular and that $\sum_{k\in\mathbb{Z}} G(u^*_h(x_k), 0)$ and $\sum_{k\in\mathbb{Z}} G(0, v^*_h(x_k)) < \infty$, since $\tilde{G}(s_1, s_2) = G(s_1, s_2) - G(s_1, 0) - G(0, s_2)$ satisfies all the assumptions of Theorem 3.1. Therefore the monotonicity of the function with respect to each variable can be removed.

Theorem 3.2 *If F is a function: $h\mathbb{N} \times \mathbb{R}_+ \to \mathbb{R}$ satisfying*

1. *$F(|x_m|, .)$ is continuous for any $m \in \mathbb{Z}$.*
2. *$-F$ is $|h\mathbb{Z}|$ supermodular.*
3. *$\sum_{k\in\mathbb{Z}} F(|x_k|, 0) < \infty$, then*

$$\sum_{k\in\mathbb{Z}} F(|x_k|, u_h(x_k)) \leq \sum_{k\in\mathbb{Z}} F(|x_k|, u^*_h(x_k)) \tag{3.10}$$

for any admissible u_h.

Moreover if $-F$ is strictly $|h\mathbb{Z}|$ supermodular and we have equality in (3.10), then $u_h = u^*_h$.

Proof the proof is identical to the one of the previous result.

Theorem 3.3 (Discrete fractional Polya Szegö inequality). *Let $u_h : h\mathbb{Z} \to \mathbb{R}_+$ be admissible, then:*

$$|\nabla_s u_h|^2_{L^2_h} = \sum_{\ell,k\in\mathbb{Z}} \frac{|u_h(x_k) - u_h(x_\ell)|^2 h}{|kh - \ell h|^{1+2s}} = \frac{1}{h^{2s}} \sum_{\ell,k\in\mathbb{Z}} \frac{|u_h(x_k) - u_h(x_\ell)|^2}{|k-\ell|^{1+2s}}$$
$$\geq \frac{1}{h^{2s}} \sum_{\ell,k\in\mathbb{Z}} \frac{|u_h^*(x_k) - u_h^*(x_\ell)|^2}{|k-\ell|^{1+2s}} = |\nabla_s u_h^*|^2_{L^2_h}. \tag{3.11}$$

If one has equality in (3.9), then $u_h(x_k) = u_h^H(x_k)$ or $= u_h^H(x_k) = u_h(\sigma_H(x_k))$ (i.e. u_h and u_h^* are equal up to a translation).

Proof Let $k, \ell \in \mathbb{Z}$ be such that $\partial H \not\subset (k, \ell)$, set $k' = \sigma_H(k)$ and $\ell' = \sigma_H(\ell)$, then $|k-\ell| = |k'-\ell'| \leq |k-\ell'| = |k'-\ell|$.

Hence

$$\frac{|u_h(x_k) - u_h(x_\ell)|^2}{|k-\ell|^{1+2s}} + \frac{|u_h(x_k) - u_h(x_{\ell'})|^2}{|k-\ell'|^{1+2s}} + \frac{|u_h(x_{k'}) - u_h(x_\ell)|^2}{|k'-\ell|^{1+2s}} +$$

$$\frac{|u_h(x_{k'}) - u_h(x_{\ell'})|^2}{|k'-\ell'|^{1+2s}} = \frac{1}{|k'-\ell|^{1+2s}} P(u_h) + \left(\frac{1}{|k-\ell|^{1+2s}} - \frac{1}{|k'-\ell|^{1+2s}}\right) Q(u_h)$$

where

$$P(u_h) = |u_h(x_k) - u_h(x_\ell)|^2 + |u_h(x_k) - u_h(x_{\ell'})|^2 + |u_h(x_{k'}) - u_h(x_\ell)|^2 +$$

$$|u_h(x_{k'}) - u_h(x_{\ell'})|^2 \text{ and } Q(u_h) = |u_h(x_k) - u_h(x_\ell)|^2 + |u_h(x_{k'}) - u_h(x_{\ell'})|^2.$$

Noticing that $P(u_h) = P(u_h^H)\ \forall\ H \in \mathcal{H}^h$ and that

$$\left(\frac{1}{|k-\ell|^{1+2s}} - \frac{1}{|k'-\ell|^{1+2s}}\right) Q(u_h) \geq \left(\frac{1}{|k-\ell|^{1+2s}} - \frac{1}{|k'-\ell|^{1+2s}}\right) Q(u_h^H)$$

and summing over k and ℓ, enables us to conclude that

$$|\nabla_s u_h|^2_{L^2_h} \geq |\nabla_s u_h^H|^2_{L^2_h}\ \forall\ H \in \mathcal{H}^h.$$

Finally thanks to Proposition 2.12, we certainly have that $|\nabla_s u_h|^2_{L^2_h} \geq |\nabla_s u_h^*|^2_{L^2_h}$ for any admissible u_h.

Cases of equality are obtained in the same way as part (ii) of Theorem 3.1.

4 Discrete Fractional Constained Variational Problem

In this section we will study the following constrained variational problem:

$$I_c^h = \inf\{E_h(u_h) : u_h \in S_c^h\}. \tag{4.1}$$

$$E_h(u_h) = \left\{ \frac{1}{2} |\nabla_s u_h|^2_{L^2_h} - \int F(|x_k|, u_h(x_k)) \right\} h$$

$$S^h_c = \left\{ u_h \in H^s_h : \sum_{k\in\mathbb{Z}} u^2_h(x_k) h = c^2 \right\},$$

where c is a prescribed real number.

Our main result in this section is:

Theorem 4.1 *Let $F : |h\mathbb{Z}| \times \mathbb{R} \to \mathbb{R}$ be a function satisfying the following assumptions:*

(F0) $F(|x_m|, t) \le F(|x_m|, |t|)\ \forall\ t \in \mathbb{R}$.

(F1) $F(|x_m|, .)$ *is continuous* $\forall\ m \in \mathbb{Z}$.

(F2) $\exists\ K > 0$ *and* $0 < \ell < 4s$ *such that* $\forall\ m \in \mathbb{Z}, t \ge 0,\ 0 \le F(|x_m|, t) \le K(t^2 + t^{\ell+2})$.

(F3) $\forall\ \varepsilon > 0, \exists\ m_0 \in \mathbb{Z}$ *and* $t_0 \in \mathbb{R}$ *such that*

$$F(|x_m|, t) \le \varepsilon t^2 \quad \forall\ m > m_0 \text{ and } |t| \le |t_0|.$$

(F4) $-F$ *is* $h\mathbb{Z}$ *supermodular.*

(F5) $F(|x_m|, \theta t) \ge \theta^2 F(|x_m|, t)\ \forall\ m \in \mathbb{Z}, \theta > 1, t \in \mathbb{R}$.

(F6) $\exists\ \delta > 0, t_1 > 0, m_1 \in \mathbb{Z}, \alpha > 0$ *such that* $F(|x_m|, t) > \delta t^\alpha$ *for any* $m > m_1$ *and* $|t| < |t_1|$, *where* $1 + 2s > \frac{\alpha}{2}$.

Then (4.1) admits a Schwarz symmetric minimizer $u^h_c = (u^h_c)^*$.

If in addition (F4) holds with a strict sign, then all minimizers of (4.1) are Schwarz symmetric.

Before proving this result, we need the following lemma:

Lemma 4.2 *Under (F6)* $I^h_c < 0\ \forall\ c \in \mathbb{R}$.

Proof Let $0 < p < 1$, $u_h \in S^h_c$, then $u^p_h(x_m) = p^{\frac{1}{2}} u_{ph}(x_m)$ is also in S^h_c.

$$E_h(u^p_h) = \sum_{k,\ell\in\mathbb{Z}} \frac{|u^p_h(x_k) - u^p_h(x_\ell)|^2}{h^{2s}|k-\ell|^{1+2s}} - \sum_{m\in\mathbb{Z}} F(|x_m|, u^p_h(x_m))$$

$$\sum_{k,\ell\in\mathbb{Z}} \frac{|p^{1/2} u p_h(x_k) - p^{1/2} u p_h(x_\ell)|^2}{|k-\ell|^{1+2s}} - \delta \sum_{|m|\ge|m_2|p^{\alpha/2}u^{p\alpha}_h} u^p_h(x_m)$$

$$\le \frac{p^{2s}}{h^{2s}} \sum_{k,\ell\in\mathbb{Z}} \frac{|u_h(x_k) - u_h(x_\ell)|^2}{|k-\ell|^{1+2s}} - \delta p^{-1} p^{\frac{\alpha}{2}} \sum_{|m|\ge|m_3|} u^\alpha_h(x_m)$$

$$\le \frac{p^{2s}}{h^{2s}} \sum_{k,\ell\in\mathbb{Z}} \frac{|u_h(x_k) - u_h(x_\ell)|^2}{|k-\ell|^{1+2s}} - p^{\frac{\alpha}{2}-1} \delta \sum_{|m|\ge|m_3|} u^\alpha_h(x_m)$$

the choice of p and α enables us to conclude.

Proof of Theorem 4.1

Step 1: (4.1) **is well posed** ($I_c^h > -\infty$ **and all minimizing sequences are bounded in** H_h^s).

By (F2), we can write:

$$\sum_{m\in\mathbb{Z}}(F(|x_m|, u_h(x_m))h \le K(\sum_{m\in\mathbb{Z}} u_h^2(x_m)h + \sum_{m\in\mathbb{Z}} u_h^{\ell+2}(x_m)h) \tag{4.2}$$

Now using the fractional discrete Gagliardo-Nirenberg inequality, (2.6), it follows:

$$\|u_h\|_{L_h^{\ell+2}} \le K'\|u_h\|_{L_h^2}^{1-\theta}\|\nabla_s u_h\|_{L_h^2}^{\theta} \quad \text{where } \theta = \frac{\ell}{2s(\ell+2)}$$

which implies that

$$\|u_h\|_{L_h^{\ell+2}}^{\ell+2} \le K''\{\|u_h\|_{L_h^2}^{(1-\theta)(\ell+2)}\|\nabla_s u_h\|_{L_h^2}^{\theta(\ell+2)}\}. \tag{4.3}$$

Now using Young inequality, we have:

$$\|u_h\|_{L_h^2}^{(1-\theta)(\ell+2)}\|\nabla_s u\|_{L_h^2}^{\theta(\ell+2)} \le \frac{1}{p}\varepsilon^p\|\nabla_s u_h\|_{L_h^2}^{p\theta(\ell+2)} + \frac{1}{q\varepsilon^q}\|u_h\|_{L_h^2}^{q(1-\theta)(\ell+2)} \tag{4.4}$$

for any $\varepsilon > 0, p > 1$ where $\frac{1}{p} + \frac{1}{q} = 1$, thus choosing $p = \frac{2}{\theta}(\ell+2) = \frac{4s}{\ell}$, we get

$$\begin{aligned}\|u_h\|_{L_h^{\ell+2}}^{\ell+2} &\le \tfrac{K''}{p}\varepsilon^p\|\nabla_s u_h\|_{L_h^2}^2 + \tfrac{K''}{q\varepsilon^q}\|u_h\|_{L_h^2}^{q(1-\theta)(\ell+2)} \\ &= \tfrac{K''}{p}\varepsilon^p\|\nabla_s u_h\|_{L_h^2}^2 + \tfrac{K''}{q\varepsilon^q}c^{q(1-\theta)(\ell+2)}\end{aligned} \tag{4.5}$$

for any $u_h^c \in S_c^h$.

Therefore

$$\begin{aligned}E_h(u_h) &\ge \frac{1}{2}\|\nabla_s u_h\|_{L_h^2}^2 - Kc^2 - K''K\varepsilon^p\|\nabla_s u_h\|_{L_h^2}^2 \\ &\quad - \frac{KK''}{q\varepsilon^q}c^{q(1-\theta)(\ell+2)} \\ &\ge (\frac{1}{2} - \frac{KK''}{p}\varepsilon^p)\|\nabla_s u_h\|_{L_h^2}^2 - Kc^2 - \frac{KK''}{q\varepsilon^q}c^{q(1-\theta)(\ell+2)}.\end{aligned}$$

Thus $I_c^h > -\infty$ and all minimizing sequences are bounded in H_h^s.

Step 2: **Existence of Schwarz symmetric minimizing sequence**.

By symmetrization inequalities proved in Sect. 3, we certainly have thanks to the assumption made on F, that

$$E_h(|u_h|) \le E(u_h)$$

so we can suppose without loss of generality that u_h is non-negative:

$$E_h(u_h^*) \leq E(u_h).$$

Step 3: Let $u_{h,n} = u_{h,n}^*$ be a Schwarz symmetric minimizing sequence of (4.1), then we can find $m_0 \in \mathbb{Z}$ such that

$$u_{n,h}^*(x_m) \leq \frac{c}{\sqrt{h}} \quad \forall\, n \in \mathbb{N}. \tag{4.6}$$

On the other hand by the weak lower semi-continuity of $\| \|_{L_h^2}$ we have that $\|\nabla_s u_h\|_{L_h^2} \leq \liminf \|\nabla_s u_h\|_{L_h^2}$.

Now fix $m_4 \in \mathbb{N}$, since $u_{h,n}$ converges weakly to u_h (up to a subsequence since it is bounded in H_h^s), it follows that it converges strongly to u_h in $L_h^{\ell+2}(|m| \leq m_4)$.

This implies that

$$\lim \sum_{|m|\leq m_4} F(|x_m|, u_{h,n}(x_m)) = \sum_{|k|\leq m_4} F(|x_m|, u_h(x_m))$$

(4.6) together with (F3) imply that

$$\sum_{|m|\geq p_0} F(|x_m|, u_{h,n}(x_m)) \text{ and } \sum_{|m|\geq p_0} F(|x_m|, u_h(x_m)) < \epsilon, \ \forall \epsilon > 0$$

for $p_0 \in \mathbb{N}$ big enough.

In conclusion

$$\lim_{n\to\infty} \sum_{m\in\mathbb{Z}} F(|x_m|, u_{h,n}(x_m)) = \sum_{m\in\mathbb{Z}} F(|x_m|, u_h(x_m)).$$

Step 4: I_c^h is achieved:

By the weak lower semi-continuity of the norm L_h^2, we know that

$$\sum_{m\in\mathbb{Z}} u_h^2(x_m)h \leq c^2.$$

Then observe that $u_h \neq 0$ since we know by Lemma 4.2 $I_c^h < 0$ and $F(., 0) = 0$ by (F2).

Now set $t^h = \frac{c^2}{\|u_h\|_{L_h^2}}$ then $t^h \geq 1$.

On the other hand

$$I_c^h \leq E_h(t^h u_h) \leq (t^h)^2 E(u_h) \leq (t^h)^2 I_c \Rightarrow t^h \leq 1$$

by the strict negativity of I_c^h.

If $-F$ is strictly $h\mathbb{Z}$ supermodular, then it follows from Theorem 3.2 that all minimizers are Schwarz symmetric.

Remark If $\ell = 4s$, it is easy to reproduce all the steps provided that c is small enough $(0 < c < (\frac{1}{2KK''})^4)$.

If $\lim_{t\to\infty} \dfrac{F(|x_m|, t)}{t^{\ell+2}} \geq A > 0$, then $I_c^h = -\infty$.

5 Some Applications

In the very special case $F(x, u) = \frac{1}{p+2}u^{p+2}$ and the fractional Laplacian is replaced by the classical one, the nonlinear Schrödinger equation (1.1) becomes:

$$i\partial_t\varphi = -\Delta\varphi - |\varphi|^p\varphi,$$

and its non-local versions arise in many domains, where one has to consider a lattice with a quantum particle sitting at each side, interacting with the others. Such lattice systems are used to understand electron transport in biopolymers like organic semiconductors, molecular crystals, and DNA.

As a first approximation, it is worth to consider a discrete model of quantum particles at lattice points with two kinds of interactions: nearest-neighbor interactions appearing as a discrete Laplacian term, representing interactions between base pairs in DNA; and self-interactions appearing as a p-nonlinear term, representing interactions within a base pair.

In a much better approximation, one has to take into account the long-range interactions (which need not be of fixed range, because DNA is constantly in flux). So we consider the same p self-interaction term as before, and inverse power-law long-range interactions for s parameters.

Our main result is another continuum limit: for certain values of s, solutions of the discrete model converge weakly to a solution of the NLS with fractional-order Laplacian:

$$i\partial_t\varphi = (-\Delta)^s\varphi - |\varphi|^p\varphi.$$

The dynamics are governed by the discrete NLS on the lattice of mesh size h, and we prove in the last section of this paper that taking the mesh size of the lattice to zero (the continuum limit), gives macroscopic behavior described by the focusing p-NLS:

It is important to have the same qualitative and quantititative properties of the discretized problems of the fractional NLS. The main difficulties are that there is no canonical discretization of the fractional derivative and that the most physical one doesn't obviously play well with the fractional derivative. We constructed a discrete fractional calculus and an interpolation of the discrete functions based on a special mollification (see Sect. 2). This framework is compatible with the fractional

derivative. We also have developed some ingenious harmonic analysis techniques to conduct the key steps of the proof in Fourier space.

For other nonlinearities covered by our integrand F, the lattice models Ferromagnets and spin glasses, Theorem 4.1 would clarify our understanding of glass and phenomena like neural networks and other models approximating Bose-Einstein condensation. A dynamic quantum icing model is the next step. Let us point out that Bose-Einstein condensates are unusual states of matter near absolute zero that can be used to slow and briefly stop light, as well as convert light to matter and back. There are excellent applications, such as quantum information processing and increased accuracy in measurements by inter-ferrometry with atom lasers instead of traditional photon lasers. But the Bose Einstein Condensate is fragile and difficult to work with, so it is vital to work out the theory.

More precisely, the particles are so super-cooled (to a few billionths of a degree Kelvin) that they all fall into the **ground state** and exhibit quantum mechanical behavior macroscopically, in effect they condense into a quantum super-particle. Only two years ago, the Bose Einstein condensates macroscopic behavior was explained mathematically: the microscopic repulsive interactions between quantum particles give rise, in the scaling limit, to quantum macro-behavior governed by the p- nonlinear Schrödinger equation (p-NLS):

$$i\partial_t \varphi = -\Delta\varphi + b|\varphi|^{p+1}\varphi.$$

Again usually the latter equation does not take into account natural anomalous diffusion phenomena, and in the most realistic model, one has to replace the classical Laplacian by the fractional Laplacian. As stated above the ground state solutions play a key role here. This special non-linearity is also covered by our main result (see Theorem 4.1).

References

1. Cho, Y., Hajaiej, H., Hwang, G., Ozawa, T.: On the Cauchy problem of fractional Schrödinger equation with Hartree type nonlinearity. Funkcialaj Ekvacioj **56**(2), 193–224 (2013)
2. Choi, Y.S., McKenna, P.J.: A mountain pass theorem for the numerical solution of semilinear elliptic problems. Nonlinear Anal. **20**, 417–437 (1993)
3. Gagliola, S., Kaypatrick, K., Lenzemann, E.: On the continuum limit of discrete NLS with long-range lattice interaction, to appear. Commun. Math. Phys
4. Hadj Selem, F.: Personal communications
5. Hajaiej, H.: Explicit constructive approximation to symmetrization via iterated polarizations. J. Convex Anal. **17**(2), 405–411 (2010)
6. Hajaiej, H.: Existence of minimizers of functional involving the fractional gradient in the absence of compactness, symmetry and monotonicity. J. Math. Anal. Appl. **399**(1), 17–26 (2013)
7. Hajaiej, H.: On the optimality of the conditions used to prove the symmetry of the minimizers of some fractional constrained variational problems. Ann. Henri. Poincare. **14**(5), 1425–1433 (2013)

8. Hajaiej, H.: Characterization of the orbit of standing waves of Hartree-type equations with external Coulomb potential. Asymptotic. Anal. **87**, 57–64 (2014)
9. Hajaiej, H.: Symmetry of minimizers of some fractional problems. Appl. Anal. **94**(4), 694–700 (2015)
10. Hajaiej, H., Molinet, L., Ozawa, T., Wang, B.: Sufficient and necessary conditions for the fractional Gagliardo-Nirenberg inequality and applications to Navier-Stokes and Boson equations. RIMS Kokyuroku Bessatsu, B **26**, 159–199 (2011)
11. Pruss, A.R.: Discrete convolution rearrangement inequalities for the Faber Kahn inequality in a regular tree. Duke Math, J. **91**(3), 463–514 (1998)
12. Tarasov, V.E.: Continuous limit of discrete systems with long-range interaction. J. Phys. A. **39**, 14895–14910 (2006)

A Scientific Tour on Orthogonal Arrays

A.S. Hedayat

Abstract This paper gives a brief introduction to orthogonal arrays, including the definitions, basic questions, important theorems and applications. It establishes the connection between coding theory and orthogonal arrays. Based on coding theory, many construction methods of orthogonal arrays and linear programming bound, which is an improvement on Rao's bound, are studied. Difference schemes and Hadamard matrices are also discussed in the paper, which contribute to the constructions of orthogonal arrays. Moreover, the paper brings in basic definitions and properties of mixed orthogonal arrays and focuses on problems and methods related to their constructions. As the main statistical application of orthogonal arrays, factorial experiments are then introduced, and ways orthogonal arrays can be used in this field is discussed. Further, the applications of orthogonal arrays in computer experiments and related structures are shown, including orthogonal Latin hypercube designs, nested orthogonal arrays, sliced orthogonal arrays, Latin squares and compound orthogonal arrays.

Keywords Orthogonal arrays · Coding theory · Factorial experiments · Computer experiments

Mathematics Subject Classification (2010): Primary 42A75 · Secondary 94A20

1 Introduction

The general theory and application of orthogonal arrays were first introduced by Rao [18–20]. Ever since their introduction, many outstanding researchers with various backgrounds have been inspired by the subject and have made great contributions to the field. Orthogonal arrays have played and continue to play a prominent role in

A.S. Hedayat (✉)
Department of Mathematics, Statistics, and Computer Science,
University of Illinois at Chicago, 851 South Morgan Street, Chicago, IL 60607-7045, USA
e-mail: hedayat@uic.edu

T. Abualrub et al. (eds.), *Mathematics Across Contemporary Sciences*,
Springer Proceedings in Mathematics & Statistics 190,
DOI 10.1007/978-3-319-46310-0_7

the design of experiments. Besides the main applications, orthogonal arrays are also closely related to finite geometry, combinatorics, error-correcting codes, Hadamard matrices and Latin squares. Orthogonal arrays and related structures have been widely used in factorial experiments, medicine, clinical trials and computer experiments.

The paper is organized as follows. In this section we will introduce how orthogonal arrays are used in factorial designs and provide basic definitions along with their properties. Then we shall introduce Rao's bounds and some famous construction methods. In Sect. 2 we will discuss the relationships between orthogonal arrays and coding theory. With the help of coding theory, we are able to find some construction methods of orthogonal arrays and improve Rao's bound using linear programming bound. Hadamard matrices and difference schemes will be presented in Sect. 3 along with their association with orthogonal arrays, which also inspires ideas in the construction methods. In Sect. 4, definitions of mixed orthogonal arrays and their parameters will be studied. Moreover, some methods of construction of mixed orthogonal arrays will be discussed. Application of orthogonal arrays in factorial experiments, which is their most important statistical application, is discussed in Sect. 5, in which we will see examples of special cases. Application of orthogonal arrays for computer experiments will be the focus of Sect. 6, along with structures derived from orthogonal arrays. Finally, in Sect. 7 we will see other topics such as Latin squares, F-squares and compound orthogonal arrays.

1.1 Definitions and Properties

Let $S = \{0, 1, ..., s-1\}$ be a set of s symbols or levels. In the design of experiments, the symbols typically indicate the levels of the factors whose effects on a response of interest are to be studied.

Definition 1.1 An $N \times k$ array A with entries from S is said to be an *orthogonal array with s levels, strength t* $(0 \leq t \leq k)$ *and index* λ if every $N \times t$ subarray of A contains each of the s^t possible $t-$tuples equally often (say λ times) as a row.

Such an array is denoted by $OA(N, k, s, t)$ or $OA(N, s^k, t)$. The integers N, k, s, t and λ are referred to as the parameters of the orthogonal array. The number of rows N is the number of runs, the number of columns k is the number of factors, s is the number of levels, t is the strength and λ is the index. In the important case when $\lambda = 1$, we usually say that the orthogonal array has index unity.

Example 1.1 The array below is an orthogonal array with eight runs, four factors each of two levels, strength three and of index unity. It's an $OA(8, 4, 2, 3)$.

$$
\begin{array}{c}
0\,0\,0\,0 \\
0\,0\,1\,1 \\
0\,1\,0\,1 \\
0\,1\,1\,0 \\
1\,0\,0\,1 \\
1\,0\,1\,0 \\
1\,1\,0\,0 \\
1\,1\,1\,1
\end{array}
$$

Some general properties can be obtained directly from the definition:

1. The parameters of an orthogonal array satisfy the equality

$$\lambda = N/s^t. \tag{1}$$

2. An orthogonal array is invariant to permutations of rows, columns, and symbols within a column. This property brings about the definitions of isomorphic and statistical equivalent arrays. Two orthogonal arrays are said to be *isomorphic* if one can be obtained from the other by a sequence of permutations of the columns, the rows, and the levels of each factor. Two orthogonal arrays are said to be *statistically equivalent* if one can be obtained from the other by a sequence permutation of the runs.
3. If A_i, $i = 1, ..., r$ is an $OA(N_i, k, s, t_i)$, then the array A obtained from the juxtaposition of these r arrays is an $OA(N, k, s, t)$ where $N = N_1 + N_2 + \ldots + N_r$ and the strength is t for some $t \geq \min\{t_1, ..., t_r\}$. Further, when $r = s$ and each A_i is an $OA(N, k, s, t)$, after appending a 0 to each row of A_1, a 1 to each row of A_2 and so on, we obtain an $OA(sN, k + 1, s, t)$.
4. Any orthogonal array of strength t is also an orthogonal array of strength t', $0 \leq t' < t$. The index of the array when considered as an array of strength t' is $\lambda s^{t-t'}$, where λ denotes the index of the array when considered to have strength t.
5. Any $N \times k'$ subarray of an $OA(N, k, s, t)$ is an $OA(N, k', s, t')$, where $t' = \min\{k', t\}$.
6. Existence of an $OA(N, k, s, t)$ implies existence of an $OA(N/s, k - 1, s, t - 1)$. This process can be achieved by permutating the runs in $OA(N,\ k,\ s,\ t)$ so that the first N/s runs all begin with 0, the second N/s runs begin with 1, ..., the last N/s runs begin with $s - 1$. Omitting the first columns yields an $OA(N/s,\ k - 1,\ s,\ t - 1)$.
7. Let C be the set of all possible runs that could have occurred in a particular orthogonal array A. For $c \in C$, let f_c be the frequency of c in A and $f = \max_{c \in C}\{f_c\}$. Then the array which contains run c with frequency $f - f_c$ for all $c \in C$ is said to be the complement of A. The complement of an $OA(N, k, s, t)$ is an $OA(fs^k - N, k, s, t)$.
8. Suppose $A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}$ is an $OA(N, k, s, t)$, where A_1 itself is an $OA(N_1, k, s, t_1)$. Then A_2 is an $OA(N - N_1, k, s, t_2)$ with $t_2 \geq \min\{t, t_1\}$.

1.2 Rao's Bounds

An important and basic problem in the study of orthogonal arrays is the existence of $OA(N, k, s, t)$ for given values of $s \geq 2, t \geq 2, k \geq t, N \equiv 0 (mod\, s^t)$. It can be treated as the problem of determining the minimal number of runs N, denoted by $F(k, s, t)$, in any $OA(N, k, s, t)$ for given values of k, s and t. The problem can also be changed to determining the maximal number of factors k, denoted by $f(N, s, t)$, in any $OA(N, k, s, t)$ for given values of N, s and t. $F(k, s, t)$ and $f(N, s, t)$ are related in the following ways:

$$\begin{aligned} F(k, s, t) &= \min\{N : f(N, s, t) \geq k\}, \\ f(N, s, t) &\leq \max\{k : F(k, s, t) \leq N\}. \end{aligned} \tag{2}$$

The way we establish values for $f(N, s, t)$ and $F(k, s, t)$ is typically through a combination of obtaining a bound and constructing an orthogonal array that attains that bound. One of the first upper bounds on the maximal number of factors in an orthogonal arrays was obtained by Rao [19], which provides an explicit lower bound for $F(k, s, t)$ and an implicit upper bound for $f(N, s, t)$. Details are shown in the theorem below.

Theorem 1.1 (Rao's Inequalities) *The parameters of an $OA(N, k, s, t)$ satisfy the following inequalities:*

$$\begin{aligned} N &\geq \sum_{i=0}^{u} \binom{k}{i} (s-1)^i, \quad \text{if} \quad t = 2u, \\ N &\geq \sum_{i=0}^{u} \binom{k}{i} (s-1)^i + \binom{k-1}{u} (s-1)^{u+1}, \quad \text{if} \quad t = 2u+1, \end{aligned} \tag{3}$$

for $u \geq 0$.

Rao's bounds apply to any $OA(N, k, s, t)$. But when considering specific values for one or more of the parameters, the bounds can be attained for some cases and sharpened for others. Bose and Bush [1] sharpened Rao's bounds for when $t = 2$ or 3. The discussions of various results of this type have been presented comprehensively in Hedayat, Sloane and Stufken's book *Orthogonal Arrays: Theory and Application*.

1.3 Constructions

A large number of techniques are known for constructing orthogonal arrays. An example is given below.

Example 1.2 (Zero-sum array) An $OA(N, t+1, s, t), t \geq 2$ can be constructed in the following way: Start with an $N \times t$ array based on 0, ..., $s-1$ which has each of the s^t possible $t-$tuples N/s^t times as rows. Then add one more column to the original array so that the entries in each row add up to 0(mod s) in the new arrays. One can verify that the resulting $N \times (t+1)$ array is an $OA(N, t+1, s, t)$. Since this array has the property that the levels in each run add up to zero, it's called a zero-sum array.

Bush [3] studied the construction of orthogonal arrays of index unity, which have the smallest number of runs for a given number of levels and strength, and are thus mathematically interesting and highly useful in statistical experiments. Main results of Bush's construction are shown in Theorems 1.2 and 1.3.

Theorem 1.2 *If $s = p^n, n \geq 1$, p is a prime power and $s \geq t$, then an $OA(s^t, s+1, s, t)$ of index unity exists.*

Theorem 1.2 gives a lower bound on the number of factors in such arrays when s is a prime power. In some cases, the result can be improved. For example, the result of Theorem 1.3 can be obtained for $p = 2, t = 3$.

Theorem 1.3 *If $s = 2^m, m \geq 1$, and $t = 3$, then there exists an $OA(s^3, s+2, s, 3)$.*

Construction: When $s = p^n, n \geq 1$ where p is a prime power, we first construct an $s^t \times s$ array whose columns are labeled by the elements of Galois field $GF(s)$ and whose rows are labeled by the s^t polynomials over $GF(s)$ of degree at most $t-1$. Denote those polynomials by ϕ_1, ..., ϕ_{s^t}:

$$\phi_j(x) = a_{j,t-1}x^{t-1} + \cdots + a_{j,1}x + a_{j,0}, \text{ where } a_{j,t-1}, \ldots, a_{j,1}, a_{j,0} \in GF(s).$$

Define the entry in the ith column and the jth row to be $\phi_j(\alpha_i)$, which is the value of the polynomial ϕ_j at α_i. Now add one additional factor to the array and take the level of this factor in the jth row to be the coefficient of x^{t-1} in ϕ_j. We can verify that the resulting $s^t \times (s+1)$ array is an $OA(s^t, s+1, s, t)$.

Specifically when $s = 2^m, m \geq 1$, we first use the previous construction method to obtain an $OA(s^3, s+1, s, 3)$, then adjoin another factor and take the level of this factor in jth row to be the coefficient of x in ϕ_j. This $s^3 \times (s+2)$ array can be verified to be an $OA(s^3, s+2, s, 3)$.

From the construction above, we can find that there are two special properties of the orthogonal arrays we constructed, one is simple and the other is linear. Definitions are listed below.

Definition 1.2 An orthogonal array is *simple* if its runs are distinct.

Definition 1.3 Let s be a prime power. If an orthogonal array $OA(N, k, s, t)$ with levels from $GF(s)$ is simple and its N runs form a vector space over $GF(s)$ when considered as $k-$tuples from $GF(s)$, then we say the orthogonal array is *linear*.

If an orthogonal array is linear, then its runs can be regarded as the codewords from a linear error-correcting codes. Relationship between codes and orthogonal arrays are discussed in Sect. 2.

Rao [18, 19] gave a construction method for $OA(s^n, (s^n - 1)/(s - 1), s, 2)$, which results in linear arrays. The same construction is also used in the Hamming [7] codes, which is one of the best-known families of error-correcting codes. This construction is thus called Rao-Hamming construction and the following results provide two versions of it.

Theorem 1.4 *If s is a prime power, then an $OA(s^n, (s^n - 1)/(s - 1), s, 2)$ exists whenever $n \geq 2$.*

Rao's Construction: Form an $s^n \times n$ array whose rows are all possible $n-$tuples from $GF(s)$. Let $C_1, ..., C_n$ denote the columns of this array. Then the columns of the full orthogonal array consist of all columns of the form

$$z_1C_1 + \cdots + z_nC_n = [C_1, \ldots, C_n]z,$$

where $z = (z_1, \ldots, z_n)^\top$ is a nonzero $n-$tuple from $GF(s)$ in which the first nonzero z_i is 1. Notice that there are $(s^n - 1)/(s - 1)$ such columns, thus forming an $s^n \times (s^n - 1)/(s - 1)$ array. One can verify that the resulting array is an $OA(s^n, (s^n - 1)/(s - 1), s, 2)$.

Hamming's Construction: Alternatively, we can start with an $n \times (s^n - 1)/(s - 1)$ matrix whose columns are all nonzero $n-$tuples $(z_1, ..., z_n)^\top$ from $GF(s)$ in which the first nonzero z_i is 1. By taking all linear combinations of the rows of this generator matrix, we can obtain an orthogonal array whose runs are comprised of all linear combinations

$$\alpha_1R_1 + \cdots + \alpha_nR_n,$$

where $R_1, ..., R_n$ are the rows of the matrix and $\alpha_1, ..., \alpha_n \in GF(s)$, thus an $OA(s^n, (s^n - 1)/(s - 1), s, 2)$ is constructed.

2 Orthogonal Arrays and Coding Theory

Error-correcting codes and orthogonal arrays are closely related. Both subjects study analogous problems and have a number of parallel constructions and theorems. Some of their basic parameters can also correspond to each other in parallel results. For example, alphabet size, length of code, number of codewords and minimal distance of dual in an error-correcting code can be treated as the number of levels, number of factors, number of runs and strength of an orthogonal array respectively, and vice verse. Basic definitions are given below.

Let S be a set of symbols of size s and S^k be the set of all s^k vectors of length k. An *error-correcting code* is any collection C of vectors in S^k (repetition allowed). S is called the *alphabet*. The vectors in C are called *codewords*.

The *Hamming distance* dist(u, v) between two vectors $u, v \in S^k$ is defined to be the number of positions where they differ. The *minimal distance* d of a code C is defined to be the minimal distance between two distinct codewords:

$$d = \min_{u,v \in C, u \neq v} \text{dist}(u, v).$$

d is undefined if the code is empty. If there is only one distinct codeword, then d is defined to be $k + 1$.

If C is a code of length k, codewords size N and minimal distance d over an alphabet of size s, then we denote C by $(k, N, d)_s$ code.

Example 2.1 A code of length $k = 4$ with an alphabet of size $s = 2$ (binary) with $N = 8$ codewords size and minimal distance $d = 2$ is

$$\begin{bmatrix} 0\,0\,0\,0 \\ 0\,0\,1\,1 \\ 0\,1\,0\,1 \\ 0\,1\,1\,0 \\ 1\,0\,0\,1 \\ 1\,0\,1\,0 \\ 1\,1\,0\,0 \\ 1\,1\,1\,1 \end{bmatrix}$$

If a code doesn't contain repeated codewords, we call it a *simple* code. A simple code with minimal distance d can correct up to $\lfloor (d-1)/2 \rfloor$ errors. One of the essential problems in coding theory is to find the maximal value of N for a simple code given s, k and d. In practice one also wants to make encoding and decoding feasible, thus codes with a rich mathematical structures are preferable.

A code C is called *linear* if its codewords are distinct and form a vector subspace of S^k. The *dimension* of the code C is n if C has $N = s^n$ for some nonnegative integer $n, 0 \leq n \leq k$. For a linear code, we can define its *dual* $C^\perp$ through the vectors in the null space of C: $C^\perp$ consists of all vectors in $v \in S^k$ such that

$$uv^\top = 0 \quad \text{for all} \quad u \in C.$$

Note that $C^\perp$ is also a linear code and can be verified to be a $(k, N', d^\perp)_s$ code where $N' = s^{k-n}$. $d^\perp$ is called the *dual distance* of C. From the definition of linearity in orthogonal arrays and codes, we can conclude that a orthogonal array is linear if and only if its associated code is also linear. Bose [2] specified the relationship between the strength of a linear orthogonal array and the associated linear code.

Theorem 2.1 *If C is a linear code $(k,\ N,\ d)_s$ over $GF(s)$ with dual distance $d^\perp$, then its codewords form the rows of an OA$(N,\ k, s,\ d^\perp - 1)$ with entries from $GF(s)$. Conversely, the runs of a linear OA(N, k, s, t) over $GF(s)$ form a $(k, N, d)_s$ linear code over $GF(s)$ with dual distance $d^\perp \geq t + 1$ (equality holds if the orthogonal array has strength t but not $t + 1$).*

The example below illustrates Theorem 2.1.

Example 2.2 The Rao-Hamming orthogonal array $OA(8, 7, 2, 2)$ is given below.

$$\begin{bmatrix} 0&0&0&0&0&0&0 \\ 1&1&1&0&1&0&0 \\ 0&1&1&1&0&1&0 \\ 0&0&1&1&1&0&1 \\ 1&0&0&1&1&1&0 \\ 0&1&0&0&1&1&1 \\ 1&0&1&0&0&1&1 \\ 1&1&0&1&0&0&1 \end{bmatrix}$$

The runs of the array form a $(7, 8, 4)_2$ code, which is a single-error correcting code. Its dual is a $(7, 16, 3)_2$ code, i.e. the dual distance of the $(7, 8, 4)_2$ code is 3. It belongs to the class of Hamming codes.

From Theorem 2.1, there is no direct relationship between d and strength t. Only the dual distance $d^{\perp}$ provides the performance of C as an orthogonal array. As a result, we need to define the dual distance of a nonlinear code in order to establish the relationship between nonlinear codes with orthogonal arrays. The introduction of the weight distribution and weight enumerator of a code can help us to accomplish the goal.

For a $(k,\ N,\ d)_s$ code C and a codeword $u \in C$, the *weight distribution with respect to u* can be defined as $(A_0(u), A_1(u), \ldots, A_k(u))$, which is a $(k+1)-$tuple of nonnegative integers. $A_i(u)$ is the number of codewords $v \in C$ such that $\text{dist}(u, v) = i$.

Definition 2.1 The *weight distribution* of a code C can be defined as the $(k+1)-$tuple $(A_0, A_1, \ldots, A_k)$, where

$$A_i = \frac{1}{N} \sum_{u \in C} A_i(u), 0 \le i \le k.$$

The minimal distance of the code is the largest positive integer d such that

$$A_1 = A_2 = \ldots = A_{d-1} = 0.$$

The *weight enumerator* of C is

$$W_C(x, y) = \sum_{i=0}^{k} A_i x^{k-i} y^i$$

which is a homogenous polynomial whose degree is equal to the length of the code.

If C is a linear code, $A_i(u)$ is independent of u and $A_i = A_i(u)$ for all $u \in C$. In any case, the weight distribution of C satisfies

$$\begin{aligned} &\sum_{i=0}^{k} A_i = N, \\ &A_0 \geq 1, \quad A_1 = A_2 = \ldots = A_{d-1} = 0, \quad A_i \geq 0, \text{ for } d \leq i \leq k. \end{aligned} \tag{4}$$

For linear codes, we can obtain the formula for the weight enumerator of the dual code, called the MacWilliams identity for linear codes. The proof can be found in MacWilliams and Sloane [14].

Theorem 2.2 *For a $(k, s^n, d)_s$ linear code C,*

$$W_{C^\perp}(x, y) = \frac{1}{N} W_C(x + (s-1)y, x - y). \tag{5}$$

Using the identity, the weight distribution $(A_0^\perp, A_1^\perp, \ldots, A_k^\perp)$ of $C^\perp$ can be expressed in terms of the weight distribution of C:

$$A_i^\perp = \frac{1}{N} \sum_{j=0}^{k} A_j P_i(j), \quad 0 \leq i \leq k, \tag{6}$$

where the $P_i(j)$ is the Krawtchouk polynomial, i.e.

$$P_i(j) = \sum_{r=0}^{i} (-1)^r (s-1)^{i-r} \binom{j}{r} \binom{k-j}{i-r}, 0 \leq i \leq k.$$

The weight distribution $(A_0^\perp, A_1^\perp, \ldots, A_k^\perp)$ of $C^\perp$ satisfies

$$\begin{aligned} &\sum_{i=0}^{k} A_i^\perp = s^k / N, \\ &A_0^\perp = 1, \quad A_1^\perp = \ldots = A_{d^\perp - 1}^\perp = 0, \quad A_i^\perp \geq 0, \text{ for } d^\perp \leq i \leq k, \end{aligned} \tag{7}$$

For a *nonlinear code*, we can still define the dual weight distribution using MacWilliam identity based on (5) and (6). Then $A_0^\perp = 1$ and $A_i^\perp \geq 0$ for $0 \leq i \leq k$ still hold. Further, we define the *dual distance* $d^\perp$ to be the largest positive integer such that

$$A_1^\perp = \ldots = A_{d^\perp - 1}^\perp = 0.$$

Now the results in Theorem 2.1 can be extended to any codes. Theorem 2.3 is due to Delsarte [4].

Theorem 2.3 *If C is a $(k, N, d)_s$ code with dual distance $d^\perp$, then the corresponding orthogonal array is an OA($N, k, s, d^\perp - 1$). Conversely, the code corresponding to*

an OA(N, k, s, t) is a $(k, N, d)_s$ code with dual distance $d^\perp \geq t+1$ (equality holds if the orthogonal array has strength t but not $t+1$).

From (4), we can find that the total number of runs or codewords in an orthogonal array or code can be expressed in the form of the weight distribution:

$$N = A_0 + A_1 + \ldots + A_k.$$

So we can obtain a lower bound on N by using linear programming to find the smallest value of $\sum_{i=0}^{k} A_i$ under the constraints of the conditions (4) and (7). Delsarte (1937) established the linear programming bound for N as indicated in Theorem 2.4.

Theorem 2.4 (The LP Bound for orthogonal arrays [4]) *Given k, s and t, let $N_{LP}(k; d^\perp)$ be the solution to the following linear programming problem:*
choose real numbers $A_0, \ldots, A_k$ so as to

$$\text{minimize} \sum_{i=0}^{k} A_i$$

subject to the constraints

$$\begin{aligned} &A_0 \geq 1; \quad A_i \geq 0, \ 1 \leq i \leq k, \\ &A_0^\perp = 1; \quad A_i^\perp \geq 0, 1 \leq i \leq k, \\ &A_1^\perp = \cdots = A_t^\perp = 0, \quad t = d^\perp - 1, \\ &\text{where } A_i^\perp \text{ are defined as in (6) for } 0 \leq i \leq k. \end{aligned} \tag{8}$$

Then the size of any orthogonal array OA(N, k, s, t) satisfies

$$N \geq N_{LP}(k; d^\perp).$$

Linear programming bound provides a lower bound on N. This bound may be very weak, thus can be improved in many cases. Using the notation in Sect. 1, we know that $F(k, s, t) \geq N_{LP}(k, d^\perp)$. Note that N_{LP} and A_i are usually non-integral, and even if they are, there is always possibility that the corresponding orthogonal array may not exist.

The linear programming bound is always at least as good as Rao's bound since Delsarte [4] showed that the general Rao's bound can be implied by the linear programming bound. Table 1 shows an example of the comparison between these two bounds for some values of k.

In Table 1, let N_{Rao} and N_{LP} denote the value given by Rao's bound and the linear programming bound, respectively. The numerical values of Rao's bound and the next multiple of 16 larger than them are given in the parentheses and the numbers before the parentheses, respectively, in the second line. Similarly, the linear programming bound and the the next multiple of 16 are presented in the third line. The last line shows the smallest N that is known under certain value of k. If it is the smallest possible N, an asterisk, $*$, is added after the number.

Table 1 Comparison of Rao and linear programming bounds for orthogonal arrays with k factors when $s = 2, t = 4$

k	6	7	8	9	10	11	12	13
N_{Rao}	32(22)	32(29)	48(37)	48(46)	64(56)	80(67)	80(79)	96(92)
N_{LP}	32(26.7)	48(42.7)	64	96(85.3)	96(85.3)	96(85.3)	112(102)	128
Known	32*	64*	64*	128	128	128	128	128*

3 Difference Schemes and Hadamard Matrices

A difference scheme, a simple but powerful tool for the construction of orthogonal arrays of strength two, was first defined by Bose and Bush [1]. Hadamard matrices are the most important examples of two-level difference schemes. Their definitions are listed below.

Let $(\mathscr{A}, +)$, or $\mathscr{A}$ denote a finite abelian group with a binary operation $+$. Denote the cardinality of $\mathscr{A}$ by s, the identity element by 0 and the inverse of an element σ by $-\sigma$. In most examples $(\mathscr{A}, +)$ will be taken to be the additive group associated with the Galois field $GF(s)$.

Definition 3.1 An $r \times c$ array D with entries from $\mathscr{A}$ is called a *difference scheme based on* $(\mathscr{A}, +)$ if it has the property that for all i and j with $1 \leq i, j \leq c, i \neq j$, the vector difference between the ith and jth columns contains every element of $\mathscr{A}$ equally often.

We denote such an array by $D(r, c, s)$, and refer to it as a difference scheme with s levels and index λ, where λ is the number of times each element of $\mathscr{A}$ occurs in the difference of two columns. Clearly, $r = \lambda s$.

Example 3.1 The 6×6 array below gives an example of $D(6, 6, 3)$, which is a difference scheme with 3 levels and index 2.

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 2 & 1 & 2 & 0 \\ 0 & 2 & 1 & 1 & 0 & 2 \\ 0 & 2 & 2 & 0 & 1 & 1 \\ 0 & 0 & 1 & 2 & 2 & 1 \\ 0 & 1 & 0 & 2 & 1 & 2 \end{bmatrix}$$

Jungnickel [13] established Theorem 3.1.

Theorem 3.1 *If a difference scheme $D(r, c, s)$ exists, then $c \leq r$.*

When $c = r, s = 2$, a difference scheme $D(r, r, 2)$ is of great importance, which is the ordinary Hadamard matrix of order r.

A *Hadamard matrix of order n* is an $n \times n$ matrix H_n with entries $+1$ and -1, whose rows are orthogonal to each other. A Hadamard matrix satisfies

$$H_n H_n^\top = nI_n \tag{9}$$

Hadamard matrices were first introduced in Hadamard [6]. In the paper Hadamard showed that if $A = (a_{ij})$ is any $n \times n$ matrix with $|a_{ij}| \leq 1$, then

$$|\det A| \leq n^{n/2},$$

with equality if and only if A is what is now called a Hadamard matrix.

Suppose H_n is a Hadamard matrix of order n. Then by definition $H_n^{-1} = n^{-1} H_n^\top$,

$$H_n^\top H_n = nI_n \tag{10}$$

which implies that the columns of H_n are also orthogonal.

Due to the orthogonality relations (9) and (10), we can find that all matrices obtained by permutations of rows or columns of H_n and negating any of rows or columns of H_n will still satisfy (9) and (10), and they are said to be isomorphic to H_n. Through this kind of transformation, we can arrange the first row and column of H_n to consist of $+1$. Such a Hadamard matrix is said to be normalized. From the orthogonality relations, we can obtain Lemma 3.1.

Lemma 3.1 *Let H_n be a normalized Hadamard matrix of order n, $n > 2$. Let $u = (u_1, \ldots, u_n)$ and $v = (v_1, \ldots, v_n)$ be any two distinct rows of H_n apart from the first one. Then:*

(a) There are $n/2$ coordinates with $u_i = +1$ and $n/2$ with $u_i = -1$.

(b) There are $n/4$ coordinates with $u_i = v_i = +1$, $n/4$ with $u_i = v_i = -1$, $n/4$ with $u_i = -1$, $v_i = +1$ and $n/4$ with $u_i = +1$, $v_i = -1$.

(c) Similar results hold for the columns of H_n.

Following Lemma 3.1, we can conclude a necessary condition for the existence of Hadamard matrices:

Corollary 3.1 *If a Hadamard matrix H_n exists, then $n = 1, 2$ or $n = 4u$ for some integer $u \geq 1$*

The converse of Corollary 3.1 leads to the Hadamard conjecture, which is a widely accepted assertion. It provides an answer to the problem concerning the existence condition of Hadamard matrices, but this problem remains unsolved.

Note 1 (*The Hadamard Conjecture*) A Hadamard matrix of order n exists if n is 1, 2 or n is a multiple of 4.

A difference scheme can be converted into an orthogonal array by following procedure: Let D be a difference scheme based on $(\mathscr{A}, +)$, where $\mathscr{A} = \{\sigma_0, \ldots, \sigma_{s-1}\}$. We use D_i to denote the array obtained from D by adding σ_i to each of its entries, clearly, D_i is a difference scheme with the same parameter as D. Then by juxtaposition of the D_i's, we obtain an orthogonal array of strength 2. Theorem 3.2 can be concluded from this process:

Theorem 3.2 *If D is a difference scheme $D(r, c, s)$, then*

$$A = \begin{bmatrix} D_0 \\ \vdots \\ D_(s-1) \end{bmatrix} = \begin{bmatrix} D+0 \\ \vdots \\ D+(s-1) \end{bmatrix} \tag{11}$$

is an $OA(rs, c, s, 2)$.

We may want to know whether an orthogonal array of strength 2 constructed through Theorem 3.2 has achieved the maximal number of factors. The answer is no. Actually, it can be shown that at least an additional factor can be added, i.e.

Corollary 3.2 *A difference scheme $D(r, c, s)$ can lead to an orthogonal array $OA(rs, c+1, s, 2)$.*

Thus the maximal number of factors in this case $f(rs, s, 2) \geq c + 1$.

As we have already discussed before, Hadamard matrices are special cases of the difference schemes.

Theorem 3.3 *A Hadamard matrix H_n exists if and only if a difference scheme $D(n, n, 2)$ exists.*

As a result, we can also find the connection between Hadamard matrices and orthogonal arrays.

Theorem 3.4 *Orthogonal arrays $OA(4\lambda, 4\lambda-1, 2, 2)$ and $OA(8\lambda, 4\lambda, 2, 3)$ exist if and only if there exists a Hadamard matrix of order 4λ.*

Construction: Suppose $H_{4\lambda}$ is a normalized Hadamard matrix. By Lemma 3.1, we can obtain an $OA(4\lambda, 4\lambda-1, 2, 2)$ by omitting the first columns of $H_{4\lambda}$:

$$H_{4\lambda} = \left[1 \; OA(4\lambda, 4\lambda-1, 2, 2) \right].$$

Similarly, an $OA(8\lambda, 4\lambda, 2, 3)$ can be obtained by juxtaposing $H_{4\lambda}$ and $-H_{4\lambda}$:

$$OA(8\lambda, 4\lambda, 2, 3) = \begin{bmatrix} H_{4\lambda} \\ -H_{4\lambda} \end{bmatrix}.$$

From Theorem 3.4, we can conclude that the study of two-level orthogonal arrays of strength 2 and 3 is essentially equivalent to the study of Hadamard matrices.

Note that we can only obtain orthogonal arrays of strength 2 or 3 from the results before. Now we bring in the definition of difference schemes of strength t, which contribute to the construction of orthogonal arrays of strength t.

Let $\mathscr{A}$ denote an abelian group of order s. When $t \geq 1$, let $\mathscr{A}^t$ denote the abelian group of order s^t consisting of all possible $t-$tuples of elements from $\mathscr{A}$, with the binary operation being the vector addition. Treat each kind of $t-$tuples as a subgroup of $\mathscr{A}^t$, i.e. start from $\mathscr{A}_0^t$, which is a subgroup of order s:

$$\mathscr{A}_0^t = \{(x_1, \ldots, x_t) : x_1 = \cdots = x_t \in \mathscr{A}\}.$$

Denote its cosets by $\mathscr{A}_i^t$, $i = 1, \ldots, s^{t-1} - 1$, then $\mathscr{A}^t = \bigcup_{i=0}^{s^{t-1}-1} \mathscr{A}_i^t$.

Definition 3.2 An $r \times c$ array D with entries from $\mathscr{A}$ is called a *difference scheme of strength t* if in every $r \times t$ subarray, members of each $\mathscr{A}_i^t$ $(i = 0, \ldots, s^{t-1} - 1)$ occur equally often when the rows of the subarray are treated as elements of $\mathscr{A}^t$. Such an array is denoted by $D_t(r, c, s)$.

When $t = 2$, the definition is equivalent to Definition 3.1. Following result illustrates the relationship between difference schemes of strength t and orthogonal arrays of strength t.

Theorem 3.5 *A $D_t(r, c, s)$ of strength t can be used to construct an OA(rs, c, s, t). If $D_t(r, c, s)$ itself is already an orthogonal arrays of strength $t - 1$, or it can be written as the juxtaposition of s difference schemes $D_{t-1}(r/s, c, s)$, then an additional factor can be added, resulting in an OA$(rs, c + 1, s, t)$.*

Example 3.2 The array below shows a difference scheme $D_3(8, 7, 2)$ of strength 3 over $(GF(2), +)$. Following Definition 3.2, the rows of any 8×3 subarray consist of two members of each of $\mathscr{A}_0^3 = \{(0, 0, 0), (1, 1, 1)\}$, $\mathscr{A}_1^3 = \{(1, 0, 0), (0, 1, 1)\}$, $\mathscr{A}_2^3 = \{(0, 1, 0), (1, 0, 1)\}$, $\mathscr{A}_3^3 = \{(0, 0, 1), (1, 1, 0)\}$.

$$\begin{array}{ccccccc}
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 1 & 1 \\
0 & 1 & 0 & 1 & 0 & 1 & 1 \\
0 & 1 & 1 & 1 & 1 & 0 & 0 \\
1 & 0 & 0 & 1 & 1 & 0 & 1 \\
1 & 0 & 1 & 1 & 0 & 1 & 0 \\
1 & 1 & 0 & 0 & 1 & 1 & 0 \\
1 & 1 & 1 & 0 & 0 & 0 & 1
\end{array}$$

Since the above array can be treated as an orthogonal array of strength 2, i.e. $OA(8, 7, 2, 2)$. From Theorem 3.5, we are able to construct an orthogonal array $OA(16, 8, 2, 3)$, which is exhibited below:

0 0 0 0 0 0 0 0 0
0 0 1 0 1 1 1 1 0
0 1 0 1 0 1 1 1 0
0 1 1 1 1 0 0 0 0
1 0 0 1 1 0 1 0 0
1 0 1 1 0 1 0 0 0
1 1 0 0 1 1 0 0 0
1 1 1 0 0 0 1 0 0
1 1 1 1 1 1 1 1 1
1 1 0 1 0 0 0 1 1
1 0 1 0 1 0 0 1 1
1 0 0 0 0 1 1 1 1
0 1 1 0 0 1 0 1 1
0 1 0 0 1 0 1 1 1
0 0 1 1 0 0 1 1 1
0 0 0 1 1 1 0 1 1

Hedayat et al. [10] studied the existence of difference schemes of strength t along with the methods for construction, and how to use them to construct orthogonal arrays of strength t.

4 Mixed Orthogonal Arrays

In previous sections we focused on orthogonal arrays in which all the factors had the same number of levels. However, in statistical application things can be more complicated and the factors in orthogonal arrays will have different number of levels. Such orthogonal arrays are called mixed or asymmetrical orthogonal arrays. While the definition and the Rao bounds for asymmetrical orthogonal arrays extend trivially from symmetrical ones, the construction is typically mathematically more challenging.

Previously the notation $OA(N, k, s, t)$ was used to denote a fixed-level orthogonal arrays. Now in order to keep the notation consistent, we use $OA(N, s^k, t)$ to denote fixed-level orthogonal arrays, where s^k indicates that there are k factors each at s levels. Thus, the following definition will be an extension of Definition 1.1, which allows the factors to have different levels.

Definition 4.1 A *mixed orthogonal orthogonal array* $OA(N, s_1^{k_1} s_2^{k_2} \cdots s_v^{k_v}, t)$ is an array of size $N \times k$, where $k = k_1 + k_2 + \cdots + k_v$ is the total number of factors, in which the first k_1 columns have symbols from $\{0, 1, \ldots, s_1 - 1\}$, the next k_2 columns have symbols from $\{0, 1, \ldots, s_2 - 1\}$ and so on, so that in any $N \times t$ subarray, every possible $t-$tuple occurs equally often as a row. For no good reasons in the literature the symbol $L_N(s_1^{k_1} s_2^{k_2} \cdots s_v^{k_v})$ has also been used.

Note that if $s_1, s_2, \ldots, s_v$ are equal, the orthogonal array will be a symmetrical one.

The Rao bound on the number of runs can also be extended to mixed orthogonal arrays. To state the bound, we first define the set that contains all $v-$tuples whose entries sum up to m, denoted by $I_m(v)$, where $m \geq 0$ and $v \geq 1$ are integers:

$$I_m(v) = \{(i_1, i_2, \ldots, i_v) : i_1 \geq 0, \ldots, i_v \geq 0, \sum_{l=1}^{v} i_l = m\}.$$

So $\sum_{I_m(v)}$ denotes the summation over all $v-$tuples in $I_m(v)$.

Theorem 4.1 (Rao's Inequalities for Mixed Orthogonal Arrays) *Consider an* $OA(N, s_1^{k_1} s_2^{k_2} \cdots s_v^{k_v}, t)$ *where, without loss of generality,* $s_1 \leq s_2 \leq \cdots \leq s_v$. *The parameters of the array satisfy the following inequalities:*

$$\begin{aligned}
N &\geq \sum_{m=0}^{u} \sum_{I_m(v)} \binom{k_1}{i_1} \cdots \binom{k_v}{i_v} (s_1 - 1)^{i_1} \cdots (s_v - 1)^{i_v}, \\
&\quad \text{if} \quad t = 2u, \\
N &\geq \sum_{m=0}^{u} \sum_{I_m(v)} \binom{k_1}{i_1} \cdots \binom{k_v}{i_v} (s_1 - 1)^{i_1} \cdots (s_{v-1} - 1)^{i_{v-1}} (s_v - 1)^{i_v} \\
&\quad + \sum_{I_u(v)} \binom{k_1}{i_1} \cdots \binom{k_{v-1}}{i_{v-1}} \binom{k_v - 1}{i_v} (s_1 - 1)^{i_1} \cdots (s_{v-1} - 1)^{i_{v-1}} (s_v - 1)^{i_v + 1}, \\
&\quad \text{if} \quad t = 2u + 1,
\end{aligned} \tag{12}$$

for $u \geq 0$.

Specifically, for an $OA(N, s_1^{k_1} s_2^{k_2}, 2)$, the Rao bound states that

$$N \geq 1 + k_1(s_1 - 1) + k_2(s_2 - 1).$$

However, the Rao bound can be improved under many parameters. Sloane and Stufken [23] established the linear programming bound for mixed orthogonal arrays using Delsarte's theory, which is an improvement to the Rao bound.

Most methods for constructing mixed orthogonal arrays apply only to arrays of strength 2. Here we mainly focus on the expansive replacement method. Difference schemes are also powerful tools for constructing orthogonal arrays of strength 2 and details can be found in Hedayat et al. [9]. A general method for construction is also established in Suen et al. [24], which is based on the construction of fixed level orthogonal arrays given by Bose and Bush [1].

Let A be an (mixed or fixed) orthogonal array of strength 2 in which factor 1 has s_1 levels. Let T be an (mixed or fixed) orthogonal array of strength 2 with s_1 runs. Then by making a one-to-one correspondence between the levels of factor 1 in A and

the runs of T, i.e. replacing each level of factor 1 in A by the corresponding run in T, we can obtain a new orthogonal array B of strength 2 which contains at least as many factors as A. This method of construction is referred to as the *expansive replacement method*.

Example 4.1 By replacing the levels of the first factor in A, an $OA(8,\ 2^4\,4^1,\ 2)$, by the corresponding runs in T, we can obtain B, an $OA(8,\ 2^7,\ 2)$.

$$A = \begin{bmatrix} 0&0&0&0&0 \\ 0&1&1&1&1 \\ 1&0&0&1&1 \\ 1&1&1&0&0 \\ 2&0&1&1&0 \\ 2&1&0&0&1 \\ 3&0&1&0&1 \\ 3&1&0&1&0 \end{bmatrix} \qquad B = \begin{bmatrix} 0&0&0&0&0&0&0 \\ 0&0&0&1&1&1&1 \\ 0&1&1&0&0&1&1 \\ 0&1&1&1&1&0&0 \\ 1&0&1&0&1&1&0 \\ 1&0&1&1&0&0&1 \\ 1&1&0&0&1&0&1 \\ 1&1&0&1&0&1&0 \end{bmatrix}$$

Here

$$T = \begin{bmatrix} 0&0&0 \\ 0&1&1 \\ 1&0&1 \\ 1&1&0 \end{bmatrix}$$

In the first factor of A, level 0, 1, 2 and 3 correspond to the first, second, third and last runs in T, respectively.

Conversely, in Example 4.1, notice that by contracting the first three factors in B, A can be obtained through mapping $(0, 0, 0) \to 0$, $(0, 1, 1) \to 1$, $(1, 0, 1) \to 2$ and $(1, 1, 0) \to 3$. This way of construction is called *contraction replacement method*. Sometimes by selecting a set of factors in an orthogonal array of strength 2 and replacing it by a single factor with a larger number of levels, the resulting array can also be an orthogonal array. However, in order to use this method, the original array should possess certain structure: A is an orthogonal array $OA(N, s_1 s_2 \cdots s_k, 2)$ (repeated s_i allowed) such that for a subarray containing u factors of A, the runs of the subarray consist of N/N_1 copies of each of the runs of an $OA(N_1, s_1 s_2 \cdots s_u, 2)$, say T. T satisfies the condition: $N_1 = 1 + \sum_{i=1}^{u}(s_i - 1)$. Then by labeling the runs of T by $0, 1, \ldots, N_1 - 1$ and replacing each runs of the subarray of A by the corresponding label, we can obtain an $OA(N, N_1 s_{u+1} \ldots s_k, 2)$, say B.

5 Application in Factorial Experiments

The goal of factorial experiments is often to study the effect of different levels of various factors on a response variable of interest. Factors that might affect the response variable are identified, then for each of the factors two or more levels are

selected in the experiment. Once the factors and the levels are determined, a collection of all possible level combinations can be obtained. However, it's often impossible to make measurements at each of the level combinations since the number of level combinations is often too large. As a result, a subset of level combinations to be used in the experiment must be selected, resulting in a fractional factorial experiment. Obviously, not all selections are good. In the selection process, an orthogonal array can help us select level combinations with desirable statistical properties.

Suppose $A_1, \ldots, A_k$ are the factors to be included in the experiment with $s_1, \ldots, s_k$ levels, respectively. The level combinations can be represented by the $k-$tuples $(j_1, \ldots, j_k)$, where $0 \leq j_l \leq s_l - 1, l = 0, \ldots, k$. Let L be the set of all possible level combinations and M be the cardinality of L. There are N experimental units and each will be assigned to a specific level combination randomly.

Let Y be $N \times 1$ vector of response random variable $Y_{\tau j}$, where $Y_{\tau j}$ denotes the response of $j-$th unit that is assigned to level combination $\tau \in L$. μ is the $M \times 1$ population mean vector and ϵ is the $N \times 1$ vector of random errors. The model can be written as

$$Y = X\mu + \epsilon \tag{13}$$

where X is an $N \times M$ matrix consists of 0's and 1's. The entries are obtained by following the procedure: Label the columns by the level combinations in reverse lexicographic order and the rows by the subscripts of corresponding entries in Y, the entries in $(\tau j, \tau')$ equals 1 if $\tau = \tau'$ and 0 otherwise.

A *treatment contrast* can be used to compare population means for different level combinations. The treatment contrast is a linear combination of the population means with coefficients add up to 0, i.e. for an $M \times 1$ vector c, $c^\top \mu$ is called a treatment contrast if $c^\top 1_M = 0$. A treatment contrast $c^\top \mu$ is said to be *estimable* under a particular model and selection of level combinations if there exists an $N \times 1$ vector b such that $E(b^\top Y) = c^\top \mu$, such a $b^\top Y$ is called an *unbiased estimator* of $c^\top \mu$. Two treatment contrasts $c_1^\top \mu$ and $c_2^\top \mu$ are said to be *orthogonal* if $c_1^\top c_2 = 0$.

In a factorial experiment, an analysis of data and how factors affect the response variable would be based on studying the main-effects and interactions of factors, which are pairwise orthogonal treatment contrasts. However, in most experiments not all included factors are important for explaining the variability in response variables. So we can assume that only a small subset of the effects will suffice to understand how important factors influence the response. This is called *effects sparsity assumption*, and it's essential in order to justify the use of fractional factorials.

Based on the concepts and assumptions, a reduced model of (13) can be built in order to estimate the components of the effects that we are interested in, in which the population mean vector μ is re-parameterized into the parameters that we want to estimate, γ_1, and the nuisance parameters, γ_2. Details of how to build the reduced model can be found in Chap. 11 of Hedayat et al. [12].

The rows of an orthogonal array can be used to specify the fractional factorials. Additionally, the strength t of an orthogonal array is related to estimability of para-

meters under the reduced model when using the runs as fractional factorials. Then the following results can be verified.

Theorem 5.1 *Under certain models, if an OA*$(N, s_1 s_2 \ldots s_k, t)$, $t \geq 2$, *is used in a factorial experiment, then:*

(i) If t is even, γ_2 *is absent and* γ_1 *consists of the intercept parameter, all components of main-effects and all components of interactions of at most* $t/2$ *factors, then all elements of* γ_1 *are estimable.*

(ii) If t is odd, γ_1 *consists of the intercept parameter, all components of main-effects and all components of interactions of at most* $(t-1)/2$ *factors while* γ_2 *consists of all components of interactions of precisely* $(t+1)/2$ *factors, then all elements of* γ_1 *are estimable.*

There are also many other desirable properties of orthogonal arrays for all kinds of experiments. Check Hedayat et al. [12] to see details and related references.

6 Application in Computer Experiments

6.1 Orthogonal Latin Hypercube Designs

Example 6.1 Consider a known function

$$Y = f(\mathbf{X}),$$

where $Y \in R, \mathbf{X} \in R^k$. Random vector $\mathbf{X} = (X^{(1)}, \ldots, X^{(k)})$ has a uniform distribution on the unit hypercube $[0, 1]^k$.

Suppose we want to estimate the mean of the random variable Y, but sometimes it takes a large amount of efforts or money to compute f, or k is very large. Then the problem can be converted to the problem of finding the integral of $f(x)$ with respect to the uniform measure on $[0, 1]^k$.

The above example shows a problem of evaluating a complex integral (maybe over a high-dimensional domain), which is a problem we often see in scientific study. Monte Carlo method is a very useful approach to solve high-dimensional integration problems, and thus a good choice for solving this kind of problems. McKay et al. [15] introduced Latin hypercube sampling (LHS) as an alternative to iid sampling when selecting the points in Monte Carlo method to tackle this kind of problems. Further research results showed that LHS lead to a smaller variance compared with the iid variance.

Definition 6.1 A *Latin hypercube* is an $N \times k$ matrix, in which each column is a permutation of $1, 2, \ldots, N$. Denoted by LH(N, k). A Latin hypercube is an $OA(N, k, N, 1)$.

Latin hypercube designs (LHD) are used in physical and computer experiments. The main property of LHS is that it stratifies each univariate margin simultaneously, thus filtering out the main effects. Similarly, one would expect that all the bivariate interactions as well as the main effects can also be filtered out if stratification is achieved on each bivariate margin. In experimental designs, the uniformity properties of OA designs are quite desirable since design points can be distributed in the design region uniformly. However, OA designs are not suitable when a large number of factors are to be studied but only a few of them are virtually effective. LHD's can be good alternatives in this case, but they can't even guarantee the projection of design points onto bivariate margins to be uniformly distributed.

Tang [25] showed that orthogonal arrays of strength t can be used to construct LHD's. He stated that such OA-based LHD's have the $t-$variate uniformity properties and can stratify each $t-$dimensional margin while keeping the univariate stratification property. As a result, such OA-based LHD's are more appropriate for computer experiments than general LHD's, and can also be applied in physical experiments. The method of construction is as follows:

Let A be an $OA(N, k, s, t)$. For each factor of A, replace the N/s positions with entries m, $m \in \{0, 1, \ldots, s-1\}$ by a permutation of $mN/s + 1, mN/s + 2, \ldots, (m+1)N/s$ for all $m = 0, 1, \ldots, s-1$. Then the resulting matrix, say U, is obviously a Latin hypercube. Additionally, inheriting from A, U achieves the uniformity in each $t-$variate margin.

How to use OA-based LH for numerical integration? For example, in the case of Example 6.1, first we can obtain a random OA-based LH, denoted by $U = (u_{ij})$, by randomizing the rows, columns and symbols in an orthogonal array $A = OA(N, k, s, r)$ and then replacing the N/s positions with entry m by a random permutation shown above, for all $m = 0, 1, ..., s-1$. Suppose in $\mathbf{X}$, each entry $X_i^{(j)} \sim Unif((i-1)/N, i/N), i = 1, \ldots, N, j = 1 \ldots, k$, are randomly generated. Then the N points to be used for integration are selected and formed by $\mathbf{X}_i = (X_{u_{i1}}^{(1)}, \ldots, X_{u_{ik}}^{(k)}), i = 1, \ldots, N$. This procedure is called U sampling.

Tang [25] also presents many other results, such as the study of U sampling constructed by $OA(s^2, k, s, 2)$, since such an array leads to the smallest possible sample size $N = s^2$ for a given s, which can save time.

6.2 *Sliced Orthogonal Arrays*

The motivation for sliced orthogonal arrays is to provide a systematic study to address the experimental planning issue in the study of computer experiments with qualitative and quantitative factors. Since the existing space-filling designs, such as Latin hypercube designs assume that all input factors to be quantitative. Another kind of space-filling designs, sliced space-filling designs are introduced by Qian and Wu [17].

The basic approach start with a Latin hypercube design for the quantitative factors. Then partition the design into groups corresponding to different levels of qualitative factors with each achieving uniformity in lower dimension.

Definition 6.2 An $OA(N_1, k, s_1, t)$, say A, is called a *sliced orthogonal array* if its rows can be partitioned into $v = N_1/N_2$ arrays B_i, each with N_2 rows, such that after mapping the s_1 levels of B_i to s_2 levels with $s_1 > s_2$, each B_i turns into an $OA(N_2, k, s_2, t)$.

Example 6.2 The example shows a sliced orthogonal array $A = OA(16, 3, 4, 2)$, after being partitioned into 4 arrays B_i (shown in (a)) and through the mapping:

$$0, 3 \to 0; \quad 1, 2 \to 1$$

each small array B_i turns into an $OA(4, 3, 2, 2)$, $1 \le i \le 4$ (shown in (b)).

$$A = \left[\begin{array}{ccc} 0&0&0\\ 0&1&1\\ 1&0&1\\ 1&1&0\\ \hline 0&3&3\\ 0&2&2\\ 1&3&2\\ 1&2&3\\ \hline 3&3&0\\ 3&2&1\\ 2&3&1\\ 2&2&0\\ \hline 3&0&3\\ 3&1&2\\ 2&0&2\\ 2&1&3 \end{array}\right] \quad (a) \qquad \left[\begin{array}{ccc} 0&0&0\\ 0&1&1\\ 1&0&1\\ 1&1&0\\ \hline 0&0&0\\ 0&1&1\\ 1&0&1\\ 1&1&0\\ \hline 0&0&0\\ 0&1&1\\ 1&0&1\\ 1&1&0\\ \hline 0&0&0\\ 0&1&1\\ 1&0&1\\ 1&1&0 \end{array}\right] \quad (b)$$

The idea of constructing a sliced space-filling design using sliced orthogonal array is as follows: Let A be a sliced orthogonal array which can be used to generate an OA-based Latin hypercube design by following the procedure in Sect. 6.1. The resulting design is an OA-based LHD for quantitative factors, denoted by D, and it's partitioned into different groups D_i, where each D_i consists of points corresponding to B_i. Different D_i's are associated with different level combinations of qualitative factors, so the design points for the quantitative factors achieve uniformity in low dimensions for any qualitative factor level combination. Now the array $(D_1, \ldots, D_v)$ is a sliced space-filling design.

Qian and Wu [17] also introduced some statistical properties of sliced space-filling designs. Further, several methods for the construction of sliced orthogonal arrays were also provided.

6.3 Nested Orthogonal Arrays

Nested orthogonal arrays have been applied in the construction of space-filling designs when we want to combine two experiments, the expensive one with higher accuracy to be nested in a relatively inexpensive one of lower accuracy.

Definition 6.3 A nested orthogonal array $NOA((N_1, N_2), k, (s_1, s_2), t)$ is an $OA(N_1, k, s_1, t)$, say A, which contains an $OA(N_2, k, s_2, t)$, where $N_2 < N_1, s_2 < s_1$.

Let $r = s_1/s_2$. Mukerjee et al. [16] studies the existence of nested orthogonal arrays and obtains a Rao-type bound, shown as follows.

Theorem 6.1 *A necessary condition on N_1 for the existence of $OA((N_1, N_2), k, (s_1, s_2), t)$ is*

$$\begin{aligned} N_1 &\geq N_2\left(1 + k(r-1) + \ldots + \binom{k}{u}(r-1)^u\right), \text{ if } t = 2u, \\ N_1 &\geq N_2\left[1 + k(r-1) + \ldots + \binom{k}{u}(r-1)^u + \binom{k-1}{u}(r-1)^{u+1}\right], \text{ if } t = 2u+1. \end{aligned} \tag{14}$$

Theorem 6.2 *A necessary condition on k for the existence of $OA((N_1, N_2), k, (s_1, s_2), t)$ is*

$$k \leq \frac{N_1 - N_2 r^{t-2}}{N_2 r^{t-2}(r-1)} + t - 2. \tag{15}$$

This bound can be attained for some cases. Mukerjee et al. [16] provides a detailed proof of this theorems with some examples discussed. We omit the details here.

Dey [5] provides some methods of construction for nested orthogonal arrays, and extends the definition and constructions to asymmetric nested orthogonal arrays.

7 Other Topics

7.1 Latin Squares and F-Squares

The areas of Latin squares and F-squares have attracted numbers of researchers. The connection between them and orthogonal arrays are also fascinating.

Definition 7.1 A *Latin square of order s* is an $s \times s$ array with entries from a set S with cardinality s, such that in every row and every column each element of S appears exactly once.

Definition 7.2 Two Latin squares of order s are said to be *orthogonal* to each other if upon superimposition of one on the other, the ordered pairs (i, j) of corresponding

entries consist of all possible s^2 pairs. A Latin square is *orthogonally mateless or isolated* if there is no Latin square orthogonal to it.

A collection of w Latin squares of order s is called a set of *pairwise orthogonal Latin squares* (or mutually orthogonal Latin squares), if any pair of Latin squares from the collection are orthogonal to each other, denote the collection by $POL(s, w)$ (or $MOLS(s, w)$).

It can be verified that $1 \leq w \leq s-1$. A $POL(s, s-1)$ is called a *complete* set of pairwise orthogonal Latin square of order s.

Pairwise orthogonal Latin squares can be converted into orthogonal arrays of strength 2 and vice versa.

Theorem 7.1 *If a $POL(s, k)$ exists, then an $OA(s^2, k+2, s, 2)$ also exists. In particular, if s is a prime power, then an $OA(s^2, s+1, s, 2)$ exists.*

Theorem 7.2 *An orthogonal array $OA(s^2, k+2, s, 2), k \geq 2$ exists only if a $POL(s, k)$ exists.*

The more general concepts, F-squares are more flexible, though not as well developed as Latin squares.

Definition 7.3 An *F-square* is an $n \times n$ array based on s symbols, such that in every row and every column each symbol appears n/s times.

Definition 7.4 Two $n \times n$ F-squares are said to be *orthogonal* to each other if upon superimposition of one on the other, the ordered pairs (i, j) of corresponding entries contain all possible s^2 pairs n^2/s^2 times.

Similarly we can define a set of *pairwise orthogonal $n \times n$ F-squares* and denote it by $POF(n, s, w)$.

The restriction on w for which $POF(n, s, w)$ exists along with the proof can also obtained, according to Hedayat et al. [8].

Theorem 7.3 *A necessary condition for the existence of w pairwise orthogonal $n \times n$ F-squares based on s symbols is*

$$w \leq (n-1)^2/(s-1). \tag{16}$$

Note that a $POF(s, s, w)$ is just a $POL(s, w)$. A $POF(n, s, w)$ that attained the equality in (16) is *complete*.

We can also convert pairwise orthogonal F-squares into orthogonal arrays of strength 2.

Theorem 7.4 *The existence of a $POF(n, s, k_1)$ implies the existence of an orthogonal array $OA(n^2, k_1+2, s, 2)$. In addition, if an $OA(n, k_2, s, 2)$ exists, then an orthogonal array $OA(n^2, k_1+2k_2, s, 2)$ can be constructed.*

However, the converse of Theorem 7.4 is not as straightforward as Theorem 7.2.

7.2 Compound Orthogonal Arrays

First proposed by Rosenbaum [21, 22] and further studied by Hedayat and Stufken [11], compound orthogonal arrays can be used in the study of dispersion effects in factorial designs.

Definition 7.5 Let

$$B = \begin{bmatrix} b_1 \\ \vdots \\ b_{N_1} \end{bmatrix}$$

be an OA($N_1, s_1 \dots s_{k_1}, t_1$) and C_i all be arrays with parameters OA($N_2, u_1 \dots u_{k_2}, t_2$), $i = 1, \dots, N_1$. Then form the $N_1N_2 \times (k_1 + k_2)$ array

$$A = \begin{bmatrix} b_1 & \\ \vdots & C_1 \\ b_1 & \\ \vdots & \vdots \\ b_{N_1} & \\ \vdots & C_{N_1} \\ b_{N_1} & \end{bmatrix}.$$

Suppose A itself is an orthogonal array of strength t_3. An array with this structure is called a compound orthogonal array of type (t_1, t_2, t_3), denoted by

$$COA((N_1, N_2), (k_1, k_2), (s_1 \dots s_{k_1}, u_1 \dots u_{k_2}), (t_1, t_2, t_3)).$$

Here some C_i could have greater strengths, and in this case t_2 represents the minimum strength of all C_i.

Compound orthogonal arrays in which the parameters satisfy

$$N_1 \leq s_1 s_2 \dots s_{k_1} \tag{17}$$

$$N_2 \leq u_1 u_2 \dots u_{k_2} \tag{18}$$

are of great interests. Four cases are distinguished in order to facilitate the discussion, which are when both (17) and (18) attain equality; or (17) doesn't attain equality but (18) does; or (17) attains equality but (18) doesn't; or both (17) and (18) fail to attain equality. The goal is to obtain the maximal values for t_1, t_2, t_3. For details of the discussions, see Hedayat and Stufken [11].

Acknowledgements Research is supported by the U.S. National Science Foundation Grant DMS-1306394. My friend Professor John Stufken and my research assistant Xindi Wang contributed immensely in the preparation of this article. I am grateful to both of them.

References

1. Bose, R.C., Bush, K.A.: Orthogonal arrays of strength two and three. Ann. Math. Statist. **23**, 508–524 (1952)
2. Bose, R.C.: On some connections between the design of experiments and information theory. Bull. Internat. Statist. Inst. **38**, 257–271 (1961)
3. Bush, K.A.: Orthogonal arrays of index unity. Ann. Math. Statist. **23**, 426–434 (1952b)
4. Delsarte, P.: An algebraic approach to the association schemes of coding theory. Philips Res. Reports Suppl. **10** (1973)
5. Dey, A.: Construction of nested orthogonal arrays. Discrete Math. **310**, 2831–2834 (2010)
6. Hadamard, J.: Résolution d'une question relative aux déterminants. Bull. des Sci. Math. **17**, 240–246 (1893)
7. Hamming, R.W.: Error-detecting and error correcting codes. Bell Syst. Tech. J. **29**, 147–160 (1950)
8. Hedayat, A.S., Raghavarao, D., Seiden, E.: Further contributions to the theory of $F-$squares design. Ann. Statist. **3**, 712–716 (1975)
9. Hedayat, A.S., Pu, K., Stufken, J.: On the construction of asymmetrical orthogonal arrays. Ann. Statist. **20**, 2142–2152 (1992)
10. Hedayat, A.S., Stufken, J., Su, G.: On difference schemes and orthogonal arrays of strength t. J. Statist. Plann. Inf. **56**, 307–324 (1996)
11. Hedayat, A.S., Stufken, J.: Compound orthogonal arrays. Technometrics **41**, 57–61 (1999)
12. Hedayat, A.S., Sloane. N.J.A., Stufken, J.: Orthogonal Arrays: Theory and Applications. Springer, New York (1999)
13. Jungnickel, D.: On difference matrices, resolvable transversal designs and generalized Hadamard matrices. Mathematische Zeitschrift **167**, 49–60 (1979)
14. MacWilliams, F.J., Sloane, N.J.A.: The Theory of Error-Correcting Codes. North-Holland, Amsterdam (1977)
15. McKay, M.D., Beckman, R.J., Conover, W.J.: A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics **21**, 239–245 (1979)
16. Mukerjee, R., Qian, P.Z.G., Wu, C.F.J.: On the existence of nested orthogonal arrays. Discrete Math. **308**, 4635–4642 (2008)
17. Qian, P.Z.G., Wu, C.F.J.: Sliced space-filling designs. Biometrika **96**, 945–956 (2009)
18. Rao, C.R.: Hypercubes of strength d leading to confounded designs in factorial experiments. Bull. Calcutta Math. Soc. **38**, 67–78 (1946a)
19. Rao, C.R.: Factorial experiments derivable from combinatorial arrangements of arrays. J. Royal Statist. Soc. (Suppl.) **9**, 128–139 (1947)
20. Rao, C.R.: On a class of arrangements. Proc. Edinburgh Math. Soc. **8**, 119–125 (1949)
21. Rosenbaum, P.R.: Dispersion effects from fractional factorials in Taguchi's method of quality design. J. Royal Statist. Soc. B **56**, 641–652 (1994)
22. Rosenbaum, P.R.: Some useful compound dispersion experiments in quality design. Technometrics **38**, 354–364 (1996)
23. Sloane, N.J.A., Stufken, J.: A linear programming bound for orthogonal arrays with mixed levels. J. Statist. Plann. Inf. **56**, 295–305 (1996)
24. Suen, C.-Y., Das, A., Dey, A.: On the construction of asymmetric orthogonal arrays. Statist. Sinica **11**, 241–260 (2001)
25. Tang, B.: Orthogonal array-based Latin hypercubes. J. Am. Statist. Assoc. **88**, 1392–1397 (1993)

Hoffman's Coclique Bound for Normal Regular Digraphs, and Nonsymmetric Association Schemes

Hadi Kharaghani and Sho Suda

Abstract We extend Hoffman's coclique bound for regular digraphs with the property that its adjacency matrix is normal, and discuss cocliques attaining the inequality. As a consequence, we characterize skew-Bush-type Hadamard matrices in terms of digraphs. We present some normal digraphs whose vertex set is decomposed into disjoint cocliques attaining the bound. The digraphs provided here are relation graphs of some nonsymmetric association schemes.

Keywords Hoffman's coclique bound · Association scheme · Skew-Bush-type · Hadamard matrix · Regular biangular matrix · Twin asymmetric design

2010 Mathematics Subject Classification: 05C62 · 05B20 · 05E30

1 Introduction

The method of discovery of certain properties of a graph from different parameters of its adjacency matrix is an area of graph theory which is referred to as *Spectral Graph Theory*. For example, using the eigenvalues or eigenspaces of the adjacency matrix of a graph, several inequalities for parameters of the graph, such as the clique size, the independent number, the chromatic number, etc. are obtained, see [4] for details.

Hoffman has given an upper bound for the independent number of regular graphs, see [4, 6]. In this paper we extend the Hoffman's bound for normal regular digraphs. Here, a normal digraph means a digraph with its adjacency matrix being normal,

H. Kharaghani (✉)
Department of Mathematics and Computer Science, University of Lethbridge, Lethbridge, AB T1K 3M4, Canada
e-mail: kharaghani@uleth.ca

S. Suda
Department of Mathematics Education, Aichi University of Education, 1 Hirosawa, Igaya-cho, Kariya, Aichi 448-8542, Japan
e-mail: suda@auecc.aichi-edu.ac.jp

T. Abualrub et al. (eds.), *Mathematics Across Contemporary Sciences*, Springer Proceedings in Mathematics & Statistics 190,
DOI 10.1007/978-3-319-46310-0_8

see Sect. 2.1. We also study normal digraphs with their coclique size attaining the upper bound. In Sect. 4, we use the bound to characterize skew-Bush-type (or skew-checkered) Hadamard matrices in terms of doubly regular asymmetric digraphs with specific properties. This result is an analog of a result of Wallis [13], where it is shown that there exists a symmetric Bush-type Hadamard matrix of order $4n^2$ if and only if a strongly regular graph with parameters $(4n^2, 2n^2 - n, n^2 - n, n^2 - n)$ exists such that the vertex set is decomposed into $2n$ disjoint cocliques of size $2n$. Note that the cocliques of size $2n$ in any strongly regular graph attain Hoffman's bound.

A coclique attaining Hoffman's coclique bound in a strongly regular graph Γ is a clique attaining the clique bound in the complement of Γ. A *spread of a strongly regular graph* is a set of disjoint cliques attaining the clique bound and covering all vertices. In [1, 4, 6], the spread of strongly regular graphs are extensively studied. We provide in Sects. 5 and 6 some normal digraphs, which are all the relation graphs of some association schemes, in such a way that their vertex set are decomposed into disjoint cocliques with their sizes attaining the upper bound.

2 Preliminaries

2.1 Digraphs

A *digraph* Γ is a pair (X, E) such that the *vertex set* X is a finite set and the *edge set* or *arc set* E is a subset of $X \times X$ with $E \cap \{(x, x) \mid x \in X\} = \emptyset$. The *adjacency matrix* of Γ is a $(0, 1)$-matrix with rows and columns indexed by the elements of X such that $A_{xy} = 1$ if $(x, y) \in E$ and $A_{xy} = 0$ otherwise. A digraph Γ is *asymmetric* if $(x, y) \in E$ implies $(y, x) \notin E$, namely $A + A^T$ is a $(0, 1)$-matrix, where A^T denotes the transpose of A. A digraph Γ is *normal* if the adjacency matrix A is normal, namely $AA^T = A^T A$ holds. A digraph Γ is *k-regular* if $|\{y \in X \mid (x, y) \in E\}| = |\{y \in X \mid (y, x) \in E\}| = k$ for any vertex x.

A digraph Γ is *normally regular with parameters* (n, k, λ, μ) if Γ is asymmetric, the number of vertices of Γ is n and the adjacency matrix A of Γ satisfies

$$AA^T = kI_n + \lambda(A + A^T) + \mu(J_n - I_n - A - A^T), \tag{1}$$

where I_n is the identity matrix of order n and J_n is the all ones matrix of order n. It was shown in [9] that a normally regular digraph is indeed normal. A *doubly regular asymmetric digraph* Γ with parameters (v, k, λ) is a normally regular digraph with parameters (v, k, λ, λ).

A subset C in X is a *coclique* (or an *independent set*) in Γ if $(x, y) \notin E$ for any $x, y \in C$.

A digraph Γ is *strongly connected* if for any distinct vertices x, y, there exist vertices $x_0, \ldots, x_s$ such that $x_0 = x, x_s = y$ and $(x_i, x_{i+1}) \in E$ for any $i \in \{0, 1, \ldots, s-1\}$.

The following lemma will be used in Proposition 4.1.

Lemma 2.1 *Let Γ be a normally regular digraph with parameters (n, k, λ, μ) with adjacency matrix A. Assume that $A + A^T = J_n - I_r \otimes J_{n/r}$ for some positive integer r dividing n. Then the eigenvalues of A are k, $\pm\sqrt{-k+\mu}$, or $-n/(2r) \pm \sqrt{k - \mu + (-\lambda+\mu)n/r - n^2/(4r^2)}$.*

Proof The valency k is an eigenvalue of A with the all-ones vector as an eigenvector. Let α be an eigenvalue whose eigenvector is orthogonal to the all-ones vector. By the Eq. (1), we have

$$\alpha\overline{\alpha} = k + \lambda(\alpha + \overline{\alpha}) + \mu(-1 - \alpha - \overline{\alpha}). \tag{2}$$

Since $A + A^T = J_n - I_r \otimes J_{n/r}$, the real part of α is $-n/(2r)$ or 0. By (2), α is the desired value. □

2.2 Association Schemes

A *commutative association scheme of class d* with vertex set X of size n is a set of non-zero $(0, 1)$-matrices $A_0, \ldots, A_d$, which are called *adjacency matrices*, with rows and columns indexed by X, such that:

(i) $A_0 = I_n$.
(ii) $\sum_{i=0}^{d} A_i = J_n$.
(iii) For any $i \in \{0, 1, \ldots, d\}$, $A_i^T \in \{A_0, A_1, \ldots, A_d\}$.
(iv) For any $i, j \in \{0, 1, \ldots, d\}$, $A_i A_j = \sum_{k=0}^{d} p_{ij}^k A_k$ for some p_{ij}^k's.
(v) For any $i, j \in \{0, 1, \ldots, d\}$, $A_i A_j = A_j A_i$.

The association scheme is said to be *symmetric* if all A_i are symmetric, *nonsymmetric* otherwise. The *intersection matrix* B_i ($i \in \{0, 1, \ldots, d\}$) is defined as follows: $B_i = (p_{ij}^k)_{j,k=0}^{d}$.

A digraph $\Gamma = (X, E)$ is a *relation graph* of an association scheme with vertex set X if the adjacency matrix of Γ is one of the adjacency matrices of the association scheme.

The vector space spanned by the A_i's forms a commutative algebra, denoted by $\mathcal{A}$ and called the *Bose-Mesner algebra* or *adjacency algebra*. There exists a basis of $\mathcal{A}$ consisting of primitive idempotents, say $E_0 = (1/n)J_n, E_1, \ldots, E_d$. Since $\{A_0, A_1, \ldots, A_d\}$ and $\{E_0, E_1, \ldots, E_d\}$ are two bases of $\mathcal{A}$, there exist the change-of-bases matrices $P = (P_{ij})_{i,j=0}^{d}$, $Q = (Q_{ij})_{i,j=0}^{d}$ so that

$$A_j = \sum_{i=0}^{d} P_{ij} E_i, \quad E_j = \frac{1}{n} \sum_{i=0}^{d} Q_{ij} A_i.$$

The matrix P (Q respectively) is said to be the *first (second respectively) eigenmatrix*.

3 Hoffman's Bound for Normal Digraphs

In this section, we give an upper bound for the size of cocliques in a normal digraph in terms of eigenvalues of the adjacency matrix of the digraph Γ. The upper bound is referred to as the *Hoffman bound*. For a digraph with adjacency matrix A, define $\theta_{\min} = \min\{\mathrm{Re}(\theta) \mid \theta \text{ is an eigenvalue of } A\}$, $\mathrm{Re}(\theta)$ is the real part of θ. Note that for a normal graph which is not the empty graph, $\theta_{\min}$ is negative since the trace of A is zero.

Proposition 3.1 *Let n, k be positive integers. Let $\Gamma = (X, E)$ be a strongly connected k-regular normal digraph with n vertices and adjacency matrix A. For a coclique C in Γ,*

$$|C| \leq \frac{n(-\theta_{\min})}{k - \theta_{\min}} \tag{3}$$

holds. Moreover the following hold.

(i) If equality holds in (3), then $|\{y \in C \mid (x, y) \in E\}| + |\{y \in C \mid (y, x) \in E\}| = -2\theta_{\min}$ for any $x \in X \setminus C$.

(ii) If equality holds in (3) and the number of eigenvalues with real part equal to $\theta_{\min}$ is exactly one, then $|\{y \in C \mid (x, y) \in E\}| = -\theta_{\min}$ for any $x \in X \setminus C$.

Proof Let $\theta_1, \ldots, \theta_{l+2m}$ be the eigenvalues of A such that $\theta_i \in \mathbb{R}$ for any $i \in \{1, \ldots, l\}$ and $\overline{\theta_{l+j}} = \theta_{l+m+j} \in \mathbb{C} \setminus \mathbb{R}$ for any $j \in \{1, \ldots, m\}$. Let E_i be the orthogonal projection onto the eigenspace of θ_i. Then $\overline{E_{l+j}} = E_{l+m+j}$ for any $j \in \{1, \ldots, m\}$. Since k is an eigenvalue, we set $\theta_1 = k$. Since Γ is strongly connected, $E_1 = \frac{1}{n} J_n$.

Let χ be the characteristic column vector of C. Since C is a coclique of Γ, it holds that

$$\chi^T A \chi = 0. \tag{4}$$

On the other hand we estimate the value $\chi^T A \chi$ to use the formula $A = \sum_{i=1}^{l+2m} \theta_i E_i$ and $\chi^T \overline{E_{l+j}} \chi = \chi^T E_{l+m+j} \chi$ for $j \in \{1, \ldots, m\}$ as follows:

$$\begin{aligned}\chi^T A\chi &= \chi^T\Big(\sum_{i=1}^{l+2m}\theta_i E_i\Big)\chi = \sum_{i=1}^{l+2m}\theta_i\chi^T E_i\chi\\ &= k\chi^T E_1\chi + \sum_{i=2}^{l}\theta_i\chi^T E_i\chi + \sum_{i=1}^{2m}\frac{\theta_{l+i}+\overline{\theta_{l+i}}}{2}\chi^T E_{l+i}\chi\\ &\geq k\chi^T E_1\chi + \theta_{\min}\sum_{i=2}^{l+2m}\chi^T E_i\chi\\ &= k\chi^T E_1\chi + \theta_{\min}\chi^T(I_n - E_1)\chi\\ &= \frac{(k-\theta_{\min})|C|^2}{n} + \theta_{\min}|C|. \end{aligned} \tag{5}$$

Combining (4) and (5), we obtain $|C| \leq n(-\theta_{\min})/(k-\theta_{\min})$.

A coclique C meets the upper bound if and only if $\chi^T E_i\chi = 0$ for i such that $i \geq 2$ and $\mathrm{Re}(\theta_i) \neq \theta_{\min}$.

(i): Let $A + A^T = \sum_{i=1}^{t}\tau_i F_t$ be the spectrum decomposition of $A + A^T$, and set $\tau_1 = 2k$ and $\tau_t = 2\theta_{\min}$. Since $F_i\chi = 0$ for $i \in \{2, 3, \ldots, t-1\}$,

$$\begin{aligned}(A + A^T)\chi &= 2kF_1\chi + \tau_t F_t\chi = 2kF_1\chi + \tau_t\sum_{i=2}^{t}F_i\chi = 2kF_1\chi + \tau_t(I_n - F_1)\chi\\ &= \tau_t\chi + (2k-\tau_t)\frac{1}{n}J_n\chi = \tau_t\chi + (2k-\tau_t)\frac{|C|}{n}\mathbf{1} = (-2\theta_{\min})(\mathbf{1}-\chi),\end{aligned} \tag{6}$$

where $\mathbf{1}$ is the all-ones vector. Equation (6) is equivalent to the condition that the $|\{y \in C \mid (x, y) \in E\}| + |\{y \in C \mid (x, y) \in E\}| = -2\theta_{\min}$ for any $x \in X \setminus C$.

(ii): Let θ_s satisfy $\mathrm{Re}(\theta_s) = \theta_{\min}$. Then $\theta_s = \theta_{\min}$. Indeed, if $\theta_s \in \mathbb{C} \setminus \mathbb{R}$, then $\overline{\theta_s}$ also satisfies $\mathrm{Re}(\overline{\theta_s}) = \theta_{\min}$. This contradicts the assumption. In this case,

$$\begin{aligned}A\chi &= kE_1\chi + \theta_s E_s\chi = kE_1\chi + \theta_s\sum_{i=2}^{s}E_i\chi = kE_1\chi + \theta_s(I_n - E_1)\chi\\ &= \theta_s\chi + (k-\theta_s)\frac{1}{n}J_n\chi = \theta_s\chi + (k-\theta_s)\frac{|C|}{n}\mathbf{1} = (-\theta_{\min})(\mathbf{1}-\chi).\end{aligned} \tag{7}$$

Equation (7) is equivalent to the condition that the size of $\{y \in C \mid (x, y) \in E\} = -\theta_{\min}$ for any $x \in X \setminus C$. □

Remark 3.2 Assume that a normally regular digraph Γ satisfies the assumptions of Lemma 2.1. By $A + A^T = J_n - I_r \otimes J_{n/r}$, the valency k of Γ is $\frac{n(r-1)}{2r}$. Thus the right hand side of the bound in Proposition 3.1 is n/r. Then the cocliques represented as the main diagonal blocks in A attain the bound in Proposition 3.1.

4 A Characterization of Skew-Bush-type Hadamard Matrices

A *Hadamard matrix of order n* is an $n \times n$ $(1, -1)$-matrix such that $HH^T = nI_n$. A Hadamard matrix H of order $4n^2$ is of *Bush-type (or checkered)* if $H = (H_{ij})_{i,j=1}^{2n}$, where each H_{ij} is a $2n \times 2n$ matrix, such that $H_{ii} = J_{2n}$ for any $i \in \{1, \ldots, 2n\}$ and $H_{ij}J_{2n} = J_{2n}H_{ij} = 0$ for any distinct $i, j \in \{1, \ldots, 2n\}$. A Bush-type Hadamard matrix $H = (H_{ij})_{i,j=1}^{2n}$ of order $4n^2$ is of *skew-Bush-type (or skew-checkered)* if $H - I_{2n} \otimes J_{2n}$ is skew-symmetric.

It was shown by Haemers and Tonchev in [6] that there exists some symmetric association scheme of class 3 if and only if there exists a strongly regular graph with vertex set being decomposed into disjoint cliques attaining Hoffmann's clique bound. It was shown by Wallis in [13] that there exists a symmetric Bush-type Hadamard matrix of order $4n^2$ if and only if there exists a strongly regular graph with parameters $(4n^2, 2n^2 - n, n^2 - n, n^2 - n)$ such that the vertex set is decomposed into $2n$ disjoint cocliques of size $2n$ (see also [12, Lemma 1.1]).

Digraph's counterpart of the result of Haemers and Tonchev by restricting the parameters to $(4n^2, 2n^2 - n, n^2 - n, n^2 - n)$ was shown in [5], which says there exists some imprimitive nonsymmetric association scheme if and only if there exists a skew-Bush-type Hadamard matrix. In this section, we show digraph counterpart of the result of Wallis [13], namely characterizing the skew-Bush-type Hadamard matrices in terms of the notion of doubly regular asymmetric digraphs with a similar property to the undirected case.

Proposition 4.1 *The following are equivalent.*

(i) There exists a skew-Bush-type Hadamard matrix of order $4n^2$.
(ii) There exists a doubly regular asymmetric digraph with parameters $(4n^2, 2n^2 - n, n^2 - n)$ *such that the vertex set is decomposed into* $2n$ *disjoint cocliques of size* $2n$.

Proof (i)⇒(ii): Let H be a skew-Bush-type Hadamard matrix of order $4n^2$. Define a $(0, 1)$-matrix $A = \frac{1}{2}(J_{4n^2} - H)$. Since $H - I_{2n} \otimes J_{2n}$ is skew-symmetric, A satisfies that $A + A^T = J_{4n^2} - I_{2n} \otimes J_{2n}$. Thus A is the adjacency matrix of a digraph whose vertex set is decomposed into disjoint $2n$ cliques of size $2n$. Since H is a regular Hadamard matrix in particular, it follows that A satisfies the equation $AA^T = n^2 I_{4n^2} + (n^2 - n)J_{4n^2}$. This shows that A is the adjacency matrix of a doubly regular asymmetric digraph with the desired parameters.

(ii)⇒(i): Let Γ be a doubly regular asymmetric digraph with parameters $(4n^2, 2n^2 - n, n^2 - n)$ with the property that the vertex set is decomposed into $2n$ disjoint cocliques of size $2n$. Let A be the adjacency matrix of Γ. Since Γ is decomposed into $2n$ disjoint cocliques of size $2n$, after a suitable rearranging the ordering of the vertices, we may assume that $A + I_{2n} \otimes J_{2n}$ is a $(0, 1)$-matrix. Let $H = A - A^T + I_{2n} \otimes J_{2n}$, and set H_{ij}, A_{ij} $(i, j \in \{1, \ldots, 2n\})$ to be $2n \times 2n$ matrices such that $H = (H_{ij})_{i,j=1}^{2n}$ and $A = (A_{ij})_{i,j=1}^{2n}$. Then H is a $(1, -1)$-matrix, and

the direct calculation shows that H is a Hadamard matrix. It is clear that each diagonal block of size $2n$ is J_n and $H - I_{2n} \otimes J_{2n}$ is skew-symmetric. By Lemma 2.1 the eigenvalues of A are $2n^2 - n, \pm\sqrt{-1}n, -n$. As is shown in Remark 3.2, the disjoint $2n$ cocliques represented as the main diagonal blocks of A attain the upper bound in Proposition 3.1, and thus by Proposition 3.1(ii) we have $A_{ij}J_{2n} = J_{2n}A_{ij} = nJ_{2n}$ for any distinct i, j, namely $H_{ij}J_{2n} = J_{2n}H_{ij} = 0$. Therefore H is a skew-Bush-type Hadamard matrix. □

5 Regular Biangular Matrices and Association Schemes

In [7] the authors constructed association schemes from a Hadamard matrix of order n and mutually orthogonal Latin squares of order $n - 1$. In this section, we construct some association scheme from a Hadamard matrix of order n and a single Latin square with some properties of order $n - 1$. Some relation graphs of the association schemes have the property that its vertex set is decomposed into disjoint cocliques attaining the bound in Proposition 3.1.

An (α, β)-*biangular matrix of order* n is an $n \times n$ $(1, -1)$-matrix H such that the inner products of its normalized rows of H are in $\{\alpha, \beta\}$ [7]. An (α, β)-biangular matrix H of order nm is called *regular* if the rows of H can be partitioned into m-classes of size n each in such a way that:

(i) $|\langle u, v\rangle| = \alpha$ for each distinct pair u, v in the same class.
(ii) $|\langle u, v\rangle| = \beta$ for each pair u, v belonging to different classes.

We will use the following lemma proven in [10].

Lemma 5.1 *If there exists a Hadamard matrix of order n, then there exist symmetric* $(1, -1)$*-matrices* $C_1, C_2, \ldots, C_n$ *of order n such that:*

(i) $C_1 = J_n$.
(ii) $C_iC_j = 0$, $1 \le i \ne j \le n$.
(iii) $C_i^2 = nC_i$, $1 \le i \le n$.
(iv) $\sum_{i=1}^{n} C_i = nI_n$.

It follows from these conditions that the row sums and column sums are 0 *for* C_i, $i \ne 1$, *and that*

$$\sum_{i=2}^{n} C_i^2 = n^2 I_n - nJ_n \ .$$

Proof Letting H be a normalized Hadamard matrix with ith row h_i for $i \in \{1, \ldots, n\}$, set $C_i = h_i^T h_i$. Then $C_1, \ldots, C_n$ satisfy the conditions (i)–(iv). □

Let $H = (H_{ij})_{i,j=1}^{n}$ be a regular (α, β)-biangular matrix of order nm, where each H_{ij} is a square matrix of order m and the rows in i-th block are in the same class for any $i \in \{1, \ldots, m\}$. The regular (α, β)-biangular matrix H is said to be of *skew-symmetric* if $H_{ij}^T = -H_{ji}$ for any distinct $i, j \in \{1, \ldots, n\}$.

Theorem 5.2 *Let n be the order of a Hadamard matrix. Then the following hold.*

(i) There is a symmetric regular $(0, \frac{1}{n-1})$-biangular matrix of order $n(n-1)$.
(ii) There is a skew-symmetric regular $(0, \frac{1}{n-1})$-biangular matrix of order $n(n-1)$.

Proof Let H be a normalized Hadamard matrix, and let L be an addition table of $\mathbb{Z}_{n-1}$. Then L is a symmetric Latin square with (i, j)-entry denoted by $l(i, j)$. We regard L as a Latin square on the set $\{2, \ldots, n\}$.

Starting with a symmetric Latin square on the set $\{2, \ldots, n\}$ and substituting i with C_i from Lemma 5.1 for $i \in \{2, \ldots, n\}$, we obtain a matrix which we will denote by M. Clearly M is a $(1, -1)$-matrix of order $n(n-1)$. It follows from Lemma 5.1 that MM^T is a block matrix with all diagonal blocks equal to $n^2 I_n - n J_n$ by Lemma 5.1, and off-diagonal blocks equal to zero matrix by Lemma 5.1(ii) and the property of L being a Latin square. This completes the proof of (i).

For a construction (ii), define $M = (M_{ij})_{i,j=1}^n$ by $M_{ij} = C_{l(i,j)}$ for $i \le j$ and $M_{ij} = -C_{l(i,j)}$ for $i > j$. Then it is easy to see that the matrix M is the desired skew-symmetric biangular matrix. □

More precisely, the matrices M in Theorem 5.2(i), (ii) satisfy the following equation:

$$MM^T = n(n-1)I_{n(n-1)} - nI_{n-1} \otimes (J_n - I_n). \tag{8}$$

We decompose M into disjoint $(0, 1)$-matrices $A_0, A_1, \ldots, A_4$ defined as follows:

$$\begin{aligned} M &= A_0 + A_1 - A_2 + A_3 - A_4 \\ A_0 &= I_{n(n-1)} \\ A_1 + A_2 &= (J_{n-1} - I_{n-1}) \otimes J_n, \qquad (9) \\ A_0 + A_3 + A_4 &= I_{n-1} \otimes J_n. \qquad (10) \end{aligned}$$

Note that $A_1 = A_1^T$, $A_2 = A_2^T$ if M is a symmetric regular biangular matrix and $A_1 = A_2^T$ if M is a skew-symmetric regular biangular matrix, and A_3, A_4 are symmetric in both cases.

Theorem 5.3 *(i) The set of matrices $\{A_0, A_1, A_2, A_3, A_4\}$ forms a symmetric association scheme if M is a symmetric regular biangular matrix.*
(ii) The set of matrices $\{A_0, A_1, A_2, A_3, A_4\}$ forms a nonsymmetric association scheme if M is a skew-symmetric regular biangular matrix.

Proof In both cases, the proof is the same as follows. Let $\mathcal{A} := \mathrm{span}_{\mathbb{R}}\{A_0, A_1, \ldots, A_4\}$. Since each block matrix of A_i for any i has a constant row and column sum, we have

$$A_i(I_{n-1} \otimes J_n) = (I_{n-1} \otimes J_n)A_i \in \mathcal{A}, \tag{11}$$
$$A_i((J_{n-1} - I_{n-1}) \otimes J_n) = ((J_{n-1} - I_{n-1}) \otimes J_n)A_i \in \mathcal{A}. \tag{12}$$

First we show $A_iA_j \in \mathcal{A}$ for $i, j \in \{3, 4\}$. By Lemma 5.1(iii), we have $(A_0 + A_3 - A_4)^2 = n(A_0 + A_3 - A_4)$. By (10) and (11) we have $A_4^2 \in \mathcal{A}$. Similarly $A_iA_j \in \mathcal{A}$ for others $i, j \in \{3, 4\}$.

Next we show $A_iA_j \in \mathcal{A}$ for $i \in \{1, 2\}$, $j \in \{3, 4\}$ or $i \in \{3, 4\}$, $j \in \{1, 2\}$. By Lemma 5.1(ii), we have $(A_1 - A_2)(A_0 + A_3 - A_4) = 0$. By (9), (10), (11) and (12), we have $A_2A_4 \in \mathcal{A}$. Similarly $A_iA_j \in \mathcal{A}$ for others $i \in \{1, 2\}$, $j \in \{3, 4\}$ or $i \in \{3, 4\}$, $j \in \{1, 2\}$.

Finally we show $A_iA_j \in \mathcal{A}$ for $i, j \in \{1, 2\}$. By (8) we have $(A_1 - A_2)^2 \in \mathcal{A}$. By (9) and (12), we have $A_iA_j \in \mathcal{A}$ for $i, j \in \{1, 2\}$. Thus this completes the proof. □

The first eigenmatrices in Theorem 5.3(i), (ii) are as follows respectively:

$$P = \begin{pmatrix} 1 & \frac{n(n-2)}{2} & \frac{n(n-2)}{2} & \frac{n-2}{2} & \frac{n}{2} \\ 1 & 0 & 0 & \frac{n-2}{2} & -\frac{n}{2} \\ 1 & -\frac{n}{2} & -\frac{n}{2} & \frac{n-2}{2} & \frac{n}{2} \\ 1 & -\frac{n}{2} & \frac{n}{2} & -1 & 0 \\ 1 & \frac{n}{2} & -\frac{n}{2} & -1 & 0 \end{pmatrix}, P = \begin{pmatrix} 1 & \frac{n(n-2)}{2} & \frac{n(n-2)}{2} & \frac{n-2}{2} & \frac{n}{2} \\ 1 & 0 & 0 & \frac{n-2}{2} & -\frac{n}{2} \\ 1 & -\frac{n}{2} & -\frac{n}{2} & \frac{n-2}{2} & \frac{n}{2} \\ 1 & -\frac{\sqrt{-1}n}{2} & \frac{\sqrt{-1}n}{2} & -1 & 0 \\ 1 & \frac{\sqrt{-1}n}{2} & -\frac{\sqrt{-1}n}{2} & -1 & 0 \end{pmatrix}.$$

See Appendices A, B for the intersection numbers and second eigenmatrices. Consider relation graphs with adjacency matrix A_1, A_2 in both association schemes. As Proposition 3.1 shows, each main diagonal block of A_1, A_2 represents a coclique and A_4 corresponds to a partition of the vertex set by cliques attaining the bound in Proposition 3.1.

6 Twin Asymmetric Designs and Association Schemes

Finally we focus on normally regular digraphs with $\lambda = \mu$, or equivalently doubly regular asymmetric graphs. If an incidence matrix N of a symmetric design is such that $N + N^T$ is a $(0, 1)$-matrix, then N is an adjacency matrix of a doubly regular asymmetric digraph, and vice versa. Our main reference for this section is [8]. We will refer to a doubly regular asymmetric digraph with parameters (v, k, λ) as a *DRAD*(v, k, λ). Symmetric (v, k, λ)-designs $\mathbf{D} = (X, \mathcal{B})$ and $\mathbf{D}' = (X, \mathcal{B}')$ are called *twin designs* if there is a bijection $f : \mathcal{B} \to \mathcal{B}'$ such that every block $B \in \mathcal{B}$ is disjoint from $f(B)$. In general, it is not easy to find twin symmetric designs. However, if Γ is a *DRAD*(v, k, λ) and Γ' is the digraph obtained by reversing the direction of every arc of Γ, then the corresponding symmetric designs are twins. The following theorem is proven in [8].

Theorem 6.1 *Let h be a positive integer such that there exists a Hadamard matrix of order $2h$. If $p = (2h - 1)^2$ is a prime power, then, for any positive integer d, there exists a*

$$DRAD\left(\frac{h(p^{2d} - 1)}{h + 1}, hp^{2d}, h(h + 1)p^{2d-1}\right). \tag{13}$$

The construction makes use of *skew balanced generalized weighing matrices* and Bush-type Hadamard matrices constructed as in Lemma 5.1 from a Hadamard matrix of order $2h$. We illustrate this by an example which relates to the special case of the theorem which used in this note.

Example 6.2 We start with a BGW(10, 9, 8) over the cyclic group C_8. Let

$$W = [w_{ij}] = \begin{pmatrix} 0&1&1&1&1&1&1&1&1&1\\ 4&0&3&7&5&6&8&1&4&2\\ 4&7&0&3&8&5&6&2&1&4\\ 4&3&7&0&6&8&5&4&2&1\\ 4&1&4&2&0&3&7&5&6&8\\ 4&2&1&4&7&0&3&8&5&6\\ 4&4&2&1&3&7&0&6&8&5\\ 4&5&6&8&1&4&2&0&3&7\\ 4&8&5&6&2&1&4&7&0&3\\ 4&6&8&5&4&2&1&3&7&0 \end{pmatrix}.$$

Then W is a skew BGW(10, 9, 8) over the cyclic group $C_8 = \langle g \rangle$ generated by the matrix

$$g = \begin{pmatrix} 0 & I_4 & 0 & 0\\ 0 & 0 & I_4 & 0\\ 0 & 0 & 0 & I_4\\ -I_4 & 0 & 0 & 0 \end{pmatrix},$$

where the number i in G denotes g^i for $i = 1, 2, \ldots, 8$. Let

$$H = \begin{pmatrix} 0 & C_2 & C_3 & C_4\\ -C_4 & 0 & C_2 & C_3\\ -C_3 & -C_4 & 0 & C_2\\ -C_2 & -C_3 & -C_4 & 0 \end{pmatrix},$$

where C_2, C_3, C_4 are those constructed in Lemma 5.1 from a normalized Hadamard matrix of order 4 and 0 denotes the zero matrix of order 16.

Let

$$R = \begin{pmatrix} 0 & 0 & 0 & I_4\\ 0 & 0 & I_4 & 0\\ 0 & I_4 & 0 & 0\\ I_4 & 0 & 0 & 0 \end{pmatrix}.$$

Let $G = [H w_{ij} R]$, then G can be splitted to parts, namely the positive and negative part, to form a twin skew symmetric (160, 54, 18) design on 160 vertices.

To do this, keep all the 1-entries in G, change all the -1-entries to 0 and let A_1 be the (0, 1)-matrix obtained. Then, A_1 is the incidence matrix of a symmetric (160, 54, 18) design. Furthermore, $A_1 + A_1^T$ is a (0, 1)-matrix. So, A_1 is the adja-

cency matrix of a doubly regular asymmetric digraph. Now change all the 1-entries in G to 0, all -1-entries to 1 and let A_2 be the $(0, 1)$-matrix obtained. Then $A_2 = A_1^T$, so A_1 and A_2 are twins. We refer the reader to [8] for the general construction.

We now use the sequence of doubly regular digraphs obtained from the above theorem for $d = 1$ to deduce the existence of some association schemes of class five. The general case corresponding to any positive integer d will appear elsewhere.

Theorem 6.3 *Let $h = 2n$ be a positive integer for which there is a Hadamard matrix of order h and $p = 2n - 1$ is a prime power. Consider the skew $BGW(p^2 + 1, p^2, p^2 - 1)$ over the cyclic group of order $4n$ and the twin design constructed in [8] for $d = 1$.*
Let A_1 be the plus and A_2 the minus twin, $A_4 = I_{2n(p^2+1)} \otimes (J_{2n} - I_{2n})$, $A_5 = I_{p^2+1} \otimes (J_{4n^2} - I_{2n} \otimes J_{2n})$.
Then $\{A_0 = I_{4n^2(p^2+1)}, A_1, A_2, A_3 = J_{4n^2(p^2+1)} - A_1 - A_2 - A_4 - A_5, A_4, A_5\}$ forms a nonsymmetric association scheme of class 5 with the following intersection numbers. Note that $A_1^T = A_2$, A_3, A_4, A_5 are symmetric.

- $A_1A_1 = A_2A_2 = (n - 1)(2n - 1)(2n^2 - n)(A_1 + A_2 + A_3 + A_5) + n^2(2n-1)^2A_4$.
- $A_1A_2 = n^2(2n - 1)^2A_0 + (2n - 1)^2(n^2 - n)J$.
- $A_1A_3 = A_2A_3 = 2n(n - 1)(2n - 1)(A_1 + A_2 + A_3) + n(2n - 1)^2A_5$.
- $A_1A_4 = (n - 1)A_1 + nA_2$.
- $A_1A_5 = A_2A_5 = 2n(n - 1)(A_1 + A_2) + n(2n - 1)A_3$.
- $A_2A_4 = nA_1 + (n - 1)A_2$.
- $A_3A_3 = 2n(2n - 1)^2A_0 + 4n(n - 1)(A_1 + A_2 + A_3) + 2n(2n - 1)^2A_4$.
- $A_3A_4 = (2n - 1)A_3$.
- $A_3A_5 = 2n(A_1 + A_2)$.
- $A_4A_4 = (2n - 1)A_0 + (2n - 2)A_4$.
- $A_4A_5 = (2n - 1)A_5$.
- $A_5A_5 = 2n(2n - 1)A_0 + 2n(2n - 1)A_4 + 4n(n - 1)A_5$.

Proof Let $W = [w_{ij}]$ be a skew $BGW(p^2 + 1, p^2, p^2 - 1)$ over a cyclic group of order $4n$ generated by a negacirculant matrix of order $4n$ as described in [8]. Let $R = R_{2n} \otimes I_{2n}$, where R_{2n} denotes the back identity matrix of order $2n$ and I_{2n} is the identity matrix of order $2n$. Then $A_3 = [|w_{ij}|R]$.
The identities for $A_1A_2 = A_2A_1$ follows from the fact that each of A_1 and A_2 are the inicdence matrices of a symmetric $((p^2 + 1)4n^2, p^2(2n^2 - n), p^2(n^2 - n))$ designs and $A_1^T = A_2$. The numbers for A_1A_1 and A_2A_2 follows from the fact that the symmetric matrix $A_1 + A_2$ has a simple structure and we make use of it in finding the numbers for other products involving A_1 and A_2. The relation related to A_3 follows from the observation that $A_3 = [|w_{ij}|R]$ and $A_1 + A_2 + A_3 = J_{4n^2(p^2+1)} - I_{p^2+1} \otimes J_{4n^2}$. The remaining numbers are not hard to calculate. □

The eigenmatrices P, Q are given as follows:

$$P = \begin{pmatrix} 1 & n(2n-1)^3 & n(2n-1)^3 & 2n(2n-1)^2 & 2n-1 & 2n(2n-1) \\ 1 & n(2n-1) & n(2n-1) & -2n(2n-1) & 2n-1 & -2n \\ 1 & n(2n-1)\sqrt{-1} & -n(2n-1)\sqrt{-1} & 0 & -1 & 0 \\ 1 & -n(2n-1)\sqrt{-1} & n(2n-1)\sqrt{-1} & 0 & -1 & 0 \\ 1 & -n(2n-1) & -n(2n-1) & -2n & 2n-1 & 2n(2n-1) \\ 1 & -n(2n-1) & -n(2n-1) & -2n(2n-1) & 2n-1 & -2n \end{pmatrix},$$

$$Q = \begin{pmatrix} 1 & (2n-1)m & 2n(2n-1)m & 2n(2n-1)m & (2n-1)^2 & (2n-1)m \\ 1 & \frac{m}{2n-1} & -\frac{2nm\sqrt{-1}}{2n-1} & \frac{2nm\sqrt{-1}}{2n-1} & -1 & -\frac{m}{2n-1} \\ 1 & \frac{m}{2n-1} & \frac{2nm\sqrt{-1}}{2n-1} & -\frac{2nm\sqrt{-1}}{2n-1} & -1 & -\frac{m}{2n-1} \\ 1 & -m & 0 & 0 & -1 & -2n+1 \\ 1 & m & -2nm & -2nm & (2n-1)^2 & m \\ 1 & -m & 0 & 0 & (2n-1)^2 & -m \end{pmatrix},$$

where $m = 2n^2 - 2n + 1$. By the definition of A_4, A_5, we have $A_4 + A_5 = I_{p^2+1} \otimes (J_{4n^2} - I_{4n^2})$. The cocliques of the digraphs whose adjacency matrices are A_1, A_2 corresponding to the main diagonal blocks of $A_4 + A_5$ attain the upper bound in Proposition 3.1.

Acknowledgements Hadi Kharaghani is supported by an NSERC Discovery Grant. Sho Suda is supported by JSPS KAKENHI Grant Number 15K21075.

Appendix 1: Parameters of the association Scheme in Theorem 5.3(i)

$$B_1 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ \frac{n^2-2n}{2} & \frac{n^2-3n}{4} & \frac{n^2-3n}{4} & \frac{n^2-4n}{4} & \frac{n^2-2n}{4} \\ 0 & \frac{n^2-3n}{4} & \frac{n^2-3n}{4} & \frac{n^2}{4} & \frac{n^2-2n}{4} \\ 0 & \frac{n}{4}-1 & \frac{n}{4} & 0 & 0 \\ 0 & \frac{n}{4} & \frac{n}{4} & 0 & 0 \end{pmatrix}$$

$$B_2 = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & \frac{n^2-3n}{4} & \frac{n^2-3n}{4} & \frac{n^2}{4} & \frac{n^2-2n}{4} \\ \frac{n^2-2n}{2} & \frac{n^2-3n}{4} & \frac{n^2-3n}{4} & \frac{n^2-4n}{4} & \frac{n^2-2n}{4} \\ 0 & \frac{n}{4} & \frac{n-4}{4} & 0 & 0 \\ 0 & \frac{n}{4} & \frac{n}{4} & 0 & 0 \end{pmatrix}$$

$$B_3 = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & \frac{n}{4}-1 & \frac{n}{4} & 0 & 0 \\ 0 & \frac{n}{4} & \frac{n}{4}-1 & 0 & 0 \\ \frac{n}{2}-1 & 0 & 0 & \frac{n}{2}-2 & 0 \\ 0 & 0 & 0 & 0 & \frac{n}{2}-1 \end{pmatrix}$$

$$B_4 = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & \frac{n}{4} & \frac{n}{4} & 0 & 0 \\ 0 & \frac{n}{4} & \frac{n}{4} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{n}{2}-1 \\ \frac{n}{2} & 0 & 0 & \frac{n}{2} & 0 \end{pmatrix}$$

$$Q = \begin{pmatrix} 1 & n-1 & n-2 & \frac{(n-1)(n-2)}{2} & \frac{(n-1)(n-2)}{2} \\ 1 & 0 & -1 & -\frac{n-1}{2} & \frac{n-1}{2} \\ 1 & 0 & -1 & \frac{n-1}{2} & -\frac{n-1}{2} \\ 1 & n-1 & n-2 & -n+1 & -n+1 \\ 1 & -n+1 & n-2 & 0 & 0 \end{pmatrix}$$

Appendix 2: Parameters of the Association Scheme in Theorem 5.3(ii)

$$B_1 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & \frac{n^2-3n}{4} & \frac{n^2-3n}{4} & \frac{n^2}{4} & \frac{n^2-2n}{4} \\ \frac{n^2-2n}{2} & \frac{n^2-3n}{4} & \frac{n^2-3n}{4} & \frac{n^2-4n}{4} & \frac{n^2-2n}{4} \\ 0 & \frac{n}{4}-1 & \frac{n}{4} & 0 & 0 \\ 0 & \frac{n}{4} & \frac{n}{4} & 0 & 0 \end{pmatrix}$$

$$B_2 = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ \frac{n^2-2n}{2} & \frac{n^2-3n}{4} & \frac{n^2-3n}{4} & \frac{n^2-4n}{4} & \frac{n^2-2n}{4} \\ 0 & \frac{n^2-3n}{4} & \frac{n^2-3n}{4} & \frac{n^2}{4} & \frac{n^2-2n}{4} \\ 0 & \frac{n}{4} & \frac{n-4}{4} & 0 & 0 \\ 0 & \frac{n}{4} & \frac{n}{4} & 0 & 0 \end{pmatrix}$$

$$B_3 = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & \frac{n}{4}-1 & \frac{n}{4} & 0 & 0 \\ 0 & \frac{n}{4} & \frac{n}{4}-1 & 0 & 0 \\ \frac{n}{2}-1 & 0 & 0 & \frac{n}{2}-2 & 0 \\ 0 & 0 & 0 & 0 & \frac{n}{2}-1 \end{pmatrix}$$

$$B_4 = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & \frac{n}{4} & \frac{n}{4} & 0 & 0 \\ 0 & \frac{n}{4} & \frac{n}{4} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{n}{2}-1 \\ \frac{n}{2} & 0 & 0 & \frac{n}{2} & 0 \end{pmatrix}$$

$$Q = \begin{pmatrix} 1 & n-1 & n-2 & \frac{(n-1)(n-2)}{2} & \frac{(n-1)(n-2)}{2} \\ 1 & 0 & -1 & -\frac{\sqrt{-1}(n-1)}{2} & \frac{\sqrt{-1}(n-1)}{2} \\ 1 & 0 & -1 & \frac{\sqrt{-1}(n-1)}{2} & -\frac{\sqrt{-1}(n-1)}{2} \\ 1 & n-1 & n-2 & -n+1 & -n+1 \\ 1 & -n+1 & n-2 & 0 & 0 \end{pmatrix}$$

References

1. Brouwer, A.E.: Distance regular graphs of diameter 3 and strongly regular graphs. Discrete Math. **49**, 101–103 (1984)
2. Brouwer, A.E., Haemers, W.H.: Spectra of graphs. Universitext. Springer, New York (2012). xiv+250 pp
3. Chang, Y.: Imprimitive Symmetric Association Schemes of Rank 4. University of Michigan, Thesis (1994)
4. Cvetković, D.M., Doob, M., Sachs, H.: Spectra of graphs. Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York (1980)
5. Goldbach, R.W., Claasen, H.L.: 3-class association schemes and Hadamard matrices of a certain block form. Europ. J. Combin. **19**, 943–951 (1998)
6. Haemers, W.H., Tonchev, V.D.: Spreads in strongly regular graphs. Des. Codes Crypt. **8**, 145–157 (1996)
7. Holzmann, W.H., Kharaghani, H., Suda, S.: Mutually unbiased biangular vectors and association schemes. In Colbourn, C. J. (ed.) Algebraic Design Theory and Hadamard Matrices, vol. 133, pp. 149–157. Springer International Publishing (2015)
8. Ionin, Y.J., Kharaghani, H.: Doubly regular digraphs and symmetric designs. J. Combin. Theory Ser. A **101**(1), 35–48 (2003)
9. Jorgensen, L.K., Jones, G.A., Klin, M.H., Song, S.Y.: Normally regular digraphs, association schemes and related combinatorial structures. Sém. Lothar. Combin. 71, Art. B71c, 39pp (2013/14)
10. Kharaghani, H.: New class of weighing matrices. Ars. Combin. **19**, 69–72 (1985)
11. Kharaghani, H., Sasani,S., Suda, S.: Mutually unbiased Bush-type Hadamard matrices and association schemes. Elec. J. Combin. **22** P3. 10 (2015)
12. Muzychuk, M., Xiang, Q.: Symmetric Bush-type Hadamard matrices of order $4m^4$ exist for all odd m. Proc. Amer. Math. Soc. **134**(8), 2197–2204 (2006)
13. Wallis, W.D.: On a problem of K. A. Bush concerning Hadamard matrices. Bull. Aust. Math. Soc. **6**, 321–326 (1971)

A Suspension Bridge Problem: Existence and Stability

Salim A. Messaoudi and Soh Edwin Mukiawa

Abstract In this work, we consider a semilinear problem describing the motion of a suspension bridge in the downward direction in the presence of its hanger restoring force $h(u)$ and a linear damping δu_t, where $\delta > 0$ is a constant. By using the semigroup theory, we establish the well posedness. We also use the multiplier method to prove a stability result.

Keywords Suspension bridge · Semigroup theory · Well posedness · Stability · Exponential decay

Mathematics Subject Classification: 35L51 · 35L71 · 35B35 · 35B41

1 Introduction

A simple model for a bending energy of a deformed thin plate $\Omega = (0, L) \times (-\ell, \ell)$ is given by

$$E_B(u) = \int_\Omega \left(\frac{K_1^2}{2} + \frac{K_2^2}{2} + \sigma K_1 K_2 \right) dxdy, \tag{1.1}$$

where $u = u(x, y)$ represents the downward vertical displacement of the plate and K_1, K_2 are the principal curvatures of the graph of u. The constant $\sigma = \frac{\lambda}{2\lambda+\mu}$ is the Poission ratio and λ, μ are called the Lamé moduli. For some physical reasons, $\lambda \geq 0$ and $\mu > 0$, hence $0 < \sigma < \frac{1}{2}$. For small deformation u, the following approximations hold

$$(K_1 + K_2)^2 \approx (\Delta u)^2, \quad K_1 K_2 \approx det(D^2 u) = u_{xx} u_{yy} - u_{xy}^2.$$

S.A. Messaoudi (✉) · S.E. Mukiawa
Department of Mathematics and Statistics, King Fahd University of Petroleum and Minerals, P.O.Box 546, Dhahran 31261, Saudi Arabia
e-mail: messaoud@kfupm.edu.sa

S.E. Mukiawa
e-mail: sohedwin2013@gmail.com

T. Abualrub et al. (eds.), *Mathematics Across Contemporary Sciences*, Springer Proceedings in Mathematics & Statistics 190,
DOI 10.1007/978-3-319-46310-0_9

As a result, we get

$$\frac{1}{2}K_1^2 + \frac{1}{2}K_2^2 + \sigma K_1 K_2 \approx \frac{1}{2}(\Delta u)^2 + (\sigma - 1)det(D^2 u).$$

Consequently, the energy functional (1.1) takes the form

$$E_B(u) = \int_\Omega \left(\frac{1}{2}(\Delta u)^2 + (\sigma - 1)det(D^2 u)\right) dxdy. \tag{1.2}$$

We note here that, for $0 < \sigma < \frac{1}{2}$, E_B is convex and is also coercive in suitable state spaces such as $H_0^2(\Omega)$ or $H^2(\Omega) \cap H_0^1(\Omega)$.
If f is an external vertical load acting on the plate Ω, then the total energy is given by

$$\begin{aligned} E_T(u) &= E_B(u) - \int_\Omega f u dxdy \\ &= \int_\Omega \left[\left(\frac{1}{2}(\Delta u)^2 + (\sigma - 1)(u_{xx}u_{yy} - u_{xy}^2)\right) - fu\right] dxdy. \end{aligned} \tag{1.3}$$

The unique minimizer u of the functional (1.3) satisfies the Euler-Lagrange equation

$$\Delta^2 u(x, y) = f(x, y), \ in \ \Omega. \tag{1.4}$$

For totally supported plate ($u = \frac{\partial u}{\partial \eta} = 0$), the problem has been first solved by Navier [17] in 1823. Since the bridge is usually simply supported on the vertical sides ($x = 0, x = L$, i.e. the $y-$ axis) only

$$u(0, y) = u_{xx}(0, y) = u(L, y) = u_{xx}(L, y) = 0,$$

then different boundary conditions should be considered for the horizontal sides ($y = -\ell, y = \ell$, i.e. $x-$axis). Various problems on a rectangular plate Ω, where only the vertical sides are simply supported, were discussed by many authors, see, for instance Mansfield [11]. Naturally, one should consider the plate Ω with free horizontal sides. In such a situation, the boundary conditions are

$$\begin{cases} u_{yy}(x, \pm\ell) + \sigma u_{xx}(x, \pm\ell) = 0, \ \ for \ x, \in \ (0, L), \\ \\ u_{yyy}(x, \pm\ell) + (2 - \sigma)u_{xxy}(x, \pm\ell) = 0, \ \ for \ x \in \ (0, L), \end{cases} \tag{1.5}$$

see Ventsel and Krauthammer [19]. Putting all pieces together (see Ferrero and Gazzola [5]), the boundary value problem for a thin plate Ω modeling a suspension bridge is

$$\begin{cases} \Delta^2 u(x,y) = f(x,y), \quad in \ \ \Omega, \\ u(0,y) = u_{xx}(0,y) = u(L,y) = u_{xx}(L,y) = 0, \quad y \in (-\ell,\ell), \\ u_{yy}(x,\pm\ell) + \sigma u_{xx}(x,\pm\ell) = 0, \quad x \in (0,L), \\ u_{yyy}(x,\pm\ell) + (2-\sigma) u_{xxy}(x,\pm\ell) = 0, \quad x \in (0,L). \end{cases} \tag{1.6}$$

In order to describe the action of the hangers (cables), Ferrero and Gazzola [5] introduced a nonlinear function $h(x,y,u)$ which admits a potential energy given by $\int_\Omega H(x,y,u)dxdy$. As a result, the total energy (1.3) becomes

$$E_T(u) = \int_\Omega \left[\left(\frac{1}{2}(\Delta u)^2 + (\sigma-1)(u_{xx}u_{yy} - u_{xy}^2) \right) + H(x,y,u) - fu \right] dxdy, \tag{1.7}$$

whose unique minimizer satisfies the stationary problem

$$\begin{cases} \Delta^2 u(x,y) + h(x,y,u(x,y)) = f(x,y), \quad in \ \ \Omega, \\ u(0,y) = u_{xx}(0,y) = u(L,y) = u_{xx}(L,y) = 0, \quad y \in (-\ell,\ell), \\ u_{yy}(x,\pm\ell) + \sigma u_{xx}(x,\pm\ell) = 0, \quad x \in (0,L), \\ u_{yyy}(x,\pm\ell) + (2-\sigma) u_{xxy}(x,\pm\ell) = 0, \quad x \in (0,L). \end{cases} \tag{1.8}$$

If the external force f depends on time, $f = f(x,y,t)$, then the kinetic energy $\frac{1}{2}\int_\Omega u_t^2 dxdy$ has to be added to the static total energy (1.7). Thus, the total energy becomes

$$\begin{aligned} E_T(u) = & \frac{1}{2}\int_\Omega u_t^2 dxdy + \int_\Omega \left[\left(\frac{1}{2}(\Delta u)^2 + (\sigma-1)(u_{xx}u_{yy} - u_{xy}^2) \right) \right. \\ & \left. + \ H(x,y,u) - fu \right] dxdy. \end{aligned} \tag{1.9}$$

Also, the equation of motion becomes

$$u_{tt}(x,y,t) + \Delta^2 u(x,y,t) + h(x,y,u(x.y,t)) = f(x,y,t). \tag{1.10}$$

Finally, we might add a damping term due to some internal friction or viscosity. In this case, Eq. (1.10) takes the form

$$u_{tt}(x,y,t) + \delta u_t(x,y,t) + \Delta^2 u(x,y,t) + h(x,y,u(x.y,t)) = f(x,y,t), \tag{1.11}$$

where $\delta > 0$ is called the friction constant. Equation (1.11) together with the boundary conditions of (1.8) and initial data has been discussed by Ferrero and Gazzola [5], for a general nonlinear restoring force h. They proved the existence of a unique solution, using the Galerkin method. In addition, they discussed several stationary problems. Recent results by Wang [20] and Al-Gwaiz et al. [2] have also made use of the above mention boundary conditions.

Early results concerning suspension bridges go back to McKenna and collaborators. For instance, Glover et al. [8] considered the damped couple system

$$\begin{cases} u_{tt} + u_t + u_{xxxx} + \gamma_1 u_t + k(u - v)^+ = f, \\ \epsilon v_{tt} - v_{xx} + \gamma_2 v_t - k(u - v)^+ = g, \end{cases} \tag{1.12}$$

where,

$$u, v : [0, L] \times \mathbb{R}^+ \longrightarrow \mathbb{R}$$

represent the downward deflection and the vertical displacement of the string. For rigid suspension bridges, Lazer and Mckenna [12] reduced the system (1.12) to the following fourth-order equation

$$u_{tt} + u_{xxxx} + u_t + k^2 u^+ = f, x \in (0, 1), \quad t > 0, \tag{1.13}$$

and established existence of periodic solutions by assuming the suspension bridge as a bending beam. Equation (1.13) has been studied by a few authors (see [1, 4]). Mckenna and Walter, [14, 15] also investigated the nonlinear oscillations of suspension bridges and the existence of travelling wave solutions have been established. To achieve this, they considered the suspension bridge as a vibrating beam. Bochicchio et al. [3] considered

$$u_{tt} + u_t + u_{xxxx} + (p - \|u_x\|^2_{L^2((0,1))})u_{xx} + ku^2 = f, \tag{1.14}$$

where p is a force that acts directly on the central axis of the bridge (axial force) and f a general external source term. They established a well-posedness as well as existence of global attractor. For more literature concerning the suspension bridges, we refer the reader to Mckenna [13], Mckenna et al. [16], Filippo et al. [7], Imhof [9], and Gazzola [6].

In this work, we consider the following fourth order semilinear plate problem

$$\begin{cases} u_{tt}(x, y, t) + \delta u_t(x, y, t) + \Delta^2 u(x, y, t) + h(u(x.y, t)) = 0, \quad in \ \ \Omega \times (0, +\infty), \\ u(0, y, t) = u_{xx}(0, y, t) = u(\pi, y, t) = u_{xx}(\pi, y, t) = 0, \quad (y, t) \in (-\ell, \ell) \times (0, +\infty), \\ u_{yy}(x, \pm\ell, t) + \sigma u_{xx}(x, \pm\ell, t) = 0, \quad (x, t) \in (0, \pi) \times (0, +\infty), \\ u_{yyy}(x, \pm\ell, t) + (2 - \sigma) u_{xxy}(x, \pm\ell, t) = 0, \quad (x, t) \in (0, \pi) \times (0, +\infty), \\ u(x, y, 0) = u_0(x, y), \ u_t(x, y, 0) = u_1(x, y), \quad in \quad \Omega. \end{cases} \tag{1.15}$$

The aim of this work is to reformulate (1.15) into a semigroup setting and then make use of the semigroup theory (see Pazy [18]) to establish the well-posedness. We also use the multiplier method (see Komornik [10]) to prove a stability result for problem (1.15). The rest of this work is organized as follows. In Sect. 2, we present some basic and fundamental materials needed to establish our main results. In Sect. 3, we establish a well-posedness result for problem (1.15). In Sect. 4, we state and prove our stability result.

2 Preliminaries

In this section we present some basic and fundamental results which will be used in proving our main results. For this, we impose the following assumptions on the function h

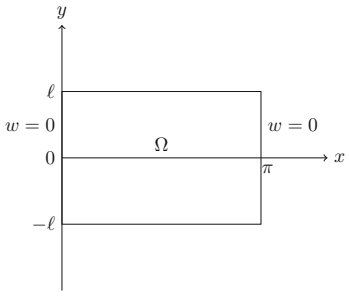
$$\begin{cases} h : \mathbb{R} \longrightarrow \mathbb{R} \quad is \quad lipschitz \quad such \quad that \quad h(0) = 0, \\ H(s) = \displaystyle\int_0^s h(\tau)d\tau \quad is \quad positive, \\ sh(s) - H(s) \geq 0, \quad \forall s \in \mathbb{R}. \end{cases} \tag{2.1}$$

Example 2.1 An example of a function satisfying (2.1) is

$$h(s) = a|s|^{p-1}s, \ a \geq 0, \ p \geq 1.$$

As in [5], we introduce the space

$$H^2_*(\Omega) = \left\{ w \in H^2(\Omega) : w = 0 \text{ on } \{0, \pi\} \times (-\ell, \ell) \right\}, \tag{2.2}$$



together with the inner product

$$(u, v)_{H_*^2} = \int_\Omega [(\Delta u \Delta v + (1-\sigma)(2u_{xy}v_{xy} - u_{xx}v_{yy} - u_{yy}v_{xx})]dxdy. \tag{2.3}$$

For the completeness of $H_*^2(\Omega)$, we have the following results by Ferrero and Gazzola [5].

Lemma 2.1 *[5] Assume $0 < \sigma < \frac{1}{2}$. Then, the norm $\|.\|_{H_*^2(\Omega)}$ given by $\|u\|^2_{H_*^2(\Omega)} = (u, u)_{H_*^2}$ is equivalent to the usual $H^2(\Omega)$-norm. Moreover, $H_*^2(\Omega)$ is a Hilbert space when endowed with the scalar product $(.,.)_{H_*^2}$.* □

Lemma 2.2 *[5] Assume $0 < \sigma < \frac{1}{2}$ and $f \in L^2(\Omega)$. Then there exists a unique $u \in H_*^2(\Omega)$ such that*

$$\int_\Omega [\Delta u \Delta v + (1-\sigma)(2u_{xy}v_{xy} - u_{xx}v_{yy} - u_{yy}v_{xx})]dxdy = \int_\Omega fv, \quad \forall v \in H_*^2(\Omega). \tag{2.4}$$

□

Remark 2.1 The function $u \in H_*^2(\Omega)$ satisfying (2.4) is called the weak solution of the stationary problem (1.6).

Lemma 2.3 *[5] The weak solution $u \in H_*^2(\Omega)$, of (2.4), is in $H^4(\Omega)$ and there exists a $C = C(l, \sigma) > 0$ such that*

$$\|u\|_{H^4(\Omega)} \leq C\|f\|_{L^2(\Omega)}. \tag{2.5}$$

In addition if $u \in C^4(\bar{\Omega})$, then u is called a classical solution of (1.6). □

Lemma 2.4 *[20] Let $u \in H_*^2(\Omega)$ and suppose $1 \leq p < +\infty$. Then, there exists a positive constant $C_e = C_e(\Omega, p)$ such that*

$$\|u\|_p^p \leq C_e\|u\|^p_{H_*^2(\Omega)}.$$

□

Lemma 2.5 *[10] Let $E : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$ be a non-increasing function. Assume that there exists $C > 0$ such that*

$$\int_s^\infty E(t)dt \leq CE(s), \ 0 < s < \infty.$$

Then, there exists $\lambda > 0$ a constant such that

$$E(t) \leq E(0)e^{-\lambda t}, \ \forall t \geq 0. \tag{2.6}$$

3 Well-Posedness

In this section we establish the well-posedness of problem (1.15) using the semigroup theory. For this, we set $u_t = v$, then problem (1.15) becomes

$$(P)\begin{cases} U_t + AU = F \\ \\ U(0) = U_0, \end{cases}$$

where

$$U = \begin{pmatrix} u \\ v \end{pmatrix}, \quad AU = \begin{pmatrix} -v \\ \Delta^2 u + \delta v \end{pmatrix}, \quad F(U) = \begin{pmatrix} 0 \\ -h(u) \end{pmatrix}, \quad U_0 = \begin{pmatrix} u_0 \\ u_1 \end{pmatrix}.$$

We define the Hilbert space

$$\mathcal{H} = H^2_*(\Omega) \times L^2(\Omega)$$

equipped with the inner product

$$(U, V)_{\mathcal{H}} = (u, \tilde{u})_{H^2_*(\Omega)} + (v, \tilde{v})_{L^2(\Omega)}, \tag{3.1}$$

where

$$U = (u, v)^T, \ V = (\tilde{u}, \tilde{v})^T \in \mathcal{H}.$$

Next, we introduce the following notation

$$\begin{cases} u_{xx}(0, y) = u_{xx}(\pi, y) = 0 \\ u_{yy}(x, \pm\ell) + \sigma u_{xx}(x, \pm\ell) = 0 \\ u_{yyy}(x, \pm\ell) + (2 - \sigma) u_{xxy}(x, \pm\ell) = 0. \end{cases} \tag{3.2}$$

The domain of the operator A is defined as

$$D(A) = \left\{(u, v) \in \mathcal{H} / u \in H^4(\Omega) \ satisfying \ (3.2), \ v \in H^2_*(\Omega)\right\}.$$

Lemma 3.1 *We have*

$$(\Delta^2 u, v)_{L^2(\Omega)} = (u, v)_{H^2_*}, \qquad \forall u, v \in D(A). \tag{3.3}$$

Proof Using Green's formula we obtain that

$$\int_\Omega v \Delta^2 u = \int_\Omega \Delta u \Delta v + \int_{\partial\Omega} [v \frac{\partial \Delta u}{\partial \eta} - \Delta u \frac{\partial v}{\partial \eta}]. \tag{3.4}$$

Integration in (3.4) leads to

$$\begin{aligned}\int_\Omega v\Delta^2 u &= \int_\Omega \Delta u\Delta v - \int_0^\pi v(x,-\ell)[u_{xxy}(x,-\ell)+u_{yyy}(x,-\ell)]dx\\ &+\int_0^\pi v(x,\ell)[u_{xxy}(x,\ell)+u_{yyy}(x,\ell)]dx\\ &+\int_0^\pi v_y(x,-\ell)[u_{xx}(x,-\ell)+u_{yy}(x,-\ell)]dx\\ &-\int_{-\ell}^{\ell} v_x(\pi,y)[\cancelto{0}{u_{xx}(\pi,y)}+\cancelto{0}{u_{yy}(\pi,y)}]dy\\ &-\int_0^\pi v_y(x,\ell)[u_{xx}(x,\ell)+u_{yy}(x,\ell)]dx\\ &+\int_{-\ell}^{\ell} v_x(0,y)[\cancelto{0}{u_{xx}(0,y)}+\cancelto{0}{u_{yy}(0,y)}]dy].\end{aligned}$$

This gives

$$\begin{aligned}\int_\Omega v\Delta^2 u &= \int_\Omega \Delta u\Delta v - \int_0^\pi v(x,-\ell)[u_{xxy}(x,-\ell)+u_{yyy}(x,-\ell)]dx\\ &+\int_0^\pi v(x,\ell)[u_{xxy}(x,\ell)+u_{yyy}(x,\ell)]dx\\ &+\int_0^\pi v_y(x,-\ell)[u_{xx}(x,-\ell)+u_{yy}(x,-\ell)]dx\\ &-\int_0^\pi v_y(x,\ell)[u_{xx}(x,\ell)+u_{yy}(x,\ell)]dx.\end{aligned}\tag{3.5}$$

By using (3.2), we obtain

$$\begin{aligned}\int_\Omega v\Delta^2 u &= \int_\Omega \Delta u\Delta v + (1-\sigma)\int_0^\pi [v(x,-\ell)u_{xxy}(x,-\ell)-v(x,\ell)u_{xxy}(x,\ell)]dx\\ &+(1-\sigma)\int_0^\pi [v_y(x,-\ell)u_{xx}(x,-\ell)-v_y(x,\ell)u_{xx}(x,\ell)]dx.\end{aligned}\tag{3.6}$$

By performing similar integration by part on the right hand side of (2.3), we obtain (3.6). Hence the result. □

Lemma 3.2 *The operator $A : D(A) \subset \mathcal{H} \longrightarrow \mathcal{H}$ is monotone.*

Proof Exploiting Lemma 3.1, we obtain, for all $U = \begin{pmatrix} u \\ v \end{pmatrix} \in D(A)$,

$$
\begin{aligned}
(AU, U)_{\mathcal{H}} &= \left(\begin{pmatrix} -v \\ \Delta^2 u + \delta_v \end{pmatrix}, \begin{pmatrix} u \\ v \end{pmatrix} \right)_{\mathcal{H}} \\
&= -(u, v)_{H^2_*(\Omega)} + \left(\Delta^2 u + \delta v, v\right)_{L^2(\Omega)} \\
&= -(u, v)_{H^2_*(\Omega)} + \left(\Delta^2 u, v\right)_{L^2(\Omega)} + \delta \|v\|^2_{L^2(\Omega)} = \delta \|v\|^2_{L^2(\Omega)} \geq 0. \quad (3.7)
\end{aligned}
$$

Thus, A is a monotone operator. □

Lemma 3.3 *The operator $A : D(A) \subset \mathcal{H} \longrightarrow \mathcal{H}$ is maximal, that is $R(I + A) = H$.*

Proof Let $G = (k, l) \in \mathcal{H}$ and consider the stationary problem

$$U + AU = G, \tag{3.8}$$

where $U = \begin{pmatrix} u \\ v \end{pmatrix}$. From (3.8) we obtain

$$\begin{cases} u - v = k, \\ v + \Delta^2 u + \delta v = l. \end{cases} \tag{3.9}$$

Combining $(3.9)_1$ and $(3.9)_2$ gives, for $\delta_0 = \delta + 1$,

$$\delta_0 u + \Delta^2 u = l + \delta_0 k. \tag{3.10}$$

The weak formulation of (3.10) is then

$$\delta_0 \int_\Omega u\phi + (u, \phi)_{H^2_*(\Omega)} = \int_\Omega (l + \delta_0 k)\phi, \quad \forall \phi \in H^2_*(\Omega). \tag{3.11}$$

We define the following bilinear and linear forms on $H^2_*(\Omega)$

$$B(u, \phi) = \delta_0 \int_\Omega u\phi + (u, \phi)_{H^2_*(\Omega)}, \tag{3.12}$$

$$\mathcal{F}(\phi) = \int_\Omega (l + \delta_0 k)\phi. \tag{3.13}$$

By using Lemmas 2.1 and 2.4, we show that B is bounded and coercive, and $\mathcal{F}$ is bounded. For this, we can easily see that

$$|B(u, \phi)| \leq C \|u\|_{H^2_*} \|\phi\|_{H^2_*}.$$

Furthermore, we have that

$$B(u,u) = \delta_0 \|u\|_{L^2}^2 + \|u\|_{H_*^2}^2 \geq \|u\|_{H_*^2}^2. \tag{3.14}$$

Therefore B is bounded and coercive.
Also,

$$|\mathcal{F}(\phi)| \leq \|l\|_{L^2}\|\phi\|_{L^2} + \delta_0\|k\|_{L^2}\|\phi\|_{L^2} \leq C(\|l\|_{L^2} + \delta_0\|k\|_{H_*^2})\|\phi\|_{H_*^2}.$$

This implies that $\mathcal{F}$ is bounded. Thus, Lax- Milgram Theorem guarantees the existence of a unique $u \in H_*^2(\Omega)$ satisfying (3.11), which yields

$$(u,\phi)_{H^{2*}(\Omega)} = \int_\Omega [l + \delta_0 k - \delta_0 u]\phi, \ \ \forall \phi \in H_*^2(\Omega). \tag{3.15}$$

Since $l + \delta_0 k - \delta_0 u \in L^2(\Omega)$, it follows from Lemma 2.3 that $u \in H^4(\Omega)$. Thus, we get $u \in H_*^2(\Omega) \cap H^4(\Omega)$. By performing similar integration by parts as in Lemma 3.1 to Eq. (3.11), we obtain

$$\begin{aligned}
&\int_\Omega [\delta_0 u + \Delta^2 u - l + \delta_0 k]\phi + \int_{-\ell}^{\ell} [u_{xx}(\pi,y)\phi_x(\pi,y) - u_{xx}(0,y)\phi_x(0,y)]dy \\
&+ \int_0^\pi \{[u_{yy}(x,\ell) + \sigma u_{xx}(x,\ell)]\phi_y(x,\ell) - [u_{yy}(x,-\ell) + \sigma u_{xx}(x,-\ell)]\phi_y(x,-\ell)\}dx \\
&+ \int_0^\pi [u_{yyy}(x,-\ell) + (2-\sigma)u_{xxy}(x,-\ell)]\phi(x,l)dx \\
&- \int_0^\pi [u_{yyy}(x,\ell) + (2-\sigma)u_{xxy}(x,\ell)]\phi(x,l)dx = 0, \ \ \forall \phi \in H_*^2(\Omega).
\end{aligned} \tag{3.16}$$

Now, by considering $\phi \in C_0^\infty(\Omega)$ ($hence \ \ \phi \in H_*^2(\Omega)$), then all the boundary terms of (3.16) vanish and we obtain

$$\int_\Omega [\delta_0 u + \Delta^2 u - l + \delta_0 k]\phi = 0, \ \ \forall \phi \in C_0^\infty(\Omega). \tag{3.17}$$

Hence (by density) we have

$$\int_\Omega [\delta_0 u + \Delta^2 u - l + \delta_0 k]\phi = 0, \ \ \forall \phi \in L^2(\Omega). \tag{3.18}$$

This implies

$$\delta_0 u + \Delta^2 u = l + \delta_0 k, \ \ in \ L^2(\Omega). \tag{3.19}$$

We take

$$v = u - k \ in \ H_*^2(\Omega)$$

and obtain

$$v + \Delta^2 u + \delta u = l, \ \ in \ L^2(\Omega).$$

Thus, $u \in H^2_*(\Omega) \cap H^4(\Omega)$ and $v \in H^2_*(\Omega)$ solves (3.9). Again, by choosing $\phi \in C^\infty(\bar{\Omega}) \cap H^2_*(\Omega)$ in (3.16) and using (3.19), we get

$$\begin{aligned}
&\int_\Omega [\delta_0 u + \Delta^2 u - w]\phi + \int_{-\ell}^{\ell} [u_{xx}(\pi, y)\phi_x(\pi, y) - u_{xx}(0, y)\phi_x(0, y)]dy \\
&+ \int_0^\pi \{[u_{yy}(x, \ell) + \sigma u_{xx}(x, \ell)]\phi_y(x, \ell) - [u_{yy}(x, -\ell) + \sigma u_{xx}(x, -\ell)]\phi_y(x, -\ell)\}dx \\
&+ \int_0^\pi [u_{yyy}(x, -\ell) + (2 - \sigma)u_{xxy}(x, -\ell)]\phi(x, l)dx \\
&- \int_0^\pi [u_{yyy}(x, \ell) + (2 - \sigma)u_{xxy}(x, \ell)]\phi(x, l)dx = 0.
\end{aligned} \tag{3.20}$$

By the arbitrary choice of $\phi \in C^\infty(\bar{\Omega}) \cap H^2_*(\Omega)$, we obtain from (3.20) the boundary conditions (3.2). Therefore there exists a unique

$$U = \begin{pmatrix} u \\ v \end{pmatrix} \in D(A)$$

satisfying (3.9). Thus, A is a maximal operator. □

Lemma 3.4 *The function F is Lipschtz.*

Proof Let $U, V \in \mathcal{H}$ and recall assumption $(2.1)_1$ to have

$$\begin{aligned}
\|F(U) - F(V)\|_{\mathcal{H}} &= \| \begin{pmatrix} 0 \\ -h(u) \end{pmatrix} - \begin{pmatrix} 0 \\ -h(\tilde{u}) \end{pmatrix} \|_{\mathcal{H}} \\
&= \| \begin{pmatrix} 0 \\ h(\tilde{u}) - h(u) \end{pmatrix} \|_{\mathcal{H}} = \|h(\tilde{u}) - h(u)\|_{L^2(\Omega)} \\
&\leq C\|u - \tilde{u}\|_{L^2(\Omega)} \leq C\|U - V\|_{\mathcal{H}}.
\end{aligned}$$

So, F is lipsctitz. □

Thus, by the semigroup theory [18], we have the following existence result.

Theorem 3.1 *Assume that (2.1) hold. Let $U_0 \in \mathcal{H}$ be given. Then the problem (P) has a unique weak solution*

$$U \in C([0, +\infty), \mathcal{H}).$$

Moreover, if h is linear and $U_0 \in D(A)$, then (P) has a unique strong solution

$$U \in C([0, +\infty), D(A)) \cap C^1([0, +\infty), \mathcal{H}).$$

Proof Follows from Lemmas 3.2, 3.3 and 3.4. □

4 Stability

In this section, we use the multiplier method (see Komornik [10]) to establish a stability result for the energy functional associated to problem (1.15).

Corollary 4.1 *We have*

$$\int_\Omega u\Delta^2 u = \|u\|^2_{H^2_*}, \quad \forall u \in D(A). \tag{4.1}$$

Proof Let $v = u$ in Lemma 3.1. □

The energy functional associated to problem (1.15) is defined by

$$E(t) = \frac{1}{2}\|u_t(t)\|^2_{L^2(\Omega)} + \frac{1}{2}\|u(t)\|^2_{H^2_*} + \int_\Omega H(u(t)). \tag{4.2}$$

Lemma 4.1 *Let $(u_0, u_1) \in D(A)$ be given and assume that (2.1) hold. Then the energy functional (4.2) satisfies*

$$\frac{dE(t)}{dt} = -\delta \int_\Omega u_t^2 \leq 0. \tag{4.3}$$

Proof Multiply $(1.15)_1$ by u_t and integrate over Ω to get

$$\frac{d}{dt}\left(\frac{1}{2}\int_\Omega u_t^2 + \frac{1}{2}\|u\|^2_{H^2_*} + \int_\Omega H(u)\right) + \delta \int_\Omega u_t^2 = 0. \tag{4.4}$$

Hence, the result. The inequality in (4.3) remains true for weak solution by simple density argument. Moreover, we get that E is a non-increasing functional. □

Theorem 4.1 *Let $(u_0, u_1) \in D(A)$ be given and assume (2.1) holds. Then, there exist constants $K > 0$, $\lambda > 0$ such that the energy functional (4.2) satisfies*

$$E(t) \leq Ke^{-\lambda t}, \quad \forall t \geq 0. \tag{4.5}$$

Proof We multiply $(1.15)_1$ by u and integrate over $\Omega \times (s, T)$, for $0 < s < T$ to get

$$\int_s^T \int_\Omega \left(u_{tt}u + u\Delta^2 u + uh(u) + \delta u u_t\right) = 0. \tag{4.6}$$

By using Corollary 4.1 we obtain

$$\int_s^T \int_\Omega (u_t u)_t - \int_s^T \int_\Omega u_t^2 + \int_s^T \|u\|_{H_*^2}^2 + \int_s^T \int_\Omega H(u) + \int_s^T \int_\Omega (uh(u) - H(u)) + \delta \int_s^T \int_\Omega u u_t = 0.$$

This gives

$$\int_s^T E(t)dt + \int_s^T \int_\Omega (u_t u)_t - \frac{3}{2}\int_s^T \int_\Omega u_t^2 + \frac{1}{2}\int_s^T \|u\|_{H_*^2}^2 + \int_s^T \int_\Omega (uh(u) - H(u)) + \delta \int_s^T \int_\Omega u u_t = 0.$$

By exploiting assumption (2.1), we obtain

$$\int_s^T E(t)dt \le -\int_s^T \int_\Omega (u_t u)_t + \frac{3}{2}\int_s^T \int_\Omega u_t^2 - \delta \int_s^T \int_\Omega u u_t. \tag{4.7}$$

Now, we estimate the terms on the right-hand side of (4.7). By using Lemma 2.4 and Young's inequality, the first term can be estimated as follows

$$\begin{aligned} |-\int_\Omega \int_s^T (u_t u)_t| &\le |\int_\Omega u_t(s)u(s)| + |\int_\Omega u_t(T)u(T)| \\ &\le \frac{1}{2}\int_\Omega u_t^2(s) + \frac{1}{2}\int_\Omega u^2(s) + \frac{1}{2}\int_\Omega u_t^2(T) + \frac{1}{2}\int_\Omega u^2(T) \\ &\le E(s) + C\|u(s)\|_{H_*^2}^2 + E(T) + C\|u(T)\|_{H_*^2}^2 \\ &\le CE(s) + CE(T) \le CE(s). \end{aligned} \tag{4.8}$$

For the second term, we have

$$\frac{3}{2}\int_s^T \int_\Omega u_t^2 = \frac{3}{2\delta}\int_s^T (-E'(t))dt = \frac{3}{2\delta}E(s) - \frac{3}{2\delta}E(T) \le \frac{3}{2\delta}E(s). \tag{4.9}$$

For the third term, we have for any $\epsilon > 0$ to be specified later

$$
\begin{aligned}
|-\delta\int_s^T\int_\Omega uu_t| &\le C_\epsilon\delta\int_s^T\int_\Omega u_t^2+\delta\frac{\epsilon}{2}\int_s^T\int_\Omega u^2\\
&\le C_\epsilon\delta\int_s^T(-E'(t))dt+\delta C_e\frac{\epsilon}{2}\int_s^T\|u\|_{H_*^2}^2\\
&\le C_\epsilon\delta E(s)+\delta C_e\frac{\epsilon}{2}\int_s^T E(t)dt. \qquad (4.10)
\end{aligned}
$$

Combining (4.8)–(4.10), we obtain

$$
\left(1-C_e\delta\frac{\epsilon}{2}\right)\int_s^T E(t)dt\le\left(C+\frac{3}{2\delta}+\delta C_\epsilon\right)E(s). \qquad (4.11)
$$

We then choose $\epsilon>0$ small enough so that $\left(1-C_e\delta\frac{\epsilon}{2}\right)>0$ and obtain

$$
\int_s^T E(t)dt\le CE(s),\ \forall\, s>0. \qquad (4.12)
$$

Letting T go to infinity and applying Lemma 2.5, we conclude from (4.12) the existence of two constants $K,\lambda>0$ such that the energy of the solution of (1.15) satisfies

$$
E(t)\le Ke^{-\lambda t},\ \ \forall t\ge 0. \qquad (4.13)
$$

This complete the proof. □

Remark 4.1 The decay estimate (4.5) remains valid for weak solutions by virtue of the density of $D(A)$ in $\mathcal{H}$.

Acknowledgements The authors thank King Fahd University of Petroleum and Mineral for its continuous support.

References

1. Ahmed, U.N., Harbi, H.: Mathematical analysis of the dynamics of suspension bridges. SIAM J. Appl. Math. **109**, 853–874 (1998)
2. Al-Gwaiz, M., Benci, V., Gazzola, F.: Bending and stretching energies in a rectangular plate modeling suspension bridges. Nonlinear Anal. **106**, 18–34 (2014)
3. Bochicchio, I., Giorgi, C., Vuk3, E.: Long-term damped dynamics of the extensible suspension bridge. Int. J. Differ. Eq. Article ID 383420 (2010)
4. Choi, Q.H., Jung, T.: A nonlinear suspension equation with nonconstant load. Nonlinear Anal. **35**, 649–668 (1999)
5. Ferrero, A., Gazzola, F.: A partially hinged rectangular plate as a model for suspension bridges. Discrete Continuous Dyn. Syst. **35**(12), 5879–5908 (2015)
6. Gazzola, F.: Nonlinearity in oscillating bridges. Electron. J. Differ. Eq. **211**, 1–47 (2013)

7. Filippo, D., Giorgi, C., Pata, V.: Asymptotic behaviour of coupled linear systems modeling suspension bridges. Z. Angew. Math. Phys. **66**, 1095–1108 (2015)
8. Glover, J., Lazer, A.C., Mckenna, P.J.: Existence and stability of of large scale nonlinear oscillation in suspension bridges. Z. Angew. Math. Phys. **40**, 172–200 (1989)
9. Imhof, D.: Risk assessment of existing bridge structure. Ph.D. Dissertation, University of Cambridge (2004)
10. Komornik, V.: Exact Controllability and Stabilization. The Multiplier Method. Masson-John Wiley, Paris (1994)
11. Mansfield, E.H.: The Bending and Stretching of Plates, 2nd edn. Cambridge Univ, Press (2005)
12. Lazer, A.C., McKenna, P.J.: Large-amplitude periodic oscillations in suspension bridges: some new connections with non-linear analysis. SIAM Rev. **32**(4), 537–578 (1990)
13. McKenna, P.J.: Torsional oscillations in suspension bridges revisited: fixing an old approximation. Amer. Math. Monthly **106**, 1–18 (1999)
14. McKenna, P.J., Walter, W.: Non-linear oscillations in a suspension bridge. Arch. Ration. Mech. Anal. **98**(2), 167–177 (1987)
15. McKenna, P.J., Walter, W.: Travelling waves in a suspension bridge. SIAM J. Appl. Math. **50**(3), 703–715 (1990)
16. McKenna, P.J., Tuama, C.: Large torsional oscillations in suspension bridges visited again: vertical forcing creates torsional response. Amer. Math. Monthly **108**, 738–745 (2001)
17. Navier C.L.: Extraits des recherches sur la flexion des plans élastiques. Bulletin des Sciences de la Société Philomathique de Paris 92–102 (1823)
18. Pazzy, A.: Semigroups of linear operators and application to PDE. Appl. Math. Sci. **44** Springer (1983)
19. Ventsel, E., Krauthammer, T.: Thin Plates and Shells: Theory, Analysis, and Applications. Marcel Dekker Inc., New York (2001)
20. Wang, Y.: Finite time blow-up and global solutions for fourth-order damped wave equations. J. Math. Anal. Appl. **418**(2), 713–733 (2014)

An Interpolation-Based Approach to American Put Option Pricing

Greg Orosi

Abstract In this paper, we discuss how to construct interpolation-based models for American put options. In particular, we derive a closed-form expression and suggest multi-parameter extensions. Our result makes no assumption about the dynamics of the underlying asset, and is constructed to satisfy the necessary no-arbitrage conditions. Finally, we discuss potential applications.

Keywords American option · Put option · Arbitrage-free conditions

MSC[2010] code: 91E45

1 Introduction

Since the seminal work of Black and Scholes [4], a common approach in the derivative pricing literature has been to model the underlying asset's price dynamics by a stochastic process. Since option prices based on the geometric Brownian motion model of Black and Scholes do not provide a reasonable fit to observed market prices, several extensions have been proposed. Well-known examples of such models include the jump-diffusion model of Merton [12], the stochastic volatility model of Heston [10], and the local volatility model of Dupire [7].

Although these extensions have a significantly better performance than the model of Black and Scholes, Epps [8], Alghalith [1] point out that there is no universal model that provides a consistently good fit to observed option prices. Moreover, Figlewski [9], Alghalith [2] point out that simple formulas, which make no assumption about the underlying process, can produce good results. This idea has been explored by Orosi [13] who finds that a nonparametric extension of Figlewski's model provides a nearly perfect fit to European call options on the S&P 500 index. Additionally,

G. Orosi (✉)
Department of Mathematics and Statistics, American University of Sharjah,
P.O. Box 26666, Sharjah, UAE
e-mail: gorosi@aus.edu

T. Abualrub et al. (eds.), *Mathematics Across Contemporary Sciences*,
Springer Proceedings in Mathematics & Statistics 190,
DOI 10.1007/978-3-319-46310-0_10

interpolation-based models can be used to extract information from European call options (see for example (Orosi [14, 15]).

Motivated by these results, in this work, we introduce an interpolation-based model for American put options. Instead of making an assumption about the underlying, we construct a suitable pricing function based on no-arbitrage conditions. The rest of the paper is organized as follows. In Sect. 2, we state the necessary no-arbitrage constraints and review how to construct interpolation-based put option prices that satisfy well-known no arbitrage conditions. Moreover, we derive a closed-form American put option formula and introduce a three-parameter model. To illustrate the applicability of our method, we calibrate our models to market quotes in Sect. 3. Section 4 discusses the practical use of our findings and we propose further extensions. Finally, Sect. 5 presents our conclusions.

2 Constructing Suitable Pricing Functions

Merton [11] shows that an American put option function, $P(K,T)$, with strike price K and time to expiry T, must satisfy the following no-arbitrage conditions: (i) $P(K,T)$ is a convex and increasing function of K; (ii) $P(K,T) \leq K$; (iii) $P(K,T)$ is an increasing and convex function of T; (iv) $\max(0, K-S) \leq P(K,T)$ where S is the stock price; (v) $P(0,T) = 0$; (vi) $\lim_{K\to\infty} \frac{P(K,T)}{K-S} = 1$. Moreover, Carr and Wu [6] show that for put options written on non-defaultable assets $\left.\frac{\partial P(K,T)}{\partial K}\right|_{K=0} = 0$. We will refer to these as the necessary no-arbitrage conditions.

To determine suitable put option functions, first, we apply the following transformations to the strikes and put option prices:

$$p = \frac{P(K,T)}{S}$$
$$k = \frac{K}{S}.$$

The main idea behind our approach is that it easier to construct suitable pricing functions in the space that is rotated counterclockwise by 45°. Furthermore, we introduce the following function:

$$y = G\frac{x^2}{\left(\frac{1}{\sqrt{2}} - x\right)} + x, \tag{1}$$

where y and x represent the rotated values of p and k, respectively. Then, the put option prices recovered from the above equation satisfy the necessary no-arbitrage properties (see the results in the Appendix).

To obtain the relation between p and k, one must apply the transformation to x and y that rotates these clockwise by 45°. This transformation gives the following

relation between the above variables:

$$p = \frac{y - x}{\sqrt{2}}$$
$$k = \frac{y + x}{\sqrt{2}}.$$

Moreover, substituting (1) for y gives

$$p = \frac{1}{\sqrt{2}} \left(G \frac{x^2}{\left(\frac{1}{\sqrt{2}} - x\right)} + x \right) - \frac{1}{\sqrt{2}} x,$$

where

$$x = \frac{k - p}{\sqrt{2}}.$$

Finally, an analytic solution for p can be obtained from

$$p = \frac{1}{\sqrt{2}} \left(G \frac{\left(\frac{k-p}{\sqrt{2}}\right)^2}{\left(\frac{1}{\sqrt{2}} - \frac{k-p}{\sqrt{2}}\right)} \right) \tag{2}$$

that is given by

$$p = \frac{-\sqrt{2(G-1)k + k^2 + 1} + (G-1)k + 1}{G - 2}. \tag{3}$$

Therefore, the arbitrage-free American put option function for a fixed maturity is given by

$$\begin{aligned} P(K, T) &= S \cdot p \\ &= \frac{-S\left(\sqrt{2(G-1)k + k^2 + 1} + (G-1)k + 1\right)}{G - 2}. \end{aligned} \tag{4}$$

Moreover, to incorporate the property that put options are an increasing function of expiry, larger values of G can be fitted to longer expiries. This is illustrated in Fig. 1 that plots arbitrage-free American put prices with two different parameters.

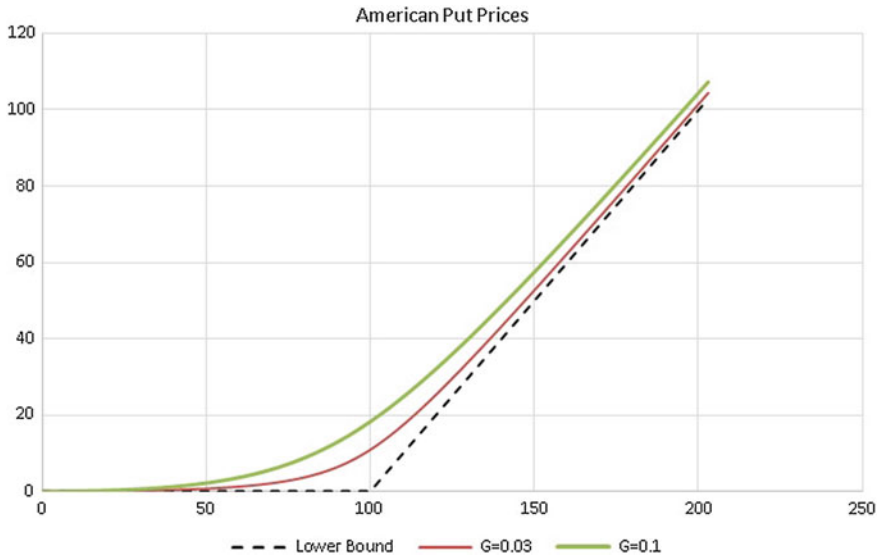


Fig. 1 American put prices generated by the single-parameter model for two different values: $G = 0.03$ and $G = 0.1$. The value of the stock price is assumed to be 100

2.1 A Multi-Parameter Extension

Instead of (1), one could consider the following relation with three parameters:

$$y = G\frac{x^{\alpha}}{\left(\frac{1}{\sqrt{2}} - x\right)^{\beta}} + x. \tag{5}$$

It can be easily shown that the necessary arbitrage-free conditions are satisfied if the following the constraints are imposed on the parameters: $\alpha > 1$ and $\beta > 0$ (see the results in the Appendix). Although this three-parameter model provides a better fit to observed option prices, there is no analytic solution for the put prices. Therefore, these have to be determined numerically.

3 An Illustrative Example

In this section, to demonstrate the applicability of the models in (4) and (5), we calibrate these to market quotes. The models are fitted to near-the-money American put options written on Apple Inc. stock. We only consider options with the fixed

expiry $T = 1.093$ on December 18, 2015. Moreover, the models are calibrated by minimizing the Non-Linear Least Squares (NLS) objective:

$$\sum_{i=1}^{n} \left(P_i\left(\theta\right) - P_i\right)^2,$$

where the P_i-s are the market prices for options, and the $P_i\left(\theta\right)$-s are the put option prices based on the model. The results are presented in Table 1 for the single-parameter model and in Table 2 for the three-parameter model. It can be observed that both models yields prices that are very close to the market prices. Moreover, all of the put option prices based on the three-parameter lie inside the bid-ask spread.

Although we leave rigorous empirical analysis of the performance of the models for further research, some advantages of the models to traditional approaches can be easily highlighted. For example, Brooks and Chance [5] point out that the most commonly used option pricing models rely on a constant interest rate as an input. However, according to them, it is difficult to determine what interest rate one should use for the purposes of option pricing. Moreover, they point out that even small errors in the interest rate used can lead to misestimated option prices and implied volatilities. In particular, the prices of American options are very sensitive because of the impact of the interest rate on early exercise. Since our models do not use interest rate as an input, option prices and hedge ratios can be efficiently and accurately calculated from observed market prices.

Table 1 The resulting call option prices of the one-parameter model with the best fit parameter G on Apple Inc. stock with $T = 1.093$ on December 18, 2015

$S = 108.98$				
$G = 0.033$				
	Market price	Model price	BID	ASK
$K = 92.5$	6.03	6.03	5.9	6.15
$K = 95$	6.83	6.83	6.75	6.9
$K = 97.5$	7.68	7.68	7.55	7.8
$K = 100$	8.63	8.63	8.5	8.75
$K = 105$	10.78	10.78	10.7	10.85
$K = 110$	13.23	13.23	13.1	13.35
$K = 115$	16.03	16.03	15.9	16.15
$K = 120$	19.15	19.15	19.05	19.25
$K = 125$	22.58	22.58	22.45	22.7
$K = 130$	26.28	26.28	26.1	26.45

Table 2 The resulting call option prices of the three-parameter with the best fit parameter G, α and β on Apple Inc. stock with $T = 1.093$ on December 18, 2015

$S = 108.98$				
$G = 0.4789$, $\alpha = 4.6233$, $\beta = 0.4425$				
	Market price	Model price	BID	ASK
$K = 92.5$	6.03	5.99	5.9	6.15
$K = 95$	6.83	6.78	6.75	6.9
$K = 97.5$	7.68	7.66	7.55	7.8
$K = 100$	8.63	8.60	8.5	8.75
$K = 105$	10.78	10.74	10.7	10.85
$K = 110$	13.23	13.20	13.1	13.35
$K = 115$	16.03	15.99	15.9	16.15
$K = 120$	19.15	19.12	19.05	19.25
$K = 125$	22.58	22.55	22.45	22.7
$K = 130$	26.28	26.28	26.1	26.45

4 Applications and Extensions

4.1 Model-Free Hedge Ratios

Bates [3], Reiss and Wystup [16] point out that a model-free deltas and gammas can be calculated if one has a continuous set of options as a function of strikes. For example, for puts and calls, it is reasonable to assume that

$$O(aK, aS, T) = aO(K, S, T),$$

where $O(K, S, T)$ is the option price, K is the strike price, S is the price of the asset, and a is a positive constant. Then, a model-free delta, O_S, can be calculated from the expression:

$$O_S S + O_K K = O(K, S, T). \tag{6}$$

Similarly, deltas and kappas satisfy the relations:

$$O_S(aK, aS, T) = O_S(K, S, T)$$
$$O_K(aK, aS, T) = O_K(K, S, T).$$

From the above, the following equations can be obtained:

$$O_{SS} S + O_{KS} K = 0$$
$$O_{KS} S + O_{KK} K = 0.$$

Then, eliminating O_{KS} from the above yields

$$O_{SS}S^2 = O_{KK}K^2. \tag{7}$$

Finally, O_{KK} and O_K can be calculated numerically or analytically, and model-free deltas and gammas can be obtained.

4.2 A Further Extension

Instead of (1) or (5), one could consider the following relation:

$$y = G\frac{s(x)}{\left(\frac{1}{\sqrt{2}} - x\right)^{\beta}} + x,$$

where $s(x)$ is a function with an arbitrary number of parameters or a nonparametric function. Note that if $\beta > 0$, $s(0) = 0$, $s'(0) = 0$, $s'(x) \geq 0$, and $s''(x) \geq 0$, then the put option prices satisfy the necessary no-arbitrage conditions (see the results in the Appendix).

5 Conclusion

In this paper, we demonstrate how to construct interpolation-based models for American put options. Our approach is constructed to satisfy the necessary no-arbitrage conditions, and makes no assumption about the dynamics of the underlying asset. We also briefly explain the advantage of the proposed models and discuss potential applications.

Appendix

In this section, we show that put prices obtained from (1) or (5) satisfy the necessary no-arbitrage conditions.

Claim 1 If a function $y = f(x)$ is convex on $x \in [0, \infty)$ and $f'(x) \geqslant 1$, then the resulting function obtained by rotating $f(x)$ clockwise by 45° is convex and increasing on $[0, \infty)$.

Proof First, note that the rotated function $p = g(k)$ is given by the equations

$$p = \frac{1}{\sqrt{2}} y - \frac{1}{\sqrt{2}} x$$
$$k = \frac{1}{\sqrt{2}} y + \frac{1}{\sqrt{2}} x.$$

From the above, the following can be obtained:

$$p = \frac{1}{\sqrt{2}} (f(x) - x) = \frac{1}{\sqrt{2}} \left(f\left(\frac{k-p}{\sqrt{2}} \right) - \frac{k-p}{\sqrt{2}} \right).$$

Then,

$$\frac{dp}{dk} = p' = \frac{1}{\sqrt{2}} \left(f'\left(\frac{k-p}{\sqrt{2}} \right) \frac{1-p'}{\sqrt{2}} - \frac{1-p'}{\sqrt{2}} \right) = \frac{1}{2} \left(f'\left(\frac{p-k}{\sqrt{2}} \right) - 1 \right) (1-p')$$

and

$$p' = \frac{\frac{1}{2} \left(f'\left(\frac{k-p}{\sqrt{2}} \right) - 1 \right)}{1 + \frac{1}{2} \left(f'\left(\frac{k-p}{\sqrt{2}} \right) - 1 \right)}.$$

Therefore, $p' \geqslant 0$ because $f'(x) \geqslant 1$. Moreover,

$$\frac{d^2 p}{dk^2} = p'' = \frac{1}{2} \left(f''\left(\frac{k-p}{\sqrt{2}} \right) \left(\frac{1-p'}{\sqrt{2}} \right) \right) (1-p') + \frac{1}{2} \left(f'\left(\frac{k-p}{\sqrt{2}} \right) - 1 \right) (-p'')$$

and

$$\frac{d^2 p}{dk^2} = p'' = \frac{\frac{1}{2} \left(f''\left(\frac{k-p}{\sqrt{2}} \right) \left(\frac{1-p'}{\sqrt{2}} \right) \right) (1-p')}{1 + \frac{1}{2} \left(f'\left(\frac{k-p}{\sqrt{2}} \right) - 1 \right)}.$$

Therefore, $p'' \geqslant 0$ because both the numerator and denominator are positive. ■

Moreover, if a continuous function $y = f(x)$ is convex on $x \in \left[0, \frac{1}{\sqrt{2}}\right)$, $f'(x) \geqslant 1$, and $\lim_{x \to \frac{1}{\sqrt{2}}} f(x) = \infty$, then it is bounded by the lines: $y = x$, $x = \frac{1}{\sqrt{2}}$, and the y-axis. Hence, the function obtained by rotating $f(x)$ clockwise by 45° is bounded by the following lines: the k-axis (the horizontal axis), $p = k$ (that line that passes through the original and has a slope of 1), and $p = k - 1$. Moreover,

$$\lim_{p \to \infty} \frac{g(k)}{k-1} = 1,$$
$$g(0) = 0 \text{ iff } f(0) = 0,$$
$$\text{and } f'(0) = 1 \text{ iff } g'(0) = 0.$$

Consequently, we require $f(x)$ to satisfy the following requirements: $f(0) = 0$, $f'(0) = 1$, $\lim_{x \to \frac{1}{\sqrt{2}}} f(x) = \infty$, $f(x)$ is convex on $x \in \left[0, \frac{1}{\sqrt{2}}\right)$, and $f'(x) \geqslant 1$. Then, the put prices obtained from

$$P(K, T) = P(k \cdot S, T) = S \cdot p = S \cdot g(k)$$

satisfy the necessary no-arbitrage conditions with the exception of condition (iii). Finally, condition (iii) is satisfied if the parameter G in (4) or (5) is and increasing function of T.

References

1. Alghalith, M.: A new stopping time model: a solution to a free-boundary problem. J. Optim. Theory Appl. **152**(1), 265–270 (2012)
2. Alghalith, M.: Option pricing: very simple formulas. J. Deriv. Hedge Funds **20**(2), 71–73 (2014)
3. Bates, D.S.: Hedging the smirk. Finan. Res. Lett. **2**(4), 195–200 (2005)
4. Black, F., Scholes, M.S.: The pricing of options and corporate liabilities. J. Polit. Econ. **81**(3), 637–659 (1973)
5. Brooks, R., Chance, D.M.: Some subtle relationships and results in option pricing. J. Appl. Finan. **24**(1), 94–110 (2014)
6. Carr, P., Wu, L.: A simple robust link between American puts and credit protection. Rev. Finan. Stud. **24**(2), 473–505 (2011)
7. Dupire, B.: Pricing with a smile. Risk **7**(1), 18–20 (1994)
8. Epps, T.W.: Quantitative Finance: Its Development, Mathematical Foundations, and Current Scope. Wiley, New Jersey (2009)
9. Figlewski, S.: Assessing the incremental value of option pricing theory relative to an informationally passive benchmark. J. Deriv. **10**(1), 80–96 (2002)
10. Heston, S.L.: A closed-form solution for options with stochastic volatility applications to bond and currency options. Rev. Financ. Stud. **6**(2), 327–343 (1993)
11. Merton, R.C.: Theory of rational option pricing. Bell J. Econ. Manag. Sci. **4**(1), 141–183 (1973)
12. Merton, R.C.: Option pricing when underlying stock returns are discontinuous. J. Financ. Econ. **3**(1–2), 125–144 (1976)
13. Orosi, G.: Arbitrage-free call option surface construction. Appl. Stoch. Models Bus. Indus. **31**(4), 515–527 (2015a)
14. Orosi, G.: Closed-form interpolation-based formulas for European call options written on defaultable assets. J. Asset Manag. **16**(4), 236–242 (2015b)
15. Orosi, G.: Estimating option-implied risk-neutral densities: a novel parametric approach. J. Deriv. **23**(1), 41–61 (2015c)
16. Reiss, O., Wystup, U.: Computing option price sensitivities using homogeneity and other tricks. J. Deriv. **9**(2), 41–53 (2001)

Stable Homotopy Groups of Moore Spaces

Inès Saihi

Abstract We determine explicitly the stable homotopy groups of Moore spaces up to the range 7, using an equivalence of categories which allows to consider each Moore space as an exact couple of $\mathbb{Z}$-modules.

Keywords Moore spaces · Stable homotopy groups · Equivalence of categories

1991 Mathematics Subject Classification Primary 55Q10 · Secondary 55U20 · 18G99

1 Introduction

Moore spaces and their stable homotopy groups were widely studied and a complete reference on this subject is the book of Baues [1].

In this paper, we propose a new approach allowing to see Moore spaces as exact couples of $\mathbb{Z}$-modules by means of an equivalence of categories. Even though a similar result is proven in [1], the approach given here is of independent interest, since it is used to determine explicitly the stable homotopy groups of Moore spaces up to the range 7.

Let G be an abelian group and n an integer greater than 1. A Moore space $M(G, n)$ is a simply connected CW-complex X such that $H_n(X) \simeq G$ and $\widetilde{H}_i(X) = 0$ for $i \neq n$. The homotopy type of $M(G,n)$ is uniquely determined by the pair (G,n) (see [6]).

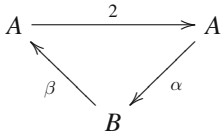
I. Saihi (✉)
Université de Tunis, École nationale supérieure d'ingénieurs de Tunis,
Tunis, Tunisia
e-mail: ines.saihi@esstt.rnu.tn

I. Saihi
Laboratoire LATAO, faculté des sciences de Tunis, université de Tunis-El Manar,
Tunis, Tunisia

T. Abualrub et al. (eds.), *Mathematics Across Contemporary Sciences*,
Springer Proceedings in Mathematics & Statistics 190,
DOI 10.1007/978-3-319-46310-0_11

Let $\mathscr{M}_n$ be the category whose objects are Moore spaces $M(A,n)$, where A is a $\mathbb{Z}$-module, and whose morphisms are homotopy classes of pointed maps between such Moore spaces. Notice that, unlike the Eilenberg-MacLane, the set of homotopy classes of pointed maps $[M(A,n)\ ,\ M(B,n)]$ between two Moore spaces is different from $\mathrm{Hom}(A,B)$ (see proposition 2.1).

Let $\mathcal{M}od$ be the category of $\mathbb{Z}$-modules and let $\mathscr{D}_e$ be the category of exact couples in $\mathcal{M}od$



such that $\alpha\beta = 2$.

There are two exact functors Φ_1 and Φ_2 from $\mathscr{D}_e$ to $\mathcal{M}od$ assigning to a diagram the $\mathbb{Z}$-module A or B respectively.

The aim of Sect. 2 is to construct, for $n \geq 3$, an equivalence of categories $\mathcal{E}$ between $\mathscr{M}_n$ and $\mathcal{D}_e$. In [1] and in a different context, Baues gave a similar result using the properties of the Whitehead Γ-functor.

In Sect. 3, the stable homotopy groups $\pi_i^S(X)$ $(0 \leq i \leq 7)$ of a Moore space X will be expressed in term of $\mathcal{E}(X)$. The same techniques can be used to determine $\pi_i^S(M(A,n))$ for $i \geq 8$, but calculations become complicated.

2 Equivalence of Categories Between Moore Spaces and Diagrams

2.1 Category of Diagrams

In this section, we propose an equivalence of categories that allows to consider Moore spaces as diagrams of $\mathbb{Z}$-modules.

Recall that the suspension functor from $\mathscr{M}_n$ to $\mathscr{M}_{n+1}$ is an equivalence of categories for $n \geq 3$, so next results are independent of n.

Consider two modules A et B. Let X be the Moore space $X = M(A,n)$, Y the Moore space $Y = M(B,n)$ and $[X,Y]$ the set of homotopy classes of pointed maps from X to Y; this set is an abelian group (see [2]). Moreover:

Proposition 2.1 (*[1], [2]*) *There is a natural exact sequence:*

$$0 \longrightarrow \mathrm{Ext}(A,B/2) \longrightarrow [X,Y] \longrightarrow \mathrm{Hom}(A,B) \longrightarrow 0. \tag{2.1}$$

Set $S = M(\mathbb{Z},n)$ and $P = M(\mathbb{Z}/2,n)$. Applying the exact sequence (2.1) to S and X, we obtain the exact sequence:

$$0 \longrightarrow \mathrm{Ext}(\mathbb{Z}, A/2) \longrightarrow [S,X] \longrightarrow \mathrm{Hom}(\mathbb{Z}, A) \longrightarrow 0$$

and $[S,X]$ is isomorphic to A. Similarly, applied to P and X, (2.1) becomes:

$$0 \longrightarrow \mathrm{Ext}(\mathbb{Z}/2, A/2) \longrightarrow [P,X] \longrightarrow \mathrm{Hom}(\mathbb{Z}/2, A) \longrightarrow 0.$$

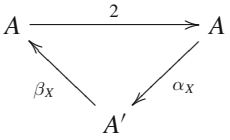
Since $\mathrm{Ext}(\mathbb{Z}/2, A/2)$ is naturally isomorphic to $A/2$ (see Proposition 2.7), we have the exact sequence:

$$0 \longrightarrow A/2 \longrightarrow A' \longrightarrow A_2 \longrightarrow 0 \tag{2.2}$$

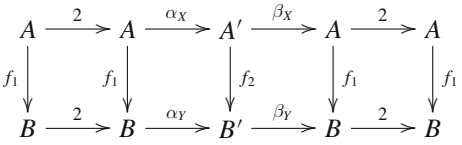
where A' is the module $[P,X]$ and A_2 is the set of order 2 elements in A. In other words, we have the long exact sequence:

$$A \xrightarrow{2} A \xrightarrow{\alpha_X} A' \xrightarrow{\beta_X} A \xrightarrow{2} A \tag{2.3}$$

or equivalently, the exact couple denoted by D_X:



Moreover, if f is a map between two Moore spaces $X = M(A,n)$ and $Y = M(B,n)$, then we can deduce a map $\bar{f} : D_X \longrightarrow D_Y$ as follows: $\bar{f} = (f_1, f_2)$ where $f_1 : A \simeq [S,X] \longrightarrow B \simeq [S,Y]$ and $f_2 : A' = [P,X] \longrightarrow B' = [P,Y]$ are the natural maps induced by f. The following diagrams commute:



2.1.1 Particular Case of P

When $X = P$, we have the next results:

Proposition 2.2 $[P,P] \simeq \mathbb{Z}/4$.

The proof of this result can be found in [2] or [7].

Lemma 2.3 *The composition $\alpha_P \beta_P$ is multiplication by 2 on $\mathbb{Z}/4$.*

Proof When $X = P$, the exact sequence (2.3) becomes:

$$\mathbb{Z}/2 \xrightarrow{2} \mathbb{Z}/2 \xrightarrow{\alpha_P} \mathbb{Z}/4 \xrightarrow{\beta_P} \mathbb{Z}/2 \xrightarrow{2} \mathbb{Z}/2$$

i.e.:

$$0 \longrightarrow \mathbb{Z}/2 \xrightarrow{\alpha_P} \mathbb{Z}/4 \xrightarrow{\beta_P} \mathbb{Z}/2 \longrightarrow 0$$

then $\alpha_P = 2$ and β_P is the canonical surjection. Hence $\alpha_P\beta_P$ is the multiplication by 2 on $\mathbb{Z}/4$.

2.1.2 General Case

Lemma 2.4 *For any Moore space $X = M(A,n)$, the composition $\alpha_X\beta_X : A' \longrightarrow A'$ is multiplication by 2.*

Proof Let $u \in A' = [P,X]$ and $f : P \longrightarrow X$ a representative of u. We have two maps f_1 and f_2 and the commutative diagram:

$$\begin{array}{ccccccccc}
\mathbb{Z}/2 & \xrightarrow{2=0} & \mathbb{Z}/2 & \xrightarrow{\alpha_P} & \mathbb{Z}/4 & \xrightarrow{\beta_P} & \mathbb{Z}/2 & \xrightarrow{2=0} & \mathbb{Z}/2 \\
\downarrow{\scriptstyle f_1} & & \downarrow{\scriptstyle f_1} & & \downarrow{\scriptstyle f_2} & & \downarrow{\scriptstyle f_1} & & \downarrow{\scriptstyle f_1} \\
A & \xrightarrow{2} & A & \xrightarrow{\alpha_X} & A' & \xrightarrow{\beta_X} & A & \xrightarrow{2} & A
\end{array}$$

If u_0 denotes the class of the identity map in $[P,P]$, then $f_2(u_0) = u$. The result is an immediate consequence of Lemma 2.3.

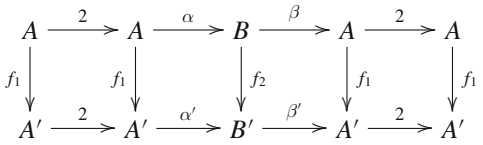
2.1.3 Category of Diagrams

Definition 2.5 Let $\mathscr{D}_e$ be the category of exact couples in the category $\mathcal{M}od$ of $\mathbb{Z}$-modules

$$\begin{array}{ccc}
A & \xrightarrow{2} & A \\
& {\scriptstyle\beta}\nwarrow \quad \swarrow{\scriptstyle\alpha} & \\
& B &
\end{array} \tag{2.4}$$

such that $\alpha\beta = 2$.

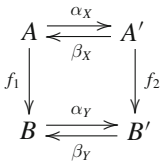
A morphism f between two objects D and D' is a couple $f = (f_1, f_2)$ such that the following diagrams commute:



Notations: For ease, an object of $\mathscr{D}_e$ will be denoted by

$$A \underset{\beta}{\overset{\alpha}{\rightleftarrows}} B$$

and a morphism between two objects D and D' will be denoted by



The previous constructions can be summarized in the following statement:

Proposition 2.6 *There is a functor $\mathcal{E} : \mathscr{M}_n \longrightarrow \mathscr{D}_e$ assigning to each Moore space X the diagram D_X, and to each homotopy class f of pointed maps between two Moore spaces X and Y the map $\bar{f} : D_X \longrightarrow D_Y$.*

In the remaining of this section, we will prove that the functor $\mathcal{E}$ is an equivalence of categories.

Notations: Let Φ_1 and Φ_2 denote the two functors from $\mathscr{D}_e$ to $\mathcal{M}od$ defined as follows: if D is an object of$\mathscr{D}$ given by:

$$A \underset{\beta}{\overset{\alpha}{\rightleftarrows}} B \tag{2.5}$$

then $\Phi_1(D) = A$ and $\Phi_2(D) = B$.

Notice that there is a natural transformation between functors Φ_1 and Φ_2 obtained by associating to a diagram D given by (2.5), the morphism α. By associating to the diagram D the morphism β, we get a natural transformation from Φ_2 to Φ_1.

2.2 Equivalence of Categories Between $\mathscr{M}_n$ and $\mathscr{D}_e$

2.2.1 Some Algebraic Results

This section is devoted to prove some general algebraic results needed to obtain the equivalence of categories announced above.

Proposition 2.7 *For every $\mathbb{Z}/2$-modules A and B, there is an isomorphism $\lambda_{(A,B)}$, natural in A and in B:*

$$\lambda_{(A,B)} : \mathrm{Ext}(A,B) \xrightarrow{\sim} \mathrm{Hom}(A,B).$$

Proof An element e of $\mathrm{Ext}(A,B)$ is represented by an extension:

$$0 \longrightarrow B \xrightarrow{f} E \xrightarrow{g} A \longrightarrow 0.$$

Each element $x \in A$ is of order 2 and g is surjective, so there is $y \in E$ such that $g(y) = x$ and $2y \in \ker g = \mathrm{Im} f$. Since f is injective, there exists a unique $z \in B$ such that $f(z) = 2y$. The map assigning to x the element z is well defined; then we obtain a morphism:

$$\lambda_{(A,B)} : \mathrm{Ext}(A,B) \longrightarrow \mathrm{Hom}(A,B).$$

Since A is free, there is a natural isomorphism $\mathrm{Ext}(A,B) \longrightarrow \mathrm{Hom}(A,\mathrm{Ext}(\mathbb{Z}/2,B))$ obtained by restriction. (Each $a \in A$ defines a map $\mathbb{Z}/2 \longrightarrow A$ which induces an extension of $\mathbb{Z}/2$ by B using a pull-back.) But $\mathrm{Ext}(\mathbb{Z}/2,B)$ is naturaly isomorphic to B, so we get an isomorphism from $\mathrm{Ext}(A,B)$ to $\mathrm{Hom}(A,B)$, which is $\lambda_{(A,B)}$.

Remark 2.8 If A and B are two $\mathbb{Z}$-modules, we construct similarly a natural morphism

$$\lambda_{(A,B)} : \mathrm{Ext}(A,B) \longrightarrow \mathrm{Hom}(A_2, B/2)$$

obtained by the composition:

$$\lambda_{(A,B)} : \mathrm{Ext}(A,B) \longrightarrow \mathrm{Ext}(A_2,B/2) \xrightarrow{\lambda_{(A_2,B/2)}} \mathrm{Hom}(A_2,B/2),$$

where the first morphism is induced by restriction to order 2 elements in A and the projection of B on $B/2$.

Corollary 2.9 *If A is a $\mathbb{Z}/2$-module and B a $\mathbb{Z}$-module, then* $\mathrm{Ext}(A,B)$ *is isomorphic to* $\mathrm{Hom}(A,B/2)$.

Proof The morphism $\lambda_{(A,B)}$ is the composition

$$\lambda_{(A,B)} : \mathrm{Ext}(A,B) \xrightarrow{pr} \mathrm{Ext}(A,B/2) \xrightarrow{\lambda_{(A,B/2)}} \mathrm{Hom}(A,B/2)$$

where pr is the morphism induced by the projection of B on $B/2$. By (2.7), $\lambda_{(A,B/2)}$ is an isomorphism; it suffices to show that $pr : \mathrm{Ext}(A,B) \longrightarrow \mathrm{Ext}(A,B/2)$ is bijective. But A is a $\mathbb{Z}/2$-module, so A is free and then can be written $A = \oplus\mathbb{Z}/2$. Since $\mathrm{Ext}(\oplus\mathbb{Z}/2,B) = \prod \mathrm{Ext}(\mathbb{Z}/2,B)$ we can show the result for $A = \mathbb{Z}/2$. Using the resolution

$$0 \longrightarrow \mathbb{Z} \xrightarrow{2} \mathbb{Z} \longrightarrow \mathbb{Z}/2 \longrightarrow 0,$$

we get the diagram:

$$\begin{array}{ccccccccc}
\mathrm{Hom}(\mathbb{Z}/2,B) & \longrightarrow & B & \xrightarrow{2} & B & \longrightarrow & \mathrm{Ext}(\mathbb{Z}/2,B) & \longrightarrow & 0 \\
\downarrow{\scriptstyle \simeq} & & \downarrow{\scriptstyle pr} & & \downarrow{\scriptstyle pr} & & \downarrow{\scriptstyle \simeq} & & \\
B_2 & \longrightarrow & B/2 & \xrightarrow{2=0} & B/2 & \longrightarrow & \mathrm{Ext}(\mathbb{Z}/2,B/2) & \longrightarrow & 0
\end{array}$$

Corollary 2.10 *If A is a $\mathbb{Z}$-module and B a $\mathbb{Z}/2$-module, then* $\mathrm{Ext}(A,B) \simeq \mathrm{Hom}(A_2,B)$.

Proof Since the morphism $\lambda_{(A,B)}$ is the composition

$$\lambda_{(A,B)} : \mathrm{Ext}(A,B) \xrightarrow{\mathrm{R}} \mathrm{Ext}(A_2,B) \xrightarrow{\lambda_{(A_2,B)}} \mathrm{Hom}(A_2,B),$$

where R is the morphism induced by the restriction to A_2, and $\lambda_{(A_2,B)}$ is an isomorphism, we have just to show that R is bijective.

But B is free, so $B = \oplus\mathbb{Z}/2$; consider the injective module $I = \oplus(\mathbb{Q}/\mathbb{Z})$; then we have the exact sequence:

$$0 \longrightarrow B \longrightarrow I \xrightarrow{2} I \longrightarrow 0.$$

Applying the functor $\mathrm{Hom}(A,\cdot)$, we obtain the following diagram:

$$\begin{array}{ccccccc}
\mathrm{Hom}(A,I) & \xrightarrow{2} & \mathrm{Hom}(A,I) & \longrightarrow & \mathrm{Ext}(A,B) & \longrightarrow & 0 \\
\downarrow{\scriptstyle R} & & \downarrow{\scriptstyle R} & & \downarrow{\scriptstyle R} & & \\
\mathrm{Hom}(A_2,I) & \xrightarrow{2=0} & \mathrm{Hom}(A_2,I) & \longrightarrow & \mathrm{Ext}(A_2,B) & \longrightarrow & 0
\end{array}$$

where R denotes the morphism induced by the restriction to A_2.

On the other hand, we have the exact sequence:

$$0 \longrightarrow A_2 \longrightarrow A \xrightarrow{2} A$$

Applying the functor $\mathrm{Hom}(\cdot,I)$, we get an isomorphism between $\mathrm{Hom}(A_2,I)$ and $\mathrm{Hom}(A,I)/2$, so $R : \mathrm{Ext}(A,B) \longrightarrow \mathrm{Ext}(A_2,B)$ is bijective.

2.2.2 Equivalence of Categories

Theorem 2.11 *The functor $\mathcal{E}$ is an equivalence of categories between $\mathscr{M}_n$ and $\mathscr{D}_e$.*

To prove this theorem, we need the next two lemmas.

Lemma 2.12 *For each diagram D in $\mathscr{D}_e$, there exists a Moore space X in $\mathscr{M}_n$ such $\mathcal{E}(X) = D$.*

Proof Let D be an object of $\mathscr{D}_e$ given by:

$$A \underset{\beta}{\overset{\alpha}{\rightleftarrows}} B$$

Set X the Moore space $X = M(A,n)$. The diagram associated to X is given by:

$$A \underset{\beta_X}{\overset{\alpha_X}{\rightleftarrows}} A'$$

Then, we have the following diagram:

$$\begin{array}{ccccccccc} 0 & \longrightarrow & A/2 & \xrightarrow{\alpha} & B & \xrightarrow{\beta} & A_2 & \longrightarrow & 0 \\ & & \downarrow{\scriptstyle Id} & & & & \downarrow{\scriptstyle Id} & & \\ 0 & \longrightarrow & A/2 & \xrightarrow{\alpha_X} & A' & \xrightarrow{\beta_X} & A_2 & \longrightarrow & 0 \end{array}$$

where the lines are exact. Each horizontal exact sequence defines an element in $\mathrm{Ext}(A_2, A/2) \simeq \mathrm{Hom}(A_2, A/2)$. Since $\beta\alpha = 2$ on B and $\beta_X\alpha_X = 2$ on A', the two extensions give the same element in $\mathrm{Hom}(A_2, A/2)$ and then the two extensions are isomorphic.

Lemma 2.13 *If X and Y are two Moore spaces, then $[X,Y]$ is isomorphic to* $\mathrm{Hom}(D_X, D_Y) = \mathrm{Hom}(\mathcal{E}(X), \mathcal{E}(Y))$.

Proof Let $X = M(A,n)$ and $Y = M(B,n)$, then there is an exact sequence:

$$0 \longrightarrow \mathrm{Ext}(A,B) \longrightarrow [X,Y] \longrightarrow \mathrm{Hom}(A,B) \longrightarrow 0$$

But we have: $\mathrm{Ext}(A, B/2) \simeq \mathrm{Ext}(A_2, B/2) \simeq \mathrm{Hom}(A_2, B/2)$, so we obtain the exact sequence:

$$0 \longrightarrow \mathrm{Hom}(A_2, B/2) \longrightarrow [X,Y] \longrightarrow \mathrm{Hom}(A,B) \longrightarrow 0.$$

On the other side, the forgetful morphism $Fr : \mathrm{Hom}(D_X, D_Y) \longrightarrow \mathrm{Hom}(A,B)$ is surjective. Recall that an element $g \in \mathrm{Hom}(D_X, D_Y)$ is given by two maps g_1 and g_2 such that:

$$\begin{array}{ccccccccc} A & \xrightarrow{2} & A & \xrightarrow{\alpha_X} & A' & \xrightarrow{\beta_X} & A & \xrightarrow{2} & A \\ \downarrow{\scriptstyle g_1} & & \downarrow{\scriptstyle g_1} & & \downarrow{\scriptstyle g_2} & & \downarrow{\scriptstyle g_1} & & \downarrow{\scriptstyle g_1} \\ B & \xrightarrow{2} & B & \xrightarrow{\alpha_Y} & B' & \xrightarrow{\beta_Y} & B & \xrightarrow{2} & B \end{array}$$

so an element $g \in \mathrm{Hom}(D_X, D_Y)$ is in the kernel of the forgetful morphism if $g_1 = 0$ and then we obtain a morphism $A_2 \longrightarrow B/2$. Hence, we get the following commutative diagram:

$$\begin{array}{ccccccccc} 0 & \longrightarrow & \mathrm{Hom}(A_2, B/2) & \longrightarrow & [X\ ,\ Y] & \longrightarrow & \mathrm{Hom}(A, B) & \longrightarrow & 0 \\ & & \downarrow{\scriptstyle f_{X,Y}} & & \downarrow & & \downarrow{\scriptstyle Id} & & \\ 0 & \longrightarrow & \mathrm{Hom}(A_2, B/2) & \longrightarrow & \mathrm{Hom}(D_X, D_Y) & \xrightarrow{Fr} & \mathrm{Hom}(A, B) & \longrightarrow & 0 \end{array}$$

To prove the isomorphism between $[X,Y]$ and $\mathrm{Hom}(D_X, D_Y)$, it suffices to verify that $f_{X,Y}$ is the identity map. Notice that $f_{X,Y}$ is a bifunctor, covariant in B and contravariant in A.

When $X = P$ and $Y = S$, the diagram becomes:

$$\begin{array}{ccccccccc} 0 & \longrightarrow & \mathbb{Z}/2 & \longrightarrow & [P\ ,\ S] \simeq \mathbb{Z}/2 & \longrightarrow & 0 & \longrightarrow & 0 \\ & & \downarrow{\scriptstyle f_{P,S}} & & \downarrow & & & & \\ 0 & \longrightarrow & \mathbb{Z}/2 & \longrightarrow & \mathrm{Hom}(D_P, D_S) \simeq \mathbb{Z}/2 & \xrightarrow{Fr} & 0 & \longrightarrow & 0 \end{array}$$

so $f_{P,S}$ is necessarily the identity map. When $X = P$ and $Y = M(B,n)$: an element $y \in B$ defines a morphism $\mathbb{Z} \longrightarrow B$ that can be realized by a map between Moore spaces $S \longrightarrow Y$ and then a map $\bar{y} : \mathbb{Z}/2 \longrightarrow B/2$. By assigning $\bar{y}$ to the generator of $\mathbb{Z}/2$, we get the commutative diagram:

$$\begin{array}{ccc} \mathbb{Z}/2 & \longrightarrow & \mathrm{Hom}(\mathbb{Z}/2, B/2) \simeq B/2 \\ \downarrow{\scriptstyle f_{P,S}=Id} & & \downarrow{\scriptstyle f_{P,Y}} \\ \mathbb{Z}/2 & \longrightarrow & \mathrm{Hom}(\mathbb{Z}/2, B/2) \simeq B/2 \end{array}$$

Since $\mathrm{Hom}(\mathbb{Z}/2, B/2)$ is naturally isomorphic to $B/2$, even in this case $f_{P,Y} = Id$.

Given $x \in A_2$, it defines a map $\mathbb{Z}/2 \longrightarrow A_2 \subset A$ which can be realized by a map of Moore spaces $P \longrightarrow X$. This map allows to have the following commutative diagram, using the functoriality of $f_{X,Y}$:

$$\begin{array}{ccc} \mathrm{Hom}(A_2, B/2) & \longrightarrow & \mathrm{Hom}(\mathbb{Z}/2, B/2) \simeq B/2 \\ \downarrow{\scriptstyle f_{X,Y}} & & \downarrow{\scriptstyle f_{P,Y}=Id} \\ \mathrm{Hom}(A_2, B/2) & \longrightarrow & \mathrm{Hom}(\mathbb{Z}/2, B/2) \simeq B/2 \end{array}$$

the horizontal maps assign to a morphism $\varphi : A_2 \longrightarrow B/2$ its evaluation $\varphi(x) \in B/2$. To conclude that $f_{X,Y}$ is the identity map on $\mathrm{Hom}(A_2, B/2)$, it suffices to notice that the module A_2 is $\mathbb{Z}/2$-free, and if $\{u_i\}_{i \in I}$ is a basis of A_2 then $\mathrm{Hom}(A_2, B/2) \simeq$

$\prod \mathrm{Hom}(\mathbb{Z}/2, B/2) \simeq \prod B/2$. Using the evaluation on each generator u_i, we deduce the desired result.

Remark 2.14 With Lemmas 2.12 et 2.13, we get the proof of Theorem 2.11.

3 Stable Homotopy Groups of Moore Spaces

Let $X = M(A,n)$ and consider the Atiyah-Hirzebruch spectral sequence in homology with coefficients in the stable homotopy groups:

$$H_p(X; \pi_q^S) \Rightarrow \pi_{p+q}^S(X).$$

This spectral sequence contains just two non trivial columns and induces the following exact sequence:

$$0 \longrightarrow A \otimes \pi_q^S \xrightarrow{\nu^X} \pi_{n+q}^S(X) \xrightarrow{\mu^X} \mathrm{Tor}(A, \pi_{q-1}^S) \longrightarrow 0 \qquad (3.1)$$

Moreover, this exact sequence is natural in X.

Notice that, if $\underline{X}$ denotes the spectrum associated to the Moore space X, then $\pi_{n+i}^S(X) = \pi_i^S(\underline{X})$. In the following, the spectrum associated to a space X will also be denoted by X.

Recall the first stable homotopy groups (see [3]):

$$\pi_0^S = \mathbb{Z}, \pi_1^S = \mathbb{Z}/2, \pi_2^S = \mathbb{Z}/2, \pi_3^S = \mathbb{Z}/24, \pi_4^S = \pi_5^S = 0, \pi_6^S = \mathbb{Z}/2, \pi_7^S = \mathbb{Z}/240,$$

so the exact sequence (3.1) allows to obtain, for any Moore space X:

$$\pi_0^S(X) = A, \qquad \pi_1^S(X) \simeq A \otimes \mathbb{Z}/2 = A/2, \qquad \pi_4^S(X) \simeq \mathrm{Tor}(A, \mathbb{Z}/24) = A_{24},$$

$$\pi_5^S(X) = 0, \qquad \pi_6^S(X) \simeq A \otimes \mathbb{Z}/2 = A/2$$

but we can't determine explicitly $\pi_2^S(X)$, $\pi_3^S(X)$ and $\pi_7^S(X)$.

To compute $\pi_i^S(X)$, for $i = 2, 3, 7$, we need the following lemma:

Lemma 3.1 $\pi_2^S(P) = \mathbb{Z}/4$, $\pi_3^S(P) = \mathbb{Z}/2 \oplus \mathbb{Z}/2$, $\pi_7^S(P) = \mathbb{Z}/2 \oplus \mathbb{Z}/2$.

Proof These groups are given in [7], but we propose an easier proof of these results using the arguments of Sect. 2.

For $q = 2, 3, 7$, the exact sequence (3.1) becomes:

$$0 \longrightarrow \mathbb{Z}/2 \longrightarrow \pi_q^S(P) \longrightarrow \mathbb{Z}/2 \longrightarrow 0$$

then $\pi_q^S(P) \simeq \mathbb{Z}/2 \oplus \mathbb{Z}/2$ or $\pi_q^S(P) \simeq \mathbb{Z}/4$.

There is a cofibration sequence:

$$\xrightarrow{\theta} P \xrightarrow{\delta} S \xrightarrow{2} S \xrightarrow{\theta} P \xrightarrow{\delta} \tag{3.2}$$

where θ is of degree 0 and δ of degree -1. If λ denotes the composition of δ by the Hopf map from S to S, then we get the Moore spectra diagram

$$\begin{array}{ccc} S & \xrightarrow{2} & S \\ & {\scriptstyle\lambda}\nwarrow \quad \swarrow{\scriptstyle\theta} & \\ & P & \end{array} \tag{3.3}$$

verifying $2\theta = 0$, $2\lambda = 0$, $\lambda\theta = 0$ et $\theta\lambda = 2$.

Applying the functor π_2^S to (3.3), we obtain the following diagram:

$$\begin{array}{ccc} \mathbb{Z}/2 & \xrightarrow{2=0} & \mathbb{Z}/2 \\ & {\scriptstyle\lambda_*}\nwarrow \quad \swarrow{\scriptstyle\theta_*} & \\ & \pi_2^S(P) & \end{array}$$

where $2\lambda_* = 0$, $2\theta_* = 0$ and $\theta_*\lambda_* = 2$. This diagram is not necessarily exact, but, the exact sequence of stable homotopy groups applied to the cofibration (3.2) gives:

$$\ker(\theta_* : \mathbb{Z}/2 \longrightarrow \pi_2^S(P)) = \mathrm{Im}(2 : \mathbb{Z}/2 \longrightarrow \mathbb{Z}/2).$$

Then it suffices to find an element $u \in \pi_2^S(P)$ such that $\lambda_*(u) = 1$. For this purpose, we can choose $n = 2$ so $S = S^2$ and $P = P_2 = \Sigma\mathbb{R}P_2$. We have the cofibration:

$$S^2 \longrightarrow P_2 \longrightarrow S^3.$$

Applying the stable homotopy functor, we get:

$$\begin{array}{ccccccccccc}
\pi_4^S(S^2) & \longrightarrow & \pi_4^S(P_2) & \longrightarrow & \pi_4^S(S^3) & \longrightarrow & \pi_3^S(S^2) & \longrightarrow & \pi_3^S(P_2) & \longrightarrow & \pi_3^S(S^3) \\
\| & & \| & & \| & & \| & & \| & & \| \\
\pi_2^S & \longrightarrow & \pi_2^S(P) & \longrightarrow & \pi_1^S & \longrightarrow & \pi_1^S & \longrightarrow & \pi_1^S(P) & \longrightarrow & \pi_0^S \\
\| & & \| & & \| & & \| & & \| & & \| \\
\mathbb{Z}/2 & \longrightarrow & \pi_2^S(P) & \longrightarrow & \mathbb{Z}/2 & \xrightarrow{0} & \mathbb{Z}/2 & \xrightarrow{Id} & \mathbb{Z}/2 & \xrightarrow{0} & \mathbb{Z} \\
 & & & \searrow & \simeq \downarrow \mathrm{Hopf} & & & & & & \\
 & & & & \mathbb{Z}/2 & & & & & &
\end{array}$$

Then $\pi_2^S(P)$ is surjected on $\mathbb{Z}/2 = \pi_1^S$ which is sent by the Hopf map on $\mathbb{Z}/2 = \pi_2^S$ by assigning to the generator η of $\pi_1^S = \mathbb{Z}/2$ the generator η^2 of $\pi_2^S = \mathbb{Z}/2$.

Now, applying the functor π_3^S to (3.3), we get:

$$\begin{array}{ccc} \mathbb{Z}/24 & \xrightarrow{2} & \mathbb{Z}/24 \\ & {\scriptstyle\lambda_*}\nwarrow \quad \swarrow{\scriptstyle\theta_*} & \\ & \pi_3^S(P) & \end{array}$$

with $2\theta_* = 0$, $2\lambda_* = 0$ and $\theta_*\lambda_* = 2$. This diagram is not necessarily exact, but

$$\ker(\theta_* : \mathbb{Z}/24 \longrightarrow \pi_3^S(P)) = \mathrm{Im}(2 : \mathbb{Z}/24 \longrightarrow \mathbb{Z}/24).$$

Let $x \in \pi_3^S(P)$, then $2\lambda_*(x) = 0$. There exists $u \in \mathbb{Z}/24$ such that $\lambda_*(x) = 12u$. So $2x = \theta_*(\lambda_*(x)) = \theta_*(12u) = 0$ since $2\theta_* = 0$. This implies that elements of $\pi_3^S(P)$ vanish when multiplied by 2 and then $\pi_3^S(P) \simeq \mathbb{Z}/2 \oplus \mathbb{Z}/2$.

The same argument shows that $\pi_7^S(P) = \mathbb{Z}/2 \oplus \mathbb{Z}/2$.

Consider a Moore space $X = M(A,n)$. The next theorems compute $\pi_i^S(X)$, for $i = 2,3,7$, in terms of the modules $\Phi_1(D_X)$ and $\Phi_2(D_X)$.

Theorem 3.2 *For each generator $\gamma \in \pi_2^S(P)$, there is a natural isomorphism $\pi_2^S(X) \simeq \Phi_2(D_X) = [P,X]$.*

Proof Consider the exact sequence (2.2) and the exact sequence (3.1) for $q = 2$:

$$0 \longrightarrow A/2 \xrightarrow{\alpha} A' \xrightarrow{\beta} A_2 \longrightarrow 0$$

$$0 \longrightarrow A/2 \xrightarrow{\nu^X} \pi_2^S(X) \xrightarrow{\mu^X} A_2 \longrightarrow 0$$

We construct a map $A' \longrightarrow \pi_2^S(X)$ as follows: choose γ a generator of $\pi_2^S(P) \simeq \mathbb{Z}/4$. Let $u \in A'$ and consider f representing the class $u \in A' = [P,X]$.Then f induces a map $f_* : \pi_2^S(P) \longrightarrow \pi_2^S(X)$ and we define $\varphi_\gamma(u) = f_*(\gamma)$. the map φ_γ relies the two exact sequences:

$$\begin{array}{ccccccccc} 0 & \longrightarrow & A/2 & \xrightarrow{\alpha} & A' & \xrightarrow{\beta} & A_2 & \longrightarrow & 0 \\ & & & & \downarrow{\scriptstyle\varphi_\gamma} & & & & \\ 0 & \longrightarrow & A/2 & \xrightarrow{\nu^X} & \pi_2^S(X) & \xrightarrow{\mu^X} & A_2 & \longrightarrow & 0 \end{array}$$

Now, we may prove that the composite map

$$A' \xrightarrow{\varphi_\gamma} \pi_2^S(X) \xrightarrow{\mu^X} A_2$$

is $\beta : A' \longrightarrow A_2$. Using the functoriality, it suffices to prove the result when $X = P$. In this case, he diagram becomes:

$$\begin{array}{ccccccccc}
0 & \longrightarrow & \mathbb{Z}/2 & \xrightarrow{\alpha} & \mathbb{Z}/4 & \xrightarrow{\beta} & \mathbb{Z}/2 & \longrightarrow & 0 \\
 & & & & \downarrow{\scriptstyle \varphi_\gamma} & & \downarrow{\scriptstyle Id} & & \\
0 & \longrightarrow & \mathbb{Z}/2 & \xrightarrow{\nu^P} & \pi_2^S(P) \simeq \mathbb{Z}/4 & \xrightarrow{\mu^P} & \mathbb{Z}/2 & \longrightarrow & 0
\end{array}$$

where we see clearly that $\mu^P \circ \varphi_\gamma = \beta$.

By functoriality, for each Moore space $X = M(A,n)$ we get the following commutative diagram:

$$\begin{array}{ccccccccc}
0 & \longrightarrow & A/2 & \xrightarrow{\alpha} & A' & \xrightarrow{\beta} & A_2 & \longrightarrow & 0 \\
 & & \downarrow{\scriptstyle h} & & \downarrow{\scriptstyle \varphi_\gamma} & & \downarrow{\scriptstyle Id} & & \\
0 & \longrightarrow & A/2 & \xrightarrow{\nu^X} & \pi_2^S(X) & \xrightarrow{\mu^X} & A_2 & \longrightarrow & 0
\end{array}$$

here h is the natural map making the diagram commute. Notice that h is functorial in X. Then, to determine $h : A/2 \longrightarrow A/2$, it suffices to study the case $X = S$. In that case, the diagram becomes:

$$\begin{array}{ccccccccc}
0 & \longrightarrow & \mathbb{Z}/2 & \xrightarrow{\alpha} & \mathbb{Z}/2 & \xrightarrow{\beta} & 0 & \longrightarrow & 0 \\
 & & \downarrow{\scriptstyle h} & & \downarrow{\scriptstyle \varphi_\gamma} & & \downarrow{\scriptstyle Id} & & \\
0 & \longrightarrow & \mathbb{Z}/2 & \xrightarrow{\nu^S} & \pi_2^S = \mathbb{Z}/2 & \xrightarrow{\mu^S} & 0 & \longrightarrow & 0
\end{array}$$

and then h is necessarily the identity map.

Let $X = M(A,n)$ be a Moore space. Each element $x \in A$ defines a maps $f : \mathbb{Z} \longrightarrow A$ given by $f(1) = x$. This map is realized by a map between Moore spaces $f : S \longrightarrow X$ and induces, by naturality of h, the following commutative diagram:

$$\begin{array}{ccc}
\mathbb{Z}/2 & \xrightarrow{\bar{f}} & A/2 \\
{\scriptstyle h=Id}\downarrow & & \downarrow{\scriptstyle h} \\
\mathbb{Z}/2 & \xrightarrow{\bar{f}} & A/2
\end{array}$$

so $h : A/2 \longrightarrow A/2$ is still the identity map.

Remark 3.3 The isomorphism $\pi_2^S(X) \simeq A'$ depends on the choice of the generator $\gamma \in \pi_2^S(P) = \mathbb{Z}/4$. Choosing the generator $-\gamma$ multiplies the isomorphisme by -1.

Theorem 3.4 *For each $\gamma \in \pi_3^S(P)$ such that $\mu^P(\gamma) = 1 \in \mathbb{Z}/2$, there is a natural isomorphism $\pi_3^S(X) \simeq A' \oplus_{A/2} A/24$ obtained by the pushout*

$$\begin{array}{ccc} A/2 & \xrightarrow{\alpha} & A' \\ {\scriptstyle \times 12}\downarrow & & \downarrow \\ A/24 & \longrightarrow & \pi_3^S(X) \end{array}$$

where $A = \Phi_1(D_X)$ and $A' = \Phi_2(D_X)$.

Proof When $q = 3$, the exact sequence (3.1) becomes:

$$0 \longrightarrow A/24 \xrightarrow{\nu^X} \pi_3^S(X) \xrightarrow{\mu^X} A_2 \longrightarrow 0$$

For $X = P$ we get

$$0 \longrightarrow \mathbb{Z}/2 \xrightarrow{\nu^P} \pi_3^S(P) \simeq \mathbb{Z}/2 \oplus \mathbb{Z}/2 \xrightarrow{\mu^P} \mathbb{Z}/2 \longrightarrow 0$$

Choose $\gamma \in \pi_3^S(P)$ such that $\mu^P(\gamma)$ is the generator of $\mathbb{Z}/2$. We construct a map $\varphi_\gamma : A' \longrightarrow \pi_3^S(X)$ as follows:

Let $u \in A' = [P, X]$ and let $f : P \longrightarrow X$ representing the class u. Then $\varphi_\gamma(u) = f_*(\gamma)$.

As in the proof of Theorem 3.2 we show that the composition of $\mu^X : \pi_3^S(X) \longrightarrow A_2$ by φ_γ is $\beta : A' \longrightarrow A_2$.

We obtain the following commutative diagram:

$$\begin{array}{ccccccccc} 0 & \longrightarrow & A/2 & \xrightarrow{\alpha} & A' & \xrightarrow{\beta} & A_2 & \longrightarrow & 0 \\ & & {\scriptstyle h}\downarrow & & \downarrow{\scriptstyle \varphi_\gamma} & & \downarrow{\scriptstyle Id} & & \\ 0 & \longrightarrow & A/24 & \xrightarrow{\nu^X} & \pi_3^S(X) & \xrightarrow{\mu^X} & A_2 & \longrightarrow & 0 \end{array}$$

Since h is natural, we need just to determine it for $X = S$. In that case, the map $h : \mathbb{Z}/2 \longrightarrow \mathbb{Z}/24$ assigns to the generator of $\mathbb{Z}/2$ an element of $\mathbb{Z}/24$ vanishing when multiplied by 2, that means 0 or 12. Then $h = 0$ or $h = \times 12$. To prove that $h = \times 12$, we consider the cofibration

$$S \xrightarrow{2} S \longrightarrow P$$

which induces the long exact sequence:

$$\cdots \longrightarrow \pi_n^S \xrightarrow{2} \pi_n^S \longrightarrow \pi_n^S(P) \longrightarrow \pi_{n-1}^S \xrightarrow{2} \pi_{n-1}^S \longrightarrow \cdots \quad (3.4)$$

For $n = 3$, we have:

$$\pi_3^S = \mathbb{Z}/24 \xrightarrow{2} \pi_3^S = \mathbb{Z}/24 \longrightarrow \pi_3^S(P) \longrightarrow \pi_2^S = \mathbb{Z}/2 \xrightarrow{2=0} \pi_2^S = \mathbb{Z}/2$$

This proves that $\pi_3^S(P) \longrightarrow \pi_2^S$ is surjective, so every map $S \longrightarrow S$ of degree 2 can be lifted to a map $S \longrightarrow P$ of degree 3.

If $\gamma \in \pi_3^S(P)$ is represented by a map, denoted also $\gamma : S^5 \longrightarrow P_2$, and since $\mu^P(\gamma) = 1 \in \mathbb{Z}/2 = \mathrm{Tor}(\mathbb{Z}/2, \pi_2^S)$, then the map $\pi_3^S(P) \longrightarrow \pi_2^S$ takes γ to the generator $1_{\mathbb{Z}/2} \in \pi_2^S = \mathbb{Z}/2$.

Let

$$u : P_2 \xrightarrow{\delta_2} S^3 \xrightarrow{Hopf} S^2$$

be a representative of the nonzero element of $[P, S] = \mathbb{Z}/2$ and $a : S^5 \longrightarrow S^3$ a representative of the generator of π_2^S. Then $\varphi_\gamma([u]) = u_*(\gamma) = \eta \times (\delta_2)_*(\gamma) = \eta \times [a]$ where η denotes the multiplication by the class of the Hopf map. But the multiplication by the Hopf map class takes the generator of π_2^S to product by 12 of the generator of π_3^S (see [3]). This allows to deduce that $\varphi_\gamma([u]) = 12 \in \mathbb{Z}/24$ and that h is multiplication by 12.

Remark 3.5 The isomorphism $\pi_3^S(X) \simeq A' \oplus_{A/2} A/24$ depends on the choice of $\gamma \in \pi_3^S(P) \simeq \mathbb{Z}/2 \oplus \mathbb{Z}/2$ verifying $\mu^P(\gamma) = 1$. There are two possible choices.

If we choose γ' such that $\mu^P(\gamma') = 1$, then $\nu^P(1_{\mathbb{Z}/2}) = \gamma - \gamma'$. We can show that

$$\varphi_{\gamma'} = \varphi_\gamma + \tilde{\lambda} \circ \beta$$

where $\tilde{\lambda} : A_2 \longrightarrow \pi_3^S(X)$ is defined as follows: if $a \in A_2$, we can represent it by a map $a : S \longrightarrow X$ such that $2a = 0$. This map induces $a_* : \pi_3^S \longrightarrow \pi_3^S(X)$ taking all generators of π_3^S to the same element $a_*(1_{\mathbb{Z}/24}) \in \pi_3^S(X)$ since $2a_* = 0$. Then we define $\tilde{\lambda}$ by $\tilde{\lambda}(a) = a_*(1_{\mathbb{Z}/24})$.

Theorem 3.6 *For each $\gamma \in \pi_7^S(P)$ such that $\mu^P(\gamma) = 1 \in \mathbb{Z}/2$, there is a natural isomorphism $\pi_7^S(X) \simeq A/240 \oplus A_2$, where $A = \Phi_1(D_X)$.*

Proof Using the same construction of the case of $\pi_3^S(X)$, we get the following commutative diagram:

$$\begin{array}{ccccccccc}
0 & \longrightarrow & A/2 & \xrightarrow{\alpha} & A' & \xrightarrow{\beta} & A_2 & \longrightarrow & 0 \\
 & & \downarrow{\scriptstyle h} & & \downarrow{\scriptstyle \varphi_\gamma} & & \downarrow{\scriptstyle Id} & & \\
0 & \longrightarrow & A/240 & \xrightarrow{\nu^X} & \pi_7^S(X) & \xrightarrow{\mu^X} & A_2 & \longrightarrow & 0
\end{array}$$

To determine h, it suffices to consider the case of $X = S$, since it is natural on X. In that case $h : \mathbb{Z}/2 \longrightarrow \mathbb{Z}/240$ is the multiplication by 0 or 120.

For $n = 7$, the long exact sequence (3.4) becomes:

$$\pi_7^S = \mathbb{Z}/240 \xrightarrow{2} \pi_7^S = \mathbb{Z}/240 \longrightarrow \pi_7^S(P) \longrightarrow \pi_6^S = \mathbb{Z}/2 \xrightarrow{2=0} \pi_6^S = \mathbb{Z}/2$$

showing that $\pi_7^S(P) \longrightarrow \pi_6^S$ is surjective. We use the same techniques of the previous theorem proof, and the fact that the product by the Hopf class on π_6^S is zero (see [3]), we deduce that $h = 0$

Remark 3.7 Using the new universal coefficient exact sequence of [4], we can represent the functor π_i^S on $\mathscr{M}_n$ as a tensor product by particular objects of an abelian category $\mathscr{D}$ containing $\mathscr{D}_e$.

References

1. Baues, H.-J.: Homotopy type and homology. Oxford Mathematical Monographs, Clarendon Press (1996)
2. Hatcher, A.: Algebraic Topol. Cambridge University Press, Cambridge (2000)
3. Kochman, S.O.: Stable homotopy groups of spheres. Lecture Notes in Math. no. 1423 (1990)
4. Saihi, I.: Homologies généralisées à coefficients. C.R. Acad. Sci. Paris, Ser. I **353**, 397–401 (2015)
5. Switzer, R.M.: Algebraic Topol.-Homol. Homotopy. Springer-Verlag, Berlin (1975)
6. Vogel, P.: A solution of the Steenrod problem for G-Moore spaces. K-theory **1**(4), 325–335 (1987)
7. Wu, J.: Homotopy theory of the suspensions of the projective plane. (2003)- books.google.com

Notes on Quasi-Cyclic Codes with Cyclic Constituent Codes

Minjia Shi, Yiping Zhang and Patrick Solé

Abstract Quasi-cyclic codes are generalizations of the familiar linear cyclic codes. By using the results of [4], the authors in [2, 3] showed that a quasi-cyclic code $\mathscr{C}$ over $\mathbb{F}_q$ of length ℓm and index ℓ with m being pairwise coprime to ℓ and the characteristic of $\mathbb{F}_q$ is equivalent to a cyclic code if the constituent codes of $\mathscr{C}$ are cyclic, where q is a prime power and the equivalence is given in [3]. In this paper, we apply an algebraic method to prove that a quasi-cyclic code with cyclic constituent codes is equivalent to a cyclic code. Moreover, the main result (see Theorem 4) includes Proposition 9 in [3] as a special case.

Keywords Quasi-cyclic codes · Constituent codes · Cyclic codes · Circulant matrix · Similar circulant matrix

MSC 2010 codes: Primary 94B05 · Secondary 94B15

M. Shi (✉)
Key Laboratory of Intelligent Computing & Signal Processing,
Ministry of Education, Anhui University, No. 3 Feixi Road, Hefei 230039,
Anhui, People's Republic of China
e-mail: smjwcl.good@163.com

M. Shi
National Mobile Communications Research Laboratory, Southeast University,
Nanjing 210096, People's Republic of China

M. Shi
School of Mathematical Sciences of Anhui University, Hefei 230601, Anhui,
People's Republic of China

Y. Zhang
School of Wendian, Anhui University, Hefei 230601, Anhui, China
e-mail: yipingzhang0123@163.com

P. Solé
CNRS/LAGA, University of Paris 8, 93 526 Saint-Denis, France
e-mail: sole@enst.fr

T. Abualrub et al. (eds.), *Mathematics Across Contemporary Sciences*,
Springer Proceedings in Mathematics & Statistics 190,
DOI 10.1007/978-3-319-46310-0_12

1 Introduction

Quasi-cyclic codes over finite fields form an important class of block codes that include cyclic codes as a special case. In [4], Ling and Solé viewed each quasi-cyclic code as a code over a polynomial ring, and extracted a description of each quasi-cyclic code as being constructed from linear codes of shorter lengths over larger fields, which are called the constituent codes of the quasi-cyclic code. It is interesting to ask what kind of codes we will obtain if constituent codes of a quasi-cyclic code are cyclic. Such codes can enjoy the ease of encoding of cyclic codes by polynomial division for instance.

In [1], quasi-cyclic codes of length 5ℓ and index ℓ over $\mathbb{F}_q$ were obtained from a pair of codes over $\mathbb{F}_q$ and $\mathbb{F}_{q^4}$, respectively, by a combinatorial construction called here the quintic construction. They enjoy a designed trellis description and a suboptimal coset decoding algorithm. They are shown to be cyclic when the constituent codes are cyclic of odd length coprime to 5. Lim [3] generalized the result in [1] to the general case by a similar method. In [2], Güneri and Özbudak considered the same issue. If the constituent codes of a quasi-cyclic code $\mathscr{C}$ of length $m\ell$ and index ℓ are cyclic, the authors show that $\mathscr{C}$ can be viewed as a 2-D cyclic code of size $m \times \ell$ over $\mathbb{F}_q$. Moreover, in case m and ℓ are also coprime to each other, $\mathscr{C}$ must be equivalent to a cyclic code. However, the results of Refs. [2], [3] relied on the structures of quasi-cyclic codes of the Ref. [4].

In this paper, we apply an algebraic method to investigate the same issue. Moreover, the equivalence in Proposition 9 of [3] is a special case of Theorem 4, which provides many equivalences. Throughout this paper we require that $(m, q) = (\ell, q) = (m, \ell) = 1$, where $q = p^k$ for some positive integer k, p is a prime.

2 The Circulant Matrix Decomposition of a Cyclic Code

Cyclic codes are generated by shift registers and play an important role in random error-correcting and burst error-correcting. Cyclic codes were first studied by Prange in 1957, and the study of the algebraic properties of cyclic codes developed rapidly since then. An $[n, k]_q$ code C is called cyclic provided that, for each codeword $\mathbf{c} = (c_0, c_1, c_2, \ldots, c_{n-1}) \in C$, the vector $(c_{n-1}, c_0, c_1, \ldots, c_{n-2}) \in C$. In this section, we require that $(n, p) = 1$.

Definition 1 Let C be a cyclic code of length n over $\mathbb{F}_q$ and $A \subseteq C$, then a *circulant matrix* A containing the codeword $(a_0, a_1, \ldots, a_{n-1})$ is defined as follows

$$A = \begin{pmatrix} a_0 & a_1 & a_2 & \ldots & a_{n-1} \\ a_{n-1} & a_0 & a_1 & \ldots & a_{n-2} \\ \ldots & & \ldots & & \ldots \\ a_1 & a_2 & a_3 & \ldots & a_0 \end{pmatrix}.$$

Remark 1 A can be considered as a set of n codewords of C. In our case, codeword repetition in A is omitted if necessary.

Lemma 1 *A cyclic code C of length n over $\mathbb{F}_q$ can be decomposed into a finite disjoint union of circulant matrices.*

Proof If $\mathbf{c} = (a_0, a_1, \ldots, a_{n-1}) \in C$, then we have $A \subseteq C$. For any $\mathbf{c}' = (b_0, b_1, \ldots, b_{n-1}) \in C$ and $\mathbf{c}' \notin A$, following the construction of the circulant matrix, then $A \cap B = \emptyset$, where B is the circulant matrix containing $\mathbf{c}'$, this operation will be stopped after finite steps.

Take the $[7, 4, 3]$ Hamming code C for example, which is a cyclic code with generator polynomial $1 + x^2 + x^3$, according to Lemma 1, we have $C =$

$$\begin{pmatrix} 1&0&1&1&1&0&0\\ 0&1&0&1&1&1&0\\ 0&0&1&0&1&1&1\\ 1&0&0&1&0&1&1\\ 1&1&0&0&1&0&1\\ 1&1&1&0&0&1&0\\ 0&1&1&1&0&0&1 \end{pmatrix} \bigcup \begin{pmatrix} 1&1&0&1&0&0&0\\ 0&1&1&0&1&0&0\\ 0&0&1&1&0&1&0\\ 0&0&0&1&1&0&1\\ 1&0&0&0&1&1&0\\ 0&1&0&0&0&1&1\\ 1&0&1&0&0&0&1 \end{pmatrix} \bigcup \begin{pmatrix} 0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0 \end{pmatrix} \bigcup \begin{pmatrix} 1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1\\ 1&1&1&1&1&1&1 \end{pmatrix}.$$

Following Definition 1, we can prove the following lemma, which plays an important role in obtaining our results.

Lemma 2 *Let C be a cyclic code of length n over $\mathbb{F}_q$, then A is a circulant matrix if and only if $A = P_n diag(f(1), f(\zeta), \ldots, f(\zeta^{n-1}))P_n^{-1}$, where*

$$P_n = \begin{pmatrix} 1 & 1 & 1 & \ldots & 1\\ 1 & \zeta & \zeta^2 & \ldots & \zeta^{n-1}\\ 1 & \zeta^2 & \zeta^{2\times 2} & \ldots & \zeta^{2(n-1)}\\ \vdots & & \vdots & & \vdots\\ 1 & \zeta^{n-1} & \zeta^{2(n-1)} & \ldots & \zeta^{(n-1)(n-1)} \end{pmatrix}$$

is a Vandermonde matrix, ζ is a primitive n-th root of unity, $(a_0, \cdots, a_{n-1})$ is the first row of A and $f(x) = a_0 + a_1x + a_2x^2 + \cdots + a_{n-1}x^{n-1}$.

Proof It is clear that P_n is invertible since ζ is a primitive n-th root of unity. Moreover, it is easy to check that

$$AP_n = \begin{pmatrix} f(1) & f(\zeta) & \ldots & f(\zeta^{n-1})\\ f(1) & \zeta f(\zeta) & \ldots & \zeta^{n-1}f(\zeta^{n-1})\\ \ldots & \ldots & \ldots & \ldots\\ f(1) & \zeta^{n-1}f(\zeta) & \ldots & \zeta^{(n-1)(n-1)}f(\zeta^{n-1}) \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 1 & \ldots & 1\\ 1 & \zeta & \ldots & \zeta^{n-1}\\ \ldots & \ldots & \ldots & \ldots\\ 1 & \zeta^{n-1} & \ldots & \zeta^{(n-1)(n-1)} \end{pmatrix} diag(f(1), f(\zeta), \ldots, f(\zeta^{n-1})).$$

Equivalently, $A = P_n diag(f(1), f(\zeta), \ldots, f(\zeta^{n-1}))P_n^{-1}$. The converse part is straightforward.

3 Quasi-cyclic Codes with Cyclic Constituent Codes

A linear code $\mathscr{C}$ is a quasi-cyclic code of length ℓm with index ℓ if $\mathscr{C}$ is invariant under a shift by ℓ places, namely, for any $(a_{00}, a_{01}, \ldots, a_{0,\ell-1}, a_{10}, \ldots, a_{1,\ell-1}, \ldots, a_{m-1,0}, \ldots, a_{m-1,\ell-1}) \in \mathscr{C}$, we have $(a_{m-1,0}, a_{m-1,1}, \ldots, a_{m-1,\ell-1}, a_{00}, \ldots, a_{0,\ell-1}, \ldots, a_{m-2,0}, \ldots, a_{m-2,\ell-1}) \in \mathscr{C}$. The *constituent codes* of such a code are codes of length ℓ over extension alphabets that appear in the CRT decomposition of [4]. See [4] for details. They are not cyclic in general. The class of quasi-cyclic codes with cyclic constituents is a strict subclass of all quasi-codes. In [2], the authors proved that if m and ℓ are both relatively prime to q, and the constituents of the quasi-cyclic code (of length ℓm and index ℓ) are all cyclic codes, then $\mathscr{C}$ is a 2-D cyclic code. Therefore, a linear code $\mathscr{C}$ of length ℓm is a quasi-cyclic code of length ℓm and index ℓ with cyclic constituent codes if $(a_{00}, a_{01}, a_{02}, \ldots, a_{0,\ell-1}, a_{10}, \ldots, a_{1,\ell-1}, \ldots, a_{m-1,0}, \ldots, a_{m-1,\ell-1}) \in \mathscr{C}$ implies that

$$(a_{m-1,\ell-1}, a_{m-1,0}, \ldots, a_{m-1,\ell-2}, a_{0,\ell-1}, \ldots, a_{0,\ell-2}, \ldots, a_{m-2,\ell-1}, \ldots, a_{m-2,\ell-2}) \in \mathscr{C}.$$

Definition 2 Let $\mathscr{C}$ be a quasi-cyclic code of length ℓm and index ℓ with cyclic constituent codes, then a *similar circulant matrix* A' containing the codeword

$$(a_{00}, a_{01}, \ldots, a_{0,\ell-1}, a_{10}, \ldots, a_{1,\ell-1}, \ldots, a_{m-1,0}, \ldots, a_{m-1,\ell-1})$$

is defined as follows

$$\begin{pmatrix}
a_{00} & a_{01} & \ldots & a_{0,\ell-1} & a_{10} & \ldots & a_{1,\ell-1} & \ldots & a_{m-1,0} & \ldots & a_{m-1,\ell-1} \\
a_{m-1,\ell-1} & a_{m-1,0} & \ldots & a_{m-1,\ell-2} & a_{0,\ell-1} & \ldots & a_{0,\ell-2} & \ldots & a_{m-2,\ell-1} & \ldots & a_{m-2,\ell-2} \\
a_{m-2,\ell-2} & a_{m-2,\ell-1} & \ldots & a_{m-2,\ell-3} & a_{m-1,\ell-2} & \ldots & a_{m-1,\ell-3} & \ldots & a_{m-3,\ell-2} & \ldots & a_{m-3,\ell-3} \\
\ldots & & \ldots & & & \ldots & & & \ldots & & \\
a_{11} & a_{12} & \ldots & a_{10} & a_{21} & \ldots & a_{20} & \ldots & a_{01} & \ldots & a_{00}
\end{pmatrix}.$$

Remark 2 A' can be considered as a set of ℓm codewords of $\mathscr{C}$. Codeword repetition in A' is omitted if necessary. Note that A' is a $\ell m \times \ell m$ matrix.

Similar to the proof of Lemma 1, we have the following corollary.

Corollary 1 *Let $\mathscr{C}$ be a quasi-cyclic code of length ℓm and index ℓ with cyclic constituent codes, then the code $\mathscr{C}$ can be decomposed into finite disjoint unions of similar circulant matrices.*

We denote by S_n the symmetric group of n elements. The following lemma will be clear from matrix theory.

Lemma 3 *Let D_1 and D_2 be $n \times n$ matrices, for $\sigma \in S_n$, $\sigma(D_1)$ represents the action of σ on coordinates of every row of D_1, $\sigma^T(D_1)$ represents the action of σ on coordinates of every column of D_1, which means if*

$$D_1 = \begin{pmatrix} d_{00} & d_{01} & d_{02} & \dots & d_{0,n-1} \\ d_{10} & d_{11} & d_{12} & \dots & d_{1,n-1} \\ \dots & & \dots & & \dots \\ d_{n-1,0} & d_{n-1,1} & d_{n-1,2} & \dots & d_{n-1,n-1} \end{pmatrix},$$

then we have

$$\sigma(D_1) = \begin{pmatrix} d_{0,\sigma(0)} & d_{0,\sigma(1)} & d_{0,\sigma(2)} & \dots & d_{0,\sigma(n-1)} \\ d_{1,\sigma(0)} & d_{1,\sigma(1)} & d_{1,\sigma(2)} & \dots & d_{1,\sigma(n-1)} \\ \dots & & \dots & & \dots \\ d_{n-1,\sigma(0)} & d_{n-1,\sigma(1)} & d_{n-1,\sigma(2)} & \dots & d_{n-1,\sigma(n-1)} \end{pmatrix},$$

$$\sigma^T(D_1) = \begin{pmatrix} d_{\sigma(0),0} & d_{\sigma(0),1} & d_{\sigma(0),2} & \dots & d_{\sigma(0),n-1} \\ d_{\sigma(1),0} & d_{\sigma(1),1} & d_{\sigma(1),2} & \dots & d_{\sigma(1),n-1} \\ \dots & & \dots & & \dots \\ d_{\sigma(n-1),0} & d_{\sigma(n-1),1} & d_{\sigma(n-1),2} & \dots & d_{\sigma(n-1),n-1} \end{pmatrix}$$

and $D_1 D_2 = \sigma(D_1)\sigma^T(D_2)$.

Lemma 4 *Let ε be a primitive ℓm-th root of unity, then there exists a permutation $\theta \in S_{\ell m}$ such that $\theta(A') = P_{\ell m} \Lambda P_{\ell m}^{-1}$, where*

$$P_{\ell m} = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & \varepsilon & \varepsilon^2 & \dots & \varepsilon^{\ell m-1} \\ 1 & \varepsilon^2 & \varepsilon^{2\times 2} & \dots & \varepsilon^{2(\ell m-1)} \\ \vdots & & \vdots & & \vdots \\ 1 & \varepsilon^{\ell m-1} & \varepsilon^{2(\ell m-1)} & \dots & \varepsilon^{(\ell m-1)(\ell m-1)} \end{pmatrix}$$

is a Vandermonde matrix, $\Lambda = diag(g(1), g(\varepsilon), g(\varepsilon^2), \dots, g(\varepsilon^{\ell m-1}))$ is a diagonal matrix, and $g(y) = a_{00} + a_{11}y + \cdots + a_{i_m,i_\ell} y^i + \cdots + a_{m-1,\ell-1} y^{\ell m-1}$ with $i_m = i \pmod m$, $i_\ell = i \pmod \ell$, $i = 0, 1, 2, \dots, \ell m - 1$.

Proof Let $\xi \in \{1, \varepsilon, \varepsilon^2, \dots, \varepsilon^{\ell m-1}\}$ and $P'_{\ell m}$ be obtained from the matrix $P_{\ell m}$ under certain row shift, then there exists a permutation θ such that $\theta^T(P'_{\ell m}) = P_{\ell m}$. Since $gcd(\ell, m) = 1$, according to the Chinese Remainder Theorem, we can establish a one-to-one correspondence between the coefficient of the term ξ^i in $g(\xi)$ and ξ^i denoted by $a_{i_m,i_\ell} \leftrightarrow \xi^i$, this correspondence can make the calculation of $g(y)$ easily. Let $P'_{\ell m}(\xi)$ be any column vector of $P'_{\ell m}$, and $A' P'_{\ell m}(\xi) = (b_0, b_1, \dots, b_{\ell m-1})^T$. Set $b_0 = g(\xi)$, by this correspondence and the elements of the first row of A', we can determine $P'_{\ell m}(\xi) = (1, \xi^{tm}, \xi^{2tm}, \dots, \xi^i, \dots, \xi^{\ell m-1})^T$, where t is the multiplicative inverse of m module ℓ. Thus θ is determined by $P'_{\ell m}(\xi)$. The elements of the j-th

row of A' can be expressed as

$$(a_{00}^{(j)}, a_{01}^{(j)}, \ldots, a_{0,\ell-1}^{(j)}, a_{10}^{(j)}, a_{11}^{(j)}, \ldots, a_{1,\ell-1}^{(j)}, \ldots, a_{m-1,0}^{(j)}, a_{m-1,1}^{(j)}, \ldots, a_{m-1,\ell-1}^{(j)}),$$

where $1 \le j \le \ell m$.

Next, we try to calculate b_j ($j = 1, 2, \ldots, \ell m - 1$). If we fix j, by the construction of the similar circulant matrix A', since $1 \le i + j \le 2\ell m - 2$, we know that in the $(j+1)$-th row of A',

$$a_{i_m,i_\ell}^{(1)} = a_{(i+j)_m,(i+j)_\ell}^{(j+1)} \leftrightarrow \xi^{(i+j)_{\ell m}},$$

and $\xi^{(i+j)_{\ell m}} = \xi^{i+j}$ for $\xi^{\ell m} = 1$. Then

$$b_j = \sum_{i=0}^{\ell m-1} a_{i_m,i_\ell}^{(j+1)} \xi^i = \sum_{i+j=0}^{i+j=\ell m-1} a_{(i+j)_m,(i+j)_\ell}^{(j+1)} \xi^{i+j} = \xi^j \sum_{i+j=0}^{i+j=\ell m-1} a_{(i+j)_m,(i+j)_\ell}^{(j+1)} \xi^i$$

$$= \xi^j \sum_{i+j=0}^{i+j=\ell m-1} a_{i_m,i_\ell}^{(1)} \xi^i = \xi^j \sum_{i=0}^{\ell m-1} a_{i_m,i_\ell}^{(1)} \xi^i = \xi^j b_0. \tag{1}$$

From (1), we have

$$A' P'_{\ell m}(\xi) = (b_0, b_1, \ldots, b_{\ell m-1})^T = g(\xi)(1, \xi, \xi^2, \ldots, \xi^{\ell m-1})^T. \tag{2}$$

Set $\xi = 1, \varepsilon, \varepsilon^2, \ldots, \varepsilon^{\ell m-1}$, from (2), we have

$$A'(P'_{\ell m}(1), P'_{\ell m}(\varepsilon), P'_{\ell m}(\varepsilon^2), \ldots, P'_{\ell m}(\varepsilon^{\ell m-1}))^T = A' P'_{\ell m},$$

then

$$A' P'_{\ell m} = \begin{pmatrix} g(1) & g(\varepsilon) & \ldots & g(\varepsilon^{\ell m-1}) \\ g(1) & \varepsilon g(\varepsilon) & \ldots & \varepsilon^{\ell m-1} g(\varepsilon^{\ell m-1}) \\ \ldots & & \ldots & \ldots \\ g(1) & \varepsilon^{\ell m-1} g(\varepsilon) & \ldots & \varepsilon^{(\ell m-1)(\ell m-1)} g(\varepsilon^{\ell m-1}) \end{pmatrix} = P_{\ell m} \Lambda. \tag{3}$$

Thus $A' P'_{\ell m} = P_{\ell m} \Lambda$. From Lemma 3, we have $A' P'_{\ell m} = \theta(A')\theta^T(P'_{\ell m}) = \theta(A') P_{\ell m} = P_{\ell m} \Lambda$. Consequently, $\theta(A') = P_{\ell m} \Lambda P_{\ell m}^{-1}$.

Corollary 2 *A similar circulant matrix A' is equivalent to a circulant matrix.*

Proof From Lemmas 4 and 2, we know that $\theta(A')$ is a circulant matrix, so A' is equivalent to a circulant matrix $\theta(A')$. Moreover, from the expressions of $f(x)$ and $g(y)$, the circulant matrix $\theta(A')$ is none other than the circulant matrix containing the codeword $(a_{00}, a_{11}, \ldots, a_{i_m,i_\ell}, \ldots, a_{m-1,\ell-1})$.

Theorem 1 *A quasi-cyclic code $\mathscr{C}$ of length ℓm and index ℓ with cyclic constituent codes is equivalent to a cyclic code.*

Proof From Corollary 1, we can write $\mathscr{C} = A'_1 \cup A'_2 \cup \cdots \cup A'_k = \cup_{i=1}^{k} A'_i$, from Lemma 4, let θ be a permutation that $\theta(A'_1)$ is a circulant matrix, and according to the proof of Lemma 4, the permutation θ is universally applicable for the matrices A'_i, thus $\theta(A'_i)(i = 1, \ldots, k)$ are all circulant matrices. Now we prove that $\theta(\mathscr{C})$ is a linear cyclic code. For $\theta(\mathbf{c}) \in \theta(\mathscr{C})$, then there exists i such that $\theta(\mathbf{c}) \in \theta(A'_i)$, from the construction of the circulant matrix, then $\theta(\mathscr{C})$ is cyclic. The linearity of $\theta(\mathscr{C})$ is obtained by the linearity of $\mathscr{C}$. In more details, for $\theta(\mathbf{c}), \theta(\mathbf{c}') \in \theta(\mathscr{C})$, there exist $\mathbf{c}, \mathbf{c}' \in \mathscr{C}$, in such a way that, for $k_1, k_2 \in \mathbb{F}_p$, $k_1\mathbf{c} + k_2\mathbf{c}' \in \mathscr{C}$ we have $\theta(k_1\mathbf{c} + k_2\mathbf{c}') = k_1\theta(\mathbf{c}) + k_2\theta(\mathbf{c}') \in \theta(\mathscr{C})$. Therefore, $\theta(\mathscr{C})$ is a linear cyclic code and $\mathscr{C}$ is equivalent to a cyclic code $\theta(\mathscr{C})$.

Theorem 1 in fact gives an alternative proof of Proposition 9 in [3] by a different method.

Lemma 5 (See Proposition 9 in [3]) *Let q be a prime power, and let $\mathbb{F}_q$ denote a finite field. Let ℓ and m be coprime positive integers with m coprime to q, and let $\mathscr{C}$ be a quasi-cyclic code of length ℓm and index ℓ with cyclic constituent codes over $\mathbb{F}_q$, let t denote the multiplicative inverse of m module ℓ, then $\mathscr{C}$ is equivalent to a cyclic code C, the equivalence is given by $\mathbf{d} = (d_0, d_1, \ldots, d_{\ell m-1}) \in C$, its pre-image $\mathbf{c}$ in $\mathscr{C}$ is given by*

$$(d_{(0)tm+0}, d_{tm+0}, d_{2tm+0}, \ldots, d_{(\ell-1)tm+0}, d_{(\ell-1)tm+1}, d_{(0)tm+1}, d_{tm+1}, \ldots, d_{(\ell-2)tm+1},$$

$$\ldots, d_{(\ell-m+1)tm+(m-1)}, d_{(\ell-m+2)tm+(m-1)}, d_{(\ell-m+3)tm+(m-1)}, \ldots, d_{(\ell-m)tm+(m-1)}).$$

Theorem 2 *The results of Theorem 1 are equivalent to those of Lemma 5.*

Proof According to Corollary 2, the codeword

$$(a_{00}, \ldots, a_{0,\ell-1}, a_{10}, \ldots, a_{1,\ell-1}, \ldots, a_{m-1,0}, \ldots, a_{m-1,\ell-1}) \in \mathscr{C}$$

is equivalent to the codeword $(a_{00}, a_{11}, \ldots, a_{i_m,i_\ell}, \ldots, a_{m-1,\ell-1}) \in \theta(\mathscr{C})$. Let

$$(a_{00}, a_{11}, \ldots, a_{i_m,i_\ell}, \ldots, a_{m-1,\ell-1}) = (y_0, y_1, y_2, \ldots, y_i, \ldots, y_{\ell m-1}),$$

in such a way that $a_{i_m,i_\ell} = y_i$, where $0 \le i \le \ell m - 1$. For any $a_{i,j}$, write

$$k_m = i, k_\ell = j \Leftrightarrow k \equiv i \pmod{m}, k \equiv j \pmod{\ell}. \tag{4}$$

Note that $mt = 1 \pmod{\ell}$, and $0 \le k \le \ell m - 1$, it is easy to check that $k = (j - i)_\ell mt + i$ is a solution of the congruence Eq. (4). Therefore

$$(a_{00}, a_{01}, a_{02}, \ldots, a_{0,\ell-1}, a_{10}, \ldots, a_{1,\ell-1}, \ldots, a_{m-1,0}, \ldots, a_{m-1,\ell-1})$$

$$= (y_{(0)tm+0}, y_{tm+0}, y_{2tm+0}, \ldots, y_{(\ell-1)tm+0}, y_{(\ell-1)tm+1}, y_{(0)tm+1}, y_{tm+1}, \ldots, y_{(\ell-2)tm+1},$$

$$\ldots, y_{(\ell-m+1)tm+(m-1)}, y_{(\ell-m+2)tm+(m-1)}, y_{(\ell-m+3)tm+(m-1)}, \ldots, y_{(\ell-m)tm+(m-1)}),$$

which is the same as Lemma 5.

4 The Generator Polynomial of $\theta(\mathscr{C})$

In this section, we make an attempt to describe the generator polynomials of $\mathscr{C}$ and $\theta(\mathscr{C})$ over $\mathbb{F}_q$ without using the results of [4].

Definition 3 For $\mathbf{c} = (a_{00}, a_{01}, a_{02}, \ldots, a_{0,\ell-1}, a_{10}, a_{11}, a_{12} \ldots, a_{1,\ell-1}, \ldots, a_{m-1,0}, \ldots, a_{m-1,\ell-1}) \in \mathscr{C}$, we define a mapping ϕ which maps from the codeword $\mathbf{c} \in \mathscr{C}$ to bivariate polynomial ring $\mathbb{F}_q[x, y]/\langle x^m - 1, y^\ell - 1\rangle$.

$$\phi : \mathbf{c} \mapsto \phi(\mathbf{c}) = a_{00} + a_{01}y + a_{02}y^2 + \cdots + a_{ij}x^i y^j + \cdots + a_{m-1,\ell-1}x^{m-1}y^{\ell-1},$$

where $0 \le i \le m-1, 0 \le j \le \ell - 1$.

Theorem 3 *J is a principal ideal of $\mathbb{F}_q[x, y]/\langle x^m - 1, y^\ell - 1\rangle$ if and only if $\mathscr{C}$ is a quasi-cyclic code of length ℓm and index ℓ with cyclic constituent codes, where $J = \phi(\mathscr{C})$.*

Proof For $\mathbf{c} = (a_{00}, a_{01}, a_{02}, \ldots, a_{0,\ell-1}, a_{10}, \ldots, a_{1,\ell-1}, \ldots, a_{m-1,0}, \ldots, a_{m-1,\ell-1}) \in \mathscr{C}$, namely, $\phi(\mathbf{c}) = a_{00} + a_{01}y + a_{02}y^2 + \cdots + a_{ij}x^i y^j + \cdots + a_{m-1,\ell-1}x^{m-1}y^{\ell-1} \in J$, then we have $x\phi(\mathbf{c}) = a_{00}x + a_{01}xy + a_{02}xy^2 + \cdots + a_{ij}x^{i+1}y^j + \cdots + a_{m-1,\ell-1}y^{\ell-1} \in J$. Therefore

$$(a_{m-1,0}, a_{m-1,1}, a_{m-1,2}, \ldots, a_{m-1,\ell-1}, a_{00}, \ldots, a_{0,\ell-1}, \ldots, a_{m-2,0}, \ldots, a_{m-2,\ell-1}) \in \mathscr{C} \quad (5)$$

and $y\phi(\mathbf{c}) = a_{00}y + a_{01}y^2 + a_{02}y^3 + \cdots + a_{ij}x^i y^{j+1} + \cdots + a_{m-1,\ell-1}x^{m-1} \in J$, then

$$(a_{0,\ell-1}, a_{00}, a_{01}, \ldots, a_{0,\ell-2}, a_{1,\ell-1}, \ldots, a_{1,\ell-2}, \ldots, a_{m-1,\ell-1}, \ldots, a_{m-1,\ell-2}) \in \mathscr{C} \quad (6)$$

Moreover, J is a principal ideal, then $x^i y^j \phi(\mathbf{c}) \in J$, and

$$\phi^{-1}(x^i y^j \phi(\mathbf{c})) \in \mathscr{C}. \quad (7)$$

Since J is a principal ideal, then $\mathscr{C}$ is linear. Moreover, $\mathscr{C}$ satisfies Eqs. (5)-(7), so that $\mathscr{C}$ is a quasi-cyclic code with cyclic constituent codes.

Next, we consider the converse part. From Theorem 1, $\theta(\mathscr{C})$ is a cyclic code, then $\theta(\mathscr{C})$ is a principal ideal of $\mathbb{F}_q[z]/\langle z^{\ell m}-1\rangle$, let the generator polynomial of $\theta(\mathscr{C})$ be

$$g(z)=\sum_{i=0}^{\ell m-1} a_{i_m,i_\ell} z^i,$$

then $\theta(\mathbf{c})=(a_{00},a_{11},\ldots,a_{i_m,i_\ell},\ldots,a_{m-1,\ell-1})\in\theta(\mathscr{C})$, according to Corollary 2, we have

$$\mathbf{c}=(a_{00},a_{01},a_{02},\ldots,a_{0,\ell-1},a_{10},\ldots,a_{1,\ell-1},\ldots,a_{m-1,0},\ldots,a_{m-1,\ell-1})\in\mathscr{C}.$$

Now we claim that $\phi(\mathscr{C})=\langle\phi(\mathbf{c})\rangle$. Clearly, $\phi(\mathbf{c})\in\phi(\mathscr{C})$, thus

$$\langle\phi(\mathbf{c})\rangle\subseteq\phi(\mathscr{C}). \tag{8}$$

It is easy to check that $xy\phi(\mathbf{c})=$

$$\phi(a_{m-1,\ell-1},a_{m-1,0},\ldots,a_{m-1,\ell-2},a_{0,\ell-1},\ldots,a_{0,\ell-2},\ldots,a_{m-2,\ell-1},\ldots,a_{m-2,\ell-2}).$$

And $(a_{m-1,\ell-1},a_{m-1,0},\ldots,a_{m-1,\ell-2},a_{0,\ell-1},\ldots,a_{0,\ell-2},\ldots,a_{m-2,\ell-1},\ldots,a_{m-2,\ell-2})$ is exactly the second row of the similar circulant matrix A' containing $\mathbf{c}$. From Lemma 4, $xy\phi(\mathbf{c})$ is equivalent to $zg(z)$, since $zg(z)$ is the second row of $\theta(A')$, similarly, $z^2g(z)$ is equivalent to $x^2y^2\phi(\mathbf{c})$, and so on.

Since the coordinate transformation θ is a linear mapping, then we can define a mapping Ψ which maps from the polynomial (codeword) of $\theta(\mathscr{C})$ to the equivalent polynomial (codeword) of $\langle\phi(\mathbf{c})\rangle$. Namely,

$$\Psi: f(z)g(z)\in\theta(\mathscr{C})\mapsto f(xy)\phi(\mathbf{c})\in\langle\phi(\mathbf{c})\rangle\subseteq\phi(\mathscr{C}).$$

Next we prove the mapping Ψ is bijective. For $\theta(\mathbf{c}')\in\theta(\mathscr{C})$, since $\theta(\mathscr{C})$ is a principal ideal, we can write $\theta(\mathbf{c}')=f_1(z)g(z)$, from the equivalence between $\mathscr{C}$ and $\theta(\mathscr{C})$, we can obtain $\phi(\mathbf{c}')=f_1(xy)\phi(\mathbf{c})\in\phi(\mathscr{C})$. It is clear that Ψ is injective. Now it is sufficient to prove that $x^iy^j\phi(\mathbf{c})$ has its pre-image in $\theta(\mathscr{C})$, rewrite

$$x^iy^j=x^{k_1m+i}y^{k_2\ell+j},$$

and it is clear that the equation $k_1m+i=k_2\ell+j$ has integer solution (k_1,k_2), one can choose the pair (k_1,k_2) such that k_1m+i is the smallest. Set $k_1m+i=k_2\ell+j=e$, then $x^iy^j\phi(\mathbf{c})$ has pre-image $z^eg(z)\in\theta(\mathscr{C})$ for some positive integer e. Thus the mapping Ψ is bijective. Consequently,

$$|\theta(\mathscr{C})|=|\phi(\mathscr{C})|=|\langle\phi(\mathbf{c})\rangle|. \tag{9}$$

Combining (8) and (9), we obtain $\langle\phi(\mathbf{c})\rangle=\phi(\mathscr{C})$.

From the proof of Theorem 3, we have the following corollaries.

Corollary 3 *Let $\mathscr{C}$ be a quasi-cyclic code of length ℓm and index ℓ with cyclic constituent codes, then $\phi(\mathscr{C})$ is a principal ideal of $\mathbb{F}_q[x, y]/\langle x^m - 1, y^\ell - 1\rangle$. Similar to the case of cyclic codes, $\phi(\mathbf{c}) = a_{00} + a_{01}y + a_{02}y^2 + \cdots + a_{ij}x^i y^j + \cdots + a_{m-1,\ell-1}x^{m-1}y^{\ell-1}$ is a generator polynomial of $\mathscr{C}$. Namely, $\mathscr{C}$ can be constructed by a principal ideal of $\mathbb{F}_q[x, y]/\langle x^m - 1, y^\ell - 1\rangle$.*

Corollary 4 *Let $\mathscr{C}$ be a quasi-cyclic code of length ℓm and index ℓ with cyclic constituent codes, and $\mathscr{C}$ has a generator polynomial $\phi(\mathbf{c}) = a_{00} + a_{01}y + a_{02}y^2 + \cdots + a_{ij}x^i y^j + \cdots + a_{m-1,\ell-1}x^{m-1}y^{\ell-1}$, then $\theta(\mathscr{C})$ is a cyclic code with the generator polynomial $g(z) = \sum_{i=0}^{\ell m-1} a_{i_m,i_\ell} z^i$.*

5 General Equivalences

In this section, we will give more general equivalences which include θ in Lemma 4 and the equivalence of Proposition 9 in [3] as a special case.

Theorem 4 *Let $\mathscr{C}$ be a quasi-cyclic code of length ℓm and index ℓ with cyclic constituent codes, then there exists another permutation θ' such that $\theta'(\mathscr{C})$ is a cyclic code and similar to the proof of Theorem 3, we can obtain another generator polynomial of $\phi(\mathscr{C})$.*

Proof If $\mathscr{C}$ is a quasi-cyclic code of length ℓm and index ℓ with cyclic constituent codes and $gcd(k_3, \ell) = gcd(k_4, m) = 1$, where k_3 and k_4 are positive integers, then for

$$(a_{00}, a_{01}, a_{02}, \ldots, a_{0,\ell-1}, a_{10}, \ldots, a_{1,\ell-1}, \ldots, a_{m-1,0}, \ldots, a_{m-1,\ell-1}) \in \mathscr{C},$$

we have

$$(a_{m-k_4,\ell-k_3}, a_{m-k_4,\ell-k_3+1}, \ldots, a_{m-k_4,\ell-1}, a_{m-k_4,0}, \ldots, a_{m-k_4,\ell-k_3-1},$$

$$a_{m-k_4+1,\ell-k_3}, \ldots, a_{m-k_4+1,\ell-k_3-1}, \ldots, a_{m-k_4-1,\ell-k_3}, \ldots, a_{m-k_4-1,\ell-k_3-1}) \in \mathscr{C}.$$

Similar to Definition 1, we can define a similar circulant matrix E' containing the codeword $(a_{00}, a_{01}, a_{02}, \ldots, a_{0,\ell-1}, a_{10}, \ldots, a_{1,\ell-1}, \ldots, a_{m-1,0}, \ldots, a_{m-1,\ell-1})$

$$E' = \begin{pmatrix} a_{00} & \cdots & a_{0,\ell-1} & \cdots & a_{m-1,0} & \cdots & a_{m-1,\ell-1} \\ a_{m-k_4,\ell-k_3} & \cdots & a_{m-k_4,\ell-k_3-1} & \cdots & a_{m-k_4-1,\ell-k_3} & \cdots & a_{m-k_4-1,\ell-k_3-1} \\ a_{m-2k_4,\ell-2k_3} & \cdots & a_{m-2k_4,\ell-2k_3-1} & \cdots & a_{m-2k_4-1,\ell-2k_3} & \cdots & a_{m-2k_4-1,\ell-2k_3-1} \\ \cdots & & \cdots & & \cdots & & \cdots \\ a_{k_4,k_3} & \cdots & a_{k_4,k_3-1} & \cdots & a_{k_4-1,k_3} & \cdots & a_{k_4-1,k_3-1} \end{pmatrix}.$$

Parallel to the proof of Lemma 4 and Corollary 2, there exists another permutation θ' such that $\theta'(E')$ is a circulant matrix.

Take $m = 5, \ell = 3, p = 2, k_3 = 2$ and $k_4 = 1$ for example. Let E' be a similar circulant matrix containing the codeword $(a_{00}, a_{01}, a_{02}, a_{10}, a_{11}, a_{12}, a_{20}, a_{21}, a_{22}, a_{30}, a_{31}, a_{32}, a_{40}, a_{41}, a_{42})$, namely,

$$E' = \begin{pmatrix}
a_{00} & a_{01} & a_{02} & a_{10} & a_{11} & a_{12} & a_{20} & a_{21} & a_{22} & a_{30} & a_{31} & a_{32} & a_{40} & a_{41} & a_{42} \\
a_{41} & a_{42} & a_{40} & a_{01} & a_{02} & a_{00} & a_{11} & a_{12} & a_{10} & a_{21} & a_{22} & a_{20} & a_{31} & a_{32} & a_{30} \\
a_{32} & a_{30} & a_{31} & a_{42} & a_{40} & a_{41} & a_{02} & a_{00} & a_{01} & a_{12} & a_{10} & a_{11} & a_{22} & a_{20} & a_{21} \\
a_{20} & a_{21} & a_{22} & a_{30} & a_{31} & a_{32} & a_{40} & a_{41} & a_{42} & a_{00} & a_{01} & a_{02} & a_{10} & a_{11} & a_{12} \\
a_{11} & a_{12} & a_{10} & a_{21} & a_{22} & a_{20} & a_{31} & a_{32} & a_{30} & a_{41} & a_{42} & a_{40} & a_{01} & a_{02} & a_{00} \\
a_{02} & a_{00} & a_{01} & a_{12} & a_{10} & a_{11} & a_{22} & a_{20} & a_{21} & a_{32} & a_{30} & a_{31} & a_{42} & a_{40} & a_{41} \\
a_{40} & a_{41} & a_{42} & a_{00} & a_{01} & a_{02} & a_{10} & a_{11} & a_{12} & a_{20} & a_{21} & a_{22} & a_{30} & a_{31} & a_{32} \\
a_{31} & a_{32} & a_{30} & a_{41} & a_{42} & a_{40} & a_{01} & a_{02} & a_{00} & a_{11} & a_{12} & a_{10} & a_{21} & a_{22} & a_{20} \\
a_{22} & a_{20} & a_{21} & a_{32} & a_{30} & a_{31} & a_{42} & a_{40} & a_{41} & a_{02} & a_{00} & a_{01} & a_{12} & a_{10} & a_{11} \\
a_{10} & a_{11} & a_{12} & a_{20} & a_{21} & a_{22} & a_{30} & a_{31} & a_{32} & a_{40} & a_{41} & a_{42} & a_{00} & a_{01} & a_{02} \\
a_{01} & a_{02} & a_{00} & a_{11} & a_{12} & a_{10} & a_{21} & a_{22} & a_{20} & a_{31} & a_{32} & a_{30} & a_{41} & a_{42} & a_{40} \\
a_{42} & a_{40} & a_{41} & a_{02} & a_{00} & a_{01} & a_{12} & a_{10} & a_{11} & a_{22} & a_{20} & a_{21} & a_{32} & a_{30} & a_{31} \\
a_{30} & a_{31} & a_{32} & a_{40} & a_{41} & a_{42} & a_{00} & a_{01} & a_{02} & a_{10} & a_{11} & a_{12} & a_{20} & a_{21} & a_{22} \\
a_{21} & a_{22} & a_{20} & a_{31} & a_{32} & a_{30} & a_{41} & a_{42} & a_{40} & a_{01} & a_{02} & a_{00} & a_{11} & a_{12} & a_{10} \\
a_{12} & a_{10} & a_{11} & a_{22} & a_{20} & a_{21} & a_{32} & a_{30} & a_{31} & a_{42} & a_{40} & a_{41} & a_{02} & a_{00} & a_{01}
\end{pmatrix}.$$

Set

$$h(y) = a_{01} + a_{10}y + a_{22}y^2 + a_{31}y^3 + a_{40}y^4 + a_{02}y^5 + a_{11}y^6 + a_{20}y^7 + a_{32}y^8 + a_{41}y^9 + a_{00}y^{10} + a_{12}y^{11} + a_{21}y^{12} + a_{30}y^{13} + a_{42}y^{14}.$$

Let ε be a primitive 15-th root of unity, and $\xi \in \{1, \varepsilon, \varepsilon^2, \ldots, \varepsilon^{14}\}$.

$$Q'_{3\times5}(\xi) = (\xi^{10}, 1, \xi^5, \xi, \xi^6, \xi^{11}, \xi^7, \xi^{12}, \xi^2, \xi^{13}, \xi^3, \xi^8, \xi^4, \xi^9, \xi^{14})^T,$$

$$P_{3\times5}(\xi) = (1, \xi, \xi^2, \xi^3, \xi^4, \xi^5, \xi^6, \xi^7, \xi^8, \xi^9, \xi^{10}, \xi^{11}, \xi^{12}, \xi^{13}, \xi^{14})^T,$$

and the correspondence between the coefficient of the term ξ^i in $h(\xi)$ and ξ^i is $a_{01} \leftrightarrow 1, a_{10} \leftrightarrow \xi, a_{22} \leftrightarrow \xi^2, a_{31} \leftrightarrow \xi^3, a_{40} \leftrightarrow \xi^4, a_{02} \leftrightarrow \xi^5, a_{11} \leftrightarrow \xi^6, a_{20} \leftrightarrow \xi^7, a_{32} \leftrightarrow \xi^8, a_{41} \leftrightarrow \xi^9, a_{00} \leftrightarrow \xi^{10}, a_{12} \leftrightarrow \xi^{11}, a_{21} \leftrightarrow \xi^{12}, a_{30} \leftrightarrow \xi^{13}, a_{42} \leftrightarrow \xi^{14}$.

It is easy to check that $E'Q'_{3\times5}(\xi) = h(\xi)P_{3\times5}(\xi)$, according to Lemma 4, there exists a permutation θ' in S_{15} such that

$$\theta'(E') = (P_{3\times5}(1), \ldots, P_{3\times5}(\xi^{14}))diag(h(1), \ldots, h(\xi^{14})(P_{3\times5}(1), \ldots, P_{3\times5}(\xi^{14}))^{-1}.$$

Consequently, E' is equivalent to the circulant matrix E containing the codeword

$$(a_{01}, a_{10}, a_{22}, a_{31}, a_{40}, a_{02}, a_{11}, a_{20}, a_{32}, a_{41}, a_{00}, a_{12}, a_{21}, a_{30}, a_{42}),$$

namely,

$$E = \begin{pmatrix}
a_{01} & a_{10} & a_{22} & a_{31} & a_{40} & a_{02} & a_{11} & a_{20} & a_{32} & a_{41} & a_{00} & a_{12} & a_{21} & a_{30} & a_{42} \\
a_{42} & a_{01} & a_{10} & a_{22} & a_{31} & a_{40} & a_{02} & a_{11} & a_{20} & a_{32} & a_{41} & a_{00} & a_{12} & a_{21} & a_{30} \\
a_{30} & a_{42} & a_{01} & a_{10} & a_{22} & a_{31} & a_{40} & a_{02} & a_{11} & a_{20} & a_{32} & a_{41} & a_{00} & a_{12} & a_{21} \\
a_{21} & a_{30} & a_{42} & a_{01} & a_{10} & a_{22} & a_{31} & a_{40} & a_{02} & a_{11} & a_{20} & a_{32} & a_{41} & a_{00} & a_{12} \\
a_{12} & a_{21} & a_{30} & a_{42} & a_{01} & a_{10} & a_{22} & a_{31} & a_{40} & a_{02} & a_{11} & a_{20} & a_{32} & a_{41} & a_{00} \\
a_{00} & a_{12} & a_{21} & a_{30} & a_{42} & a_{01} & a_{10} & a_{22} & a_{31} & a_{40} & a_{02} & a_{11} & a_{20} & a_{32} & a_{41} \\
a_{41} & a_{00} & a_{12} & a_{21} & a_{30} & a_{42} & a_{01} & a_{10} & a_{22} & a_{31} & a_{40} & a_{02} & a_{11} & a_{20} & a_{32} \\
a_{32} & a_{41} & a_{00} & a_{12} & a_{21} & a_{30} & a_{42} & a_{01} & a_{10} & a_{22} & a_{31} & a_{40} & a_{02} & a_{11} & a_{20} \\
a_{20} & a_{32} & a_{41} & a_{00} & a_{12} & a_{21} & a_{30} & a_{42} & a_{01} & a_{10} & a_{22} & a_{31} & a_{40} & a_{02} & a_{11} \\
a_{11} & a_{20} & a_{32} & a_{41} & a_{00} & a_{12} & a_{21} & a_{30} & a_{42} & a_{01} & a_{10} & a_{22} & a_{31} & a_{40} & a_{02} \\
a_{02} & a_{11} & a_{20} & a_{32} & a_{41} & a_{00} & a_{12} & a_{21} & a_{30} & a_{42} & a_{01} & a_{10} & a_{22} & a_{31} & a_{40} \\
a_{40} & a_{02} & a_{11} & a_{20} & a_{32} & a_{41} & a_{00} & a_{12} & a_{21} & a_{30} & a_{42} & a_{01} & a_{10} & a_{22} & a_{31} \\
a_{31} & a_{40} & a_{02} & a_{11} & a_{20} & a_{32} & a_{41} & a_{00} & a_{12} & a_{21} & a_{30} & a_{42} & a_{01} & a_{10} & a_{22} \\
a_{22} & a_{31} & a_{40} & a_{02} & a_{11} & a_{20} & a_{32} & a_{41} & a_{00} & a_{12} & a_{21} & a_{30} & a_{42} & a_{01} & a_{10} \\
a_{10} & a_{22} & a_{31} & a_{40} & a_{02} & a_{11} & a_{20} & a_{32} & a_{41} & a_{00} & a_{12} & a_{21} & a_{30} & a_{42} & a_{01}
\end{pmatrix}.$$

And the equivalence is given by $\theta' = (1\ 11\ 4\ 2)(3\ 6\ 12\ 9)(5\ 7\ 8\ 13)(10\ 14)(15)$ in S_{15}. However, $\theta = (2\ 11\ 14\ 5)(3\ 6\ 12\ 9)(4\ 7\ 13\ 10)$ in S_{15} by Lemma 4 and Corollary 2.

Similar to the proof of Theorem 1, $\theta'(\mathscr{C})$ is a cyclic code. Now we try to give another generator polynomial of $\phi(\mathscr{C})$. According to Definition 3,

$$\phi : \mathbf{c} \mapsto \phi(\mathbf{c}) = a_{00} + a_{01}y + a_{02}y^2 + \cdots + a_{ij}x^i y^j + \cdots + a_{m-1,\ell-1}x^{m-1}y^{\ell-1}.$$

And the linear mapping $\Psi_{(k_3,k_4)}$ (similar to Ψ in Theorem 3) is defined as follows,

$$\Psi_{(k_3,k_4)} : f(z)g(z) \in \theta(\mathscr{C}) \mapsto f(x^{k_4}y^{k_3})\phi(\mathbf{c}) \in \langle \phi(\mathbf{c}) \rangle \subseteq \phi(\mathscr{C}).$$

According to the proof of Theorem 3, $\Psi_{(k_3,k_4)}$ is one-to-one since $gcd(k_3, \ell) = gcd(k_4, m) = 1$. Then parallel to the proof of Theorem 3, the generator polynomial of $\phi(\mathscr{C})$ can be obtained.

Remark 3 According to the proof of Theorem 4, θ' relies on k_3 and k_4, and the similar circulant matrix A' in Sect. 3 is the case when $k_3 = k_4 = 1$.

6 Application Examples

In this section, we are ready to give some examples to illustrate the discussed results.

Example 1 If $\mathscr{C}$ is a quasi-cyclic code over $\mathbb{F}_q$ of length 6 and index 2 with cyclic constituent codes, where $(q, 6) = 1$, and let

$$B' = \begin{pmatrix} a_{00} & a_{01} & a_{10} & a_{11} & a_{20} & a_{21} \\ a_{21} & a_{20} & a_{01} & a_{00} & a_{11} & a_{10} \\ a_{10} & a_{11} & a_{20} & a_{21} & a_{00} & a_{01} \\ a_{01} & a_{00} & a_{11} & a_{10} & a_{21} & a_{20} \\ a_{20} & a_{21} & a_{00} & a_{01} & a_{10} & a_{11} \\ a_{11} & a_{10} & a_{21} & a_{20} & a_{01} & a_{00} \end{pmatrix}$$

be a similar circulant matrix of $\mathscr{C}$, where $\ell = 2, m = 3$, ε is a primitive 6-th root of unity, and $g(y) = a_{00} + a_{11}y + a_{20}y^2 + a_{01}y^3 + a_{10}y^4 + a_{21}y^5$. According to the proof of Lemma 4, the correspondence is $a_{00} \leftrightarrow 1, a_{11} \leftrightarrow \varepsilon, a_{20} \leftrightarrow \varepsilon^2, a_{01} \leftrightarrow \varepsilon^3, a_{10} \leftrightarrow \varepsilon^4, a_{21} \leftrightarrow \varepsilon^5$. Write

$$B' P'_{2\times 3}(\varepsilon) = (b_0, b_1, b_2, b_3, b_4, b_5)^T.$$

Set $b_0 = g(\varepsilon)$, then we have $P'_{2\times 3}(\varepsilon) = (1, \varepsilon^3, \varepsilon^4, \varepsilon, \varepsilon^2, \varepsilon^5)^T$. Then

$$B'(1, \varepsilon^3, \varepsilon^4, \varepsilon, \varepsilon^2, \varepsilon^5)^T = g(\varepsilon)(1, \varepsilon, \varepsilon^2, \varepsilon^3, \varepsilon^4, \varepsilon^5)^T.$$

Therefore

$$B' = \begin{pmatrix} a_{00} & a_{01} & a_{10} & a_{11} & a_{20} & a_{21} \\ a_{21} & a_{20} & a_{01} & a_{00} & a_{11} & a_{10} \\ a_{10} & a_{11} & a_{20} & a_{21} & a_{00} & a_{01} \\ a_{01} & a_{00} & a_{11} & a_{10} & a_{21} & a_{20} \\ a_{20} & a_{21} & a_{00} & a_{01} & a_{10} & a_{11} \\ a_{11} & a_{10} & a_{21} & a_{20} & a_{01} & a_{00} \end{pmatrix} \Leftrightarrow \theta(B') = \begin{pmatrix} a_{00} & a_{11} & a_{20} & a_{01} & a_{10} & a_{21} \\ a_{21} & a_{00} & a_{11} & a_{20} & a_{01} & a_{10} \\ a_{10} & a_{21} & a_{00} & a_{11} & a_{20} & a_{01} \\ a_{01} & a_{10} & a_{21} & a_{00} & a_{11} & a_{20} \\ a_{20} & a_{01} & a_{10} & a_{21} & a_{00} & a_{11} \\ a_{11} & a_{20} & a_{01} & a_{10} & a_{21} & a_{00} \end{pmatrix}.$$

And the equivalence is given by $\theta = (24)(35)$ in S_6.

Example 2 Let $\mathscr{C}$ be a quasi-cyclic code over $\mathbb{F}_5$ of length 6 and index 2 with cyclic constituent codes and the generator polynomial of $\phi(\mathscr{C})$ is $1 + xy + x^2(100110) \in \mathbb{F}_5[x, y]/\langle x^3 - 1, y^2 - 1\rangle$, where the codeword $\mathbf{c} = (100110)$ is the corresponding polynomial $1 + xy + x^2$ by Definition 3. Equivalently, $\phi(\mathscr{C}) = \langle \phi(\mathbf{c})\rangle$, then from Corollary 4, $\theta(\mathscr{C}) = \langle 1 + z + z^2\rangle(111000) \in \mathbb{F}_5[z]/\langle z^6 - 1\rangle$. And the linear mapping is

$$\Psi : \langle \phi(1 + z + z^2)\rangle \mapsto \langle 1 + xy + x^2\rangle,$$

according to the mapping Ψ, we have

$$1 \mapsto 1, z \mapsto xy = xy, z^2 \mapsto x^2y^2 = x^2, z^3 \mapsto x^3y^3 = y, z^4 \mapsto x^4y^4 = x, z^5 \mapsto x^5y^5 = x^2y$$

In more details:

$$\begin{aligned}
\phi(\mathbf{c}) = 1 + xy + x^2\ (100110) &\Leftrightarrow g(z) = 1 + z + z^2\ (111000)\\
xy\phi(\mathbf{c}) = y + xy + x^2\ (010110) &\Leftrightarrow zg(z) = z^3 + z + z^2\ (011100)\\
x^2\phi(\mathbf{c}) = x + y + x^2\ (011010) &\Leftrightarrow z^2g(z) = z^3 + z^4 + z^2\ (001110)\\
y\phi(\mathbf{c}) = y + x + x^2y\ (011001) &\Leftrightarrow z^3g(z) = z^3 + z^4 + z^5\ (000111)\\
x\phi(\mathbf{c}) = x + x^2y + 1\ (101001) &\Leftrightarrow z^4g(z) = 1 + z^4 + z^5\ (100011)\\
x^2y\phi(\mathbf{c}) = 1 + xy + x^2y\ (100101) &\Leftrightarrow z^5g(z) = 1 + z + z^5\ (110001)
\end{aligned}$$

and $f(z)g(z) \mapsto f(xy)\phi(\mathbf{c})$ is given by the linearity of $\mathscr{C}$ and $\theta(\mathscr{C})$. And the equivalence is given by $\theta = (24)(35)$ in S_6.

Example 3 Let $\mathscr{C}$ be a quasi-cyclic code over $\mathbb{F}_5$ of length 12 and index 4 with cyclic constituent codes, and

$$\phi(\mathscr{C}) = \langle 1 + y^3 + xy + x^2y^2\rangle(100101000010) \in \mathbb{F}_5[x, y]/\langle x^3 - 1, y^4 - 1\rangle,$$

then $\theta(\mathscr{C}) = \langle 1 + z + z^2 + z^3\rangle(111100000000) \in \mathbb{F}_5[z]/\langle z^{12} - 1\rangle$, the linear mapping is $\Psi : \langle\phi(1 + z + z^2 + z^3)\rangle \mapsto \langle 1 + y^3 + xy + x^2y^2\rangle$, and

$$1 \mapsto 1, z \mapsto xy, z^2 \mapsto x^2y^2, z^3 \mapsto x^3y^3 = y^3, z^4 \mapsto x^4y^4 = x, z^5 \mapsto x^5y^5 = x^2y, z^6 \mapsto x^6y^6 = y^2,$$

$$z^7 \mapsto x^7y^7 = xy^3, z^8 \mapsto x^8y^8 = x^2, z^9 \mapsto x^9y^9 = y, z^{10} \mapsto x^{10}y^{10} = xy^2, z^{11} \mapsto x^{11}y^{11} = x^2y^3.$$

And the equivalence is given by $\theta = (2\ 10\ 6)(3\ 7\ 11)$ in S_{12}.

Acknowledgements This research is supported by National Natural Science Foundation of China (61202068, 61672036 and 11526045), the Open Research Fund of National Mobile Communications Research Laboratory, Southeast University (2015D11), Technology Foundation for Selected Overseas Chinese Scholar, Ministry of Personnel of China (05015133) and Key Projects of Support Program for outstanding young talents in Colleges and Universities (gxyqZD2016008).

References

1. Bracco, A.D., Natividad, A.M., Solé, P.: On quintic quasi-cyclic codes. Discrete Appl. Math. **156**, 3362–3375 (2008)
2. Güneri, C., Özbudak, F.: A relation between quasi-cyclic codes and 2-D cyclic codes. Finite Fields Their Appl. **18**, 123–132 (2012)
3. Lim, C.J.: Quasi-cyclic codes with cyclic constituent codes. Finite Fields Their Appl. **13**, 516–534 (2007)
4. Ling, S., Solé, P.: On the algebraic structure of quasi-cyclic codes I: Finite fields. IEEE Trans. Inform. Theory **47**, 2751–2760 (2001)

Factorization of Computations in Bayesian Networks: Interpretation of Factors

Linda Smail and Zineb Azouz

Abstract Given a Bayesian network (BN) relative to a set I of discrete random variables, we are interested in computing the probability distribution P_S, where the target S is a subset of I. The general idea is to express P_S in the form of a product of factors whereby each factor is easily computed and can be interpreted in terms of conditional probabilities. In this paper, a condition stating when P_S can be written as a product of conditional probability distributions is called a non-pathology condition. This paper also considers an interpretation of the factors involved in computing marginal probabilities in BNs and a representation of the probability target as a Bayesian network of level two. Establishing such a factorization and interpretations is indeed interesting and relevant in the case of large BNs.

Keywords Bayesian networks · Bayesian networks of level two · Inference · Pathological bayesian networks

2010 Mathematics Subject Classification: 62F15

1 Introduction

Given a set I and a directed acyclic graph (DAG) $\mathscr{G}$ on I, a Bayesian Network (BN) (see [5, 6]) is a family of random variables $X_I = (X_i)_{i \in I}$, where X_i has values in Ω_i, implying that X_I has values in $\Omega_I = \prod_{i \in I} \Omega_i$. By definition, a BN is such that, for all i, the conditional probability X_i, which is conditioned on the set of all random variables other than itself and its descendants (denoted by $d(i)$), depends only on the

L. Smail (✉)
Mathematics and Statistics Department, Zayed University, P.O. Box 19282, Dubai, United Arab Emirates
e-mail: linda.smail@zu.ac.ae

Z. Azouz
Mathematics Department Mentouri University, route d'Ain El Bey, 25000 Constantine, Algeria
e-mail: zineb.azouz@gmail.com

T. Abualrub et al. (eds.), *Mathematics Across Contemporary Sciences*, Springer Proceedings in Mathematics & Statistics 190,
DOI 10.1007/978-3-319-46310-0_13

value of $x_{p(i)}$ taken by the set of its parents: $\forall\, i\ \forall\, x_I\ \ P_{i|I-(\{i\}\cup d(i))}(x_i|x_{I-(\{i\}\cup d(i))}) = P_{i|p(i)}(x_i|x_{p(i)})$ with the convention that i is a root of the graph $\mathcal{G}$ (that is $p(i) = \emptyset$) and $P_{i|p(i)}$ is the probability of X_i. The result is that the expression of the joint probability distribution of the family X_I is as follows:

$$P_I(x_I) = \prod_{i\in I} P_{i|p(i)}(x_i|x_{p(i)}). \tag{1}$$

We are interested in the computation of the restrictions of the probability distribution P_I of the BN. In other words, given a subset S of I, we are interested in the computation of P_S, the joint probability distribution of $X_S = (X_i)_{i\in S}$. Given $x_S \in \Omega_S$ and the decomposition of x_I into (x_S, x_{I-S}), we consider the following computation:

$$P_S(x_S) = \sum_{x_{I-S}\in\Omega_{I-S}} P_I(x_S, x_{I-S}). \tag{2}$$

By definition, the restriction of a BN to an initial part J of I (which is a subset where $\forall j \in J\ p(j) \subset J$) has the structure of a BN. Thus, for any subset S, P_S can be computed by considering the restriction to the initial part generated by S, which is denoted as S^+:

$$P_S(x_S) = \sum_{x_{S^+-S}\in\Omega_{S^+-S}} P_I(x_S, x_{S^+-S}). \tag{3}$$

Therefore, without loss of generality, we suppose that $S^+ = I$; in other words, all the leaves of I are in S. If $p(i)$ is the set of the parents of i, then

$$P_S(x_S) = \sum_{x_{I-S}\in\Omega_{I-S}} P_I(x_S, x_{I-S}) = \sum_{x_{I-S}\in\Omega_{I-S}} \prod_{i\in I} P_{i|p(i)}(x_i|x_{p(i)}). \tag{4}$$

It is important to construct an ordering for some simplification techniques of the summations over $x_{I-S} \in \Omega_{I-S}$. Finding a way to simplify and order such summations, segment them into several computations that can be run in parallel, and find an interpretation of each of the intermediate factors is considered the main issue related to inference in BNs.

A wide range of literature on BNs concerns the performance of methods used to compute P_S, where S is small compared to I (see [7], [9], [10], [14], [13]). It is particularly useful to write the expression of P_S in the form of a product of factors whereby each factor is easy to compute and has an interpretation in terms of conditional probabilities.

In the special case of Markov chains where $I = \{1, ..., n\}$, $S = \{s_1, s_2, ..., s_m\}$ (with $s_1 < s_2 < ... < s_m$) and where the probability distribution is defined by P_1 (the probability of X_1) and by the conditional probability distributions $P_{i|i-1}$ (for $i \geq 2$), it is evident that $P_S = \prod_{k=1}^{m} A_k(x_s)$.

Furthermore, with two computations for $A_k(x_k)$, we obtain the following: if $s_k = s_{k-1} + 1$, then $A_k(x_k) = P_{k|k-1}(x_k|x_{k-1})$ (which is $P_1(x_1)$ if $k = 1$ and $s_1 = 1$). In addition, if $s_k > s_{k-1} + 1$, then

$$A_k(x_S) = \sum_{(x_{s_{k-1}+1},\ldots,x_{s_k-1})} \prod_{i=s_{k-1}+1}^{s_k} P_{i|i-1}(x_i|x_{i-1}). \tag{5}$$

This factor depends exclusively on $x_{s_{k-1}}$ and x_{s_k} (indeed, it is exactly $P_1(x_{s_1})$ if $s_1 > 1$ and $k = 1$) and can be interpreted as $P_{s_k|s_{k-1}}(x_{s_k}|x_{s_{k-1}})$.

In the general case, it remains true that there exists a partition $\mathscr{C}$ of $I - S$ where, if we denote by L the set of variables of S that have no parent in $I - S$, $P_S(x_S) = \prod_{\ell \in L} A_\ell(x_s) \prod_{C \in \mathscr{C}} A_C(x_S)$, where $A_\ell(x_s) = P_{\ell|p(\ell)}(x_\ell|x_{p(\ell)})$ (given by the BN). Moreover, if we denote by $T(C)$ the set of children of variables of C that are not themselves in C, then

$$A_C(x_S) = \sum_{x_C \in \Omega_C} \prod_{i \in C \cup T(C)} P_{i|p(i)}(x_i|x_{p(i)}). \tag{6}$$

We would like to evaluate the existence of an interpretation of $A_C(x_S)$, similar to the interpretation of $A_k(x_s)$ for the Markov chain in the case where $s_k > s_{k-1} + 1$. In most BNs, it is true that

$$A_C(x_S) = P_{T(C)|R(C)}(x_{T(C)}|x_{R(C)}), \tag{7}$$

where $R(C)$ is the set of elements of the Markov blanket of C (see [5]) being neither in C nor in $T(C)$.

We will show that for this interpretation to hold, it is sufficient that the graph on I relative to the case in which the set $(X_I)_{i \in I}$ is a BN possesses a property that we will call non-pathological. This property is necessary in the sense that, if it is not satisfied, we can construct a BN for which the desirable interpretation of the factors $A_C(x_S)$ is impossible.

Butz et al. in [3] analyzed the inference task in the junction tree algorithm (see [9] for additional details about the junction tree algorithm) and was able to write each intermediate computation, namely, the product of potentials, as probability distributions but only for BNs satisfying a specific condition of topological order called Butz's condition. (Presented in Sect. 2).

In this paper, we focus on the interpretation of the factors involved in computing marginal probabilities in BNs. We characterize all BNs where Butz's condition does not apply, which are denoted as pathological BNs (See Definition 1). We show in Theorem 1, the main result of this paper, that the desired factorization is not possible for such networks. For non-pathological BNs, we prove the possibility of writing the target probability distribution as a product of conditional probability

distributions. Furthermore, we relate the factorization of computations for these BNs to the structure of Bayesian networks of level two (BN2), introduced in [17].

This paper is organized as follows. Section 2 explains the motivation of our work and presents the problem of the computation of restrictions in BNs in both the special case of Markov chains and the general case of BNs. Section 3 introduces Pathological Bayesian networks and gives the main theorem and its proof. Section 4 recalls Bayesian networks of level two and discusses their relationship with non-pathological BNs, that is writing the target probability distribution as a product of conditional probability distribution associated to a BN2.

2 Presentation of the Problem: Computation of Restrictions in Bayesian Networks

2.1 Motivation

Here, we extend the work presented by Smail in [17]. In that work, a new computation algorithm called the Successive Restriction Algorithm was presented along with the possible factorizations of those computations. That method was in contrast to classical methods, which, regardless of the applied computing procedure, result in target probability distributions under an extensive form and obtain intermediate computations as potentials lacking clear probabilistic interpretations (see [4], [8]). The successive restriction algorithm has the advantage of being able to present, at each stage, interpretable results in terms of conditional probabilities under the form of a Bayesian network of level two and thus is technically usable (see [16]).

The same concerns were presented in [18], in which Studeny presented a factorization formula on the largest chain graph equivalent to the BN found based on chain graphs and d-Separation properties. In addition, Shafer [12] provided a condition under which each product of two potential functions, namely, $\phi(X_1|Y_1)$ and $\phi(X_2|Y_2)$, yields a conditional probability table $\phi(X_1, X_2|Y_1, Y_2)$. In other words, $\phi(X_1, X_2|Y_1, Y_2) = \phi(X_1|Y_1)\phi(X_2|Y_2)$ provided that X_2 is disjoint of $\{X_1, Y_1\}$.

In [1], Butz and Yan presented the following example (Fig. 1) wherein the Shafer condition does not apply and introduced a new method for the semantics of the conditional probability tables in the variable elimination algorithm.

Example 1 Eliminating variable c from the graph in Fig. 1 yields the following:

$$\sum_c P(c)P(d|a,b,c)P(e|d,c) = \sum_c [P(c)P(e|d,c)]\,P(d|a,b,c) \tag{8}$$

$$= \sum_c \phi(c,e|d)P(d|a,b,c) \tag{9}$$

$$= \phi(a,b,c,d,e).$$

Fig. 1 A Bayesian network

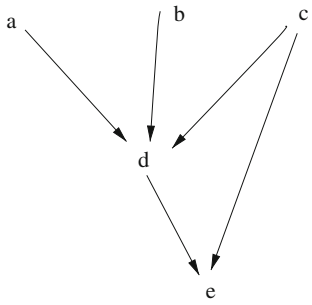
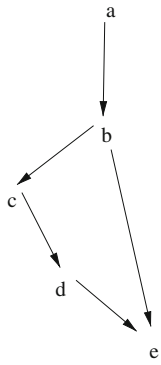


Fig. 2 A Bayesian network



As we notice, the Shafer condition applies to the product in Eq. (8) but does not apply to the product in Eq. (9); therefore, the final result should be noted as a potential (extensive form) and not as a probability distribution table, it means as $\phi(a, b, c, d, e)$.

In [3], Butz et al. presented a condition on the BN graph allowing potentials, as the one above, to be written as conditional probability tables. Butz's condition is given as follows:

$P(X|Y) = \prod_{x_i \in X} P(x_i|p(x_i))$, where $Y = (\bigcup_{x_i \in X} p(x_i)) - X$, if there is a topological order of I in which the variables of X appear consecutively.

However, in [2], Butz et al. stated that, using the following example (Fig. 2), the topological condition is sufficient but not necessary to ensure a result as a conditional probability table.

Example 2 Let us consider eliminating the variable b from the graph above (Fig. 2).

$$\sum_b P(b|a)P(c|b)P(d|c)P(e|d,b) = \sum_b [P(b|a)P(c|b)P(d|c)]\, P(e|d,b)$$
$$= \sum_b \phi(b,c|a)P(e|d,b) \quad (10)$$
$$= \sum_b \phi(b,c,e|a,d) \quad (11)$$
$$= \phi(c,e|a,d).$$

By Butz's condition, $\phi(b,c|a)$ in Eq. (10) is $P(b,c|a)$ because (b,c) is a topological order; however, $\phi(b,c,e|a,d)$ in Eq. (11) is not $\phi(b,c,e|a,d)$ because all topological orders between c and e have d in between.

This work was undertaken to address the semantics and interpretations of marginal probability distributions computed from BNs similar to the example above. A characterization of such BNs is given along with a necessary and sufficient condition that enables the factorization under the form of conditional probability distributions.

2.2 A Special Case: A Markov Chain

The simplification of the computation of P_S is a classic problem in the case of a finite Markov chain, where $I = \{1, .., n\}$ and $X_I = (X_1, X_2, ..., X_n)$. The graph $\mathcal{G}$ in this case is composed of the elementary pairs $(i-1, i)$ $(2 \leq i \leq n)$. Therefore, for each $i \in \{1, 2, ..., n\}$, its unique parent is $i-1$ $(p(i) = \{i-1\})$; furthermore, for each $i \in \{1, 2, ..., n-1\}$, its descendant is $d(i) = \{i+1, i+2, ..., n\}$. Thus, if $S = \{s_1, s_2, ...s_m\}$, where $(s_1 < s_2 < ... < s_m)$, and if $s_m = n$ (so that $S^+ = \{1, ..., n\}$), then $X_S = (X_{s_1}, ..., X_{s_m})$ is also a Markov chain, and P_S can be written as follows:

$$P_S(x_{s_1}, ..., x_{s_m}) = \prod_{k=1}^{m} P_{s_k|s_{k-1}}(x_{s_k}|x_{s_{k-1}}) \quad (12)$$

with the convention that if $k = 1$, $P_{s_1|s_0}(x_{s_1}|x_{s_0}) = P_{s_1}(x_{s_1})$.

Let L be the set of indices ℓ such that s_ℓ has no parents in $I - S$ (thus, either $\ell = 1$ with $s_1 = 1$ or $s_{\ell-1} \in S$). In other words, if $\ell \in L$, then s_ℓ is either 1 or the child of $s_{\ell-1}$, and $p_{s_\ell|s_{\ell-1}}$ is one of the entries of the Markov chain in the latter case. If $k \notin L$, the computation of $P_{s_k|s_{k-1}}$ is simple and uses only the elements of $C_k = \{s_{k-1}+1, ..., s_k - 1\}$ (such that if $s_1 > 1$, we obtain the particular case $C_1 = \{1, ..., s_1 - 1\}$). The subsets C_k (where $k \notin L$) form a partition of $I - S$, and hence, we have the following:

$$P_{s_k|s_{k-1}}(x_{s_k}|x_{s_{k-1}}) = \sum_{x_{C_k} \in \Omega_{C_k}} \prod_{i=s_{k-1}+1}^{s_k} P_{i|i-1}(x_i|x_{i-1}). \quad (13)$$

In summary, due to Eq. (12), the computation of $P_S(x_{s_1}, ..., x_{s_m})$ can be conducted in the following form:

$$P_S(x_{s_1}, ..., x_{s_m}) = \left[\prod_{k\in L} P_{s_k|s_{k-1}}(x_{s_k}|x_{s_{k-1}})\right] \times \left[\prod_{k\notin L}\sum_{x_{C_k}\in\Omega_{C_k}}\prod_{i=s_{k-1}+1}^{s_k} P_{i|i-1}(x_i|x_{i-1})\right]. \tag{14}$$

Example 3 Let us consider a Markov chain consisting of $\{X_0, X_1, X_2, X_3, X_4, X_5, X_6, X_7\}$. The joint probability distribution in this case can be written as follows:

$$\begin{aligned} P(x_0, \ldots, x_7) = {} & P(x_0)P(x_1|x_0)P(x_2|x_1)P(x_3|x_2)P(x_4|x_3)P(x_5|x_4) \\ & \times P(x_6|x_5)P(x_7|x_6). \end{aligned}$$

Let us consider the subset $S = \{X_1, X_4, X_5\}$, and let us attempt to write its probability distribution $P(S)$. To do so, we need to eliminate the variables $\{X_0, X_2, X_3, X_6, X_7\}$ from the joint probability distribution $P(x_0, \ldots, x_7)$ as follows:

$$\begin{aligned} P(S) &= \sum_{x_0,x_2,x_3,x_6,x_7} P(x_0, \ldots, x_7) \\ &= P(x_5|x_4) \times \left[\sum_{x_0} P(x_0)P(x_1|x_0)\right] \\ &\quad \times \left[\sum_{x_2,x_3} P(x_2|x_1)P(x_3|x_2)P(x_4|x_3)\right]\left[\sum_{x_6,x_7} P(x_6|x_5)P(x_7|x_6)\right]. \end{aligned}$$

Using probability distribution properties, $\left[\sum_{x_6,x_7} P(x_6|x_5)P(x_7|x_6)\right] = 1$; therefore,

$$P(S) = P(x_5|x_4)\left[\sum_{x_0} P(x_0)P(x_1|x_0)\right]\left[\sum_{x_2,x_3} P(x_2|x_1)P(x_3|x_2)P(x_4|x_3)\right].$$

Using the above notations, let us consider $s_1 = 1, s_2 = 4$, and $s_3 = 5$, namely, the indices of the variables in S. We notice in this case that only 3 has no parents in $\{X_0, X_2, X_3, X_6, X_7\}$. Thus, for each k in $\{1, 2\}$, complementary of $L = \{3\}$, we obtain $C_1 = \{0, 1\}$ and $C_2 = \{2, 3\}$.

Thus, the composition of P_S can be written as follows:

$$P(S) = P(x_5|x_4) \times \left[\sum_{C_1}\prod_{i=0}^{i=1} P(x_i|x_{p(i)})\right]\left[\sum_{C_2}\prod_{i=2}^{i=4} P(x_i|x_{p(i)})\right].$$

2.3 The General Case

The generalization of the Markov chain composition for BNs will use the following remark: in a Markov chain, if i' and i'' (where $i' < i''$) are two indices of I not in S that belong to $C_{k'}$ and $C_{k''}$, respectively, where $k' \neq k''$, then the only path that connects i' to i'' contains at least one element of S.

Therefore, in the general case of BNs, the subsets C_k are introduced as equivalent classes associated with an equivalence relation on $I - S$. For this relation, i' and i'' are equivalent if and only if they are not d-Separated by S (the classic notion introduced by [11]). In other words, i' and i'' are equivalent if and only if there exists, in the moral graph associated to $\mathscr{G}$, a Markov chain that links i' to i''.
We recall that the moral graph associated with the DAG $\mathscr{G}$ is the undirected graph $\mathscr{H}$ in which the links are the pairs $\{i', i\}$ such that one of the elements is a parent of the other one, or they have a common child.

We notice that, in a Markov chain, the moral graph $\mathscr{H}$ has as links the pairs $\{i-1, i\}$ $(2 \leq i \leq n)$.
Therefore, the S-conditional partition, denoted by $\mathscr{C}$, is defined as the partition of $I - S$ on equivalent classes for this relation. Furthermore, let us denote by L the set of elements of S that have no parent in $I - S$ (regardless of whether they are roots or their parents are all in S).

For each part $C \in \mathscr{C}$, we introduce the following:

- $M(C)$, the Markov blanket of C, defined as the set of elements of I that are either in C or are neighbors to at least one element of C in the moral graph. In other words, the elements of $M(C)$ which are in C, are parents or children of at least one element of C, or have a common child with at least one element in C.
- $F(C)$, the Markov boundary of C, defined as the set of elements of $M(C)$ that are not in C.
  We notice that $F(C)$ is a subset of S. Consider $i \in F(C)$: if i does not belong to S, then it belongs to an equivalent set C' other than C, which is absurd because it is linked to at least one element of C in the moral graph.
- $(T(C), R(C))$ the partition of $F(C)$, which is called the canonical partition, where $R(C)$ is the set of elements of $F(C)$ that are not children of at least one element of C.

We previously made the assumption that $I = S^+$. It immediately follows that, for each nonempty C of $S^+ - S$ (which is the case for the elements of the $S-$conditional partition), $T(C)$ is also nonempty, and C is a subset of $T(C)^+$.

In the case of a Markov chain of length n, if $C = \{s_{k-1} + 1, ..., s_k - 1\}$ (where $k \geq 2$ and $s_k > s_{k-1}$), then we have $M(C) = \{s_{k-1}, ..., s_k\}$, $F(C) = \{s_{k-1}, s_k\}$, $T(C) = \{s_k\}$, and $R(C) = \{s_{k-1}\}$. In contrast, if $s_1 > 1$ and $C = \{1, ..., s_1 - 1\}$, then $M(C) = \{1, ..., s_1\}$, $F(C) = T(C) = \{s_1\}$, and $R(C) = \varnothing$.

For each BN, we show that (see [17])

$$P_S(x_S) = \prod_{\ell \in L} P_{\ell|p(\ell)}(x_\ell | p(x_\ell)) \times \prod_{C \in \mathscr{C}} \sum_{x_C \in \Omega_C} \prod_{i \in C \cup T(C)} P_{i|p(i)}(x_i | x_{p(i)}). \quad (15)$$

The above equation generalizes Eq. (14), which is well known for Markov chains. From Eq. (15), consider the factor $\sum_{x_C \in \Omega_C} \prod_{i \in C \cup T(C)} P_{i|p(i)}(x_i | x_{p(i)})$. We will check whether this factor depends exclusively on $x_{F(C)}$. For each i in $C \cup T(C)$, $P_{i|p(i)}$ involves only variables in the Markov blanket $M(C)$ by the definition of $M(C)$. After computing summations over the variables in C, this expression reduces to a function of $x_{M(C)-C}$, which means a function of only $x_{F(C)}$.
However, in the case of a Markov chain where $C = \{s_{k-1} + 1, ..., s_k - 1\}$ and $T(C) = \{s_k\}$, it is true that

$$\sum_{x_C \in \Omega_C} \prod_{i \in C \cup T(C)} P_{i|p(i)}(x_i | x_{p(i)}) = P_{s_k|s_{k-1}}(x_{s_k} | x_{s_{k-1}}).$$

We now check whether this result remains valid in the general case:

$$\sum_{x_C \in \Omega_C} \prod_{i \in C \cup T(C)} P_{i|p(i)}(x_i | x_{p(i)}) = P_{T(C)|R(C)}(x_{T(C)} | x_{R(C)}). \quad (16)$$

We first establish that this last equation does not hold in general and second find a characterization for the BNs for which the above equation holds. The BNs for which Eq. (16) does not hold are called pathological; this terminology is justified because there are situations related to the graph $\mathscr{G}$ where researchers perform modeling in terms of a BN to avoid such situations.

3 Pathological Bayesian Networks

Consider a fixed class C of the S-Conditional partition. Denote its Markov blanket by M, and denote the canonical partition of its Markov boundary F by (T, R). We recall that T is defined as the set of vertices that are children of at least one element of C. It seems natural to see the vertices appear "generally" after the elements of C and R. In other words, it is "abnormal" to find a parent belonging to T for an element of T^+ (the initial part generated by T) that is neither in T nor in C (it belongs to $R \cup J$, where $J = T^+ - M$).

This situation may occur as in the following examples (Fig. 3), which are the simplest cases where there exists an $i \in R \cup J$ such that the parent is in T (in graph 1 (left), $3 \in J$ and in graph 2 (right), $3 \in R$).

This circumstance is necessary and sufficient for the partition of the pathology and is the subject of the theorem being presented in this section.

For the following sections, we define the partitions (C, T, R, J) of T^+ with the following groupings:

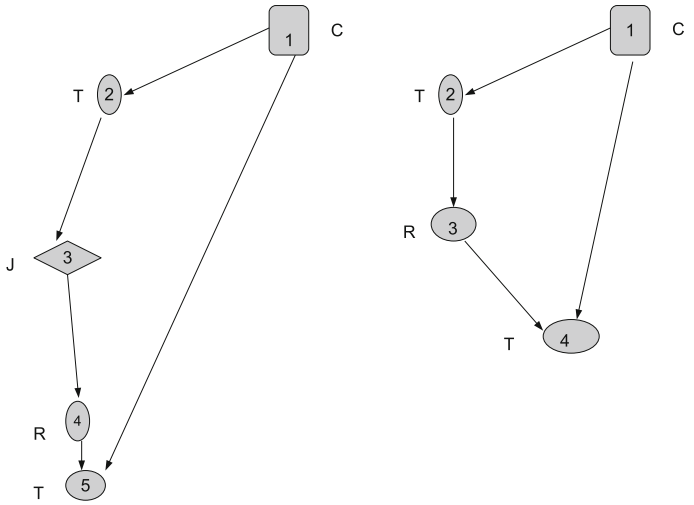


Fig. 3 Pathological BNs

$$F = T \cup R, M = C \cup F = C \cup T \cup R, T^{+} = M \cup J = C \cup T \cup R \cup J.$$

Definition 1 Given a subset C of I, the graph $\mathcal{G}$ presents a pathology on C if and only if there exists an $i \in R \cup J$ such that $p(i) \cap T \neq \varnothing$.

The following two lemmas provide necessary and sufficient conditions of the pathology property.

Lemma 1 *The graph $\mathcal{G}$ presents a pathology on C if and only if $R \cup J$ is not an initial part.*

Proof For $R \cup J$ to be an initial part, it is necessary and sufficient that there exists $i \in R \cup J$ such that at least one parent does not belong to $R \cup J$, in other words, it is enough that $p(i) \cap (C \cup T) \neq \varnothing$. However, we know that if $i \in R \cup J$, then $p(i) \cap C = \varnothing$ because, by the definition of T, a child of an element of C is necessarily in $C \cup T$. Hence, for each $i \in R \cup J$, the conditions $p(i) \cap (C \cup T) \neq \varnothing$ and $p(i) \cap T \neq \varnothing$ are equivalent. □

Lemma 2 *For the graph $\mathcal{G}$ to present a pathology on C, it is necessary and sufficient that there is a path in $\mathcal{G}$, say, $(j_0, \ldots j_n)$ (where $n \geq 1$) such that*
- $j_0 \in T$.
- $j_n \in R$.
- *If $n \geq 2$, then, for every k such that $1 \leq k \leq n-1$, we have $j_k \in J$.*

Proof The sufficient condition

Given that $(j_0, \ldots j_n)$ is a path as described in the lemma, we have $p(j_1) \cap T \neq \varnothing$, where $j_1 \in R \cup J$ (in other words, j_1 belongs to R if $n = 1$ and to J if $n > 1$).

The necessary condition

By assumption, there exists $i \in R \cup J$ such that $p(i) \cap T \neq \varnothing$. We assume that $j_1 = i$, and we consider j_0 to be a parent of i belonging to T.

If $i \in R$, then the lemma is proven (with the path (j_0, j_1)).

Let us assume now that $i \in J$.

Because $i \in T^+$, there exists $t \in T$ and a path $(j_1, \ldots, j_m)$ (with $m \geq 2$) such that $j_1 = i$ and $j_m = t$.

Assume that j_ℓ (where $2 \leq \ell \leq m$) is the first element that belongs to T on this path (it necessarily exists because $j_m \in T$). In other words, on the path $(j_1, \ldots, j_\ell)$, only the last element belongs to T.

By construction, $j_{\ell-1}$ has a child in common with an element of C (because $j_\ell \in T$), and therefore, $j_{\ell-1} \in M$. However, $j_{\ell-1} \notin T$; thus, $j_{\ell-1} \in C \cup R$.

Assume that j_n is the first element that belongs to $C \cup R$ on the path $(j_1, \ldots, j_{\ell-1})$; it is necessary that $n \geq 2$ because $j_1 \in J$. By its construction, the sequence $(j_1, \ldots, j_n)$ is such that, for any $k \leq n-1$, j_k can belong neither to T nor to $C \cup R$; therefore, it belongs to J.

It remains to be shown that $j_n \in R$. Because $j_n \in C \cup R$ by construction, j_n cannot belong to C. If j_n belonged to C, then j_{n-1}, which is one of its parents, must be in the Markov blanket of C or M. However, this situation is impossible because we have just shown that all j_k such that $1 \leq k \leq n-1$ belong to $J = T^+ - M$. □

The following theorem applies to any subset $C \subset I$ that satisfies the same hypotheses as the equivalent classes of the $S-$conditional partition. In section (4), we will comment on this theorem's use in the computation of the probability P_S.

Theorem 1 *Let C be a nonempty subset of I such that T is a nonempty set of children not in C and that $C \subset T^+$. For the existence of a BN $(X_i)_{i \in I}$ such that*

$$\sum_{x_C \in \Omega_C} \prod_{i \in C \cup T(C)} P_{i|p(i)}(x_i|x_{p(i)}) \neq P_{T|R}(x_T|x_R),$$

it is necessary and sufficient that the graph $\mathcal{G}$ presents a pathology on C.

Proof The expression $P_{T|R}(x_T|x_R)$ appearing in the theorem is equal to $\frac{P_F(x_F)}{P_R(x_R)}$. It will be computed in the general case by first computing $P_M(x_M)$, then $P_F(x_F) = \sum_{x_C \in \Omega_C} P_M(x_C, x_F)$, and finally $P_R(x_R) = \sum_{x_T \in \Omega_T} P_F(x_T, x_R)$.

We will observe where the condition of the theorem, meaning where the presence of a pathology on C, contradicts the equality between $P_{T|R}(x_T|x_R)$ and $\sum_{x_C \in \Omega_C} \prod_{i \in C \cup T(C)} P_{i|p(i)}(x_i|x_{p(i)})$.

The computation of $P_M(x_M)$

M and J are subsets of T^+, by definition an initial subset and thus preserving the structure of a BN. We are permitted to obtain $P_M(x_M)$ based only on computations on T^+.

$$P_M(x_M) = \sum_{x_J \in \Omega_J} \prod_{i \in T^+} P_{i|p(i)}(x_i|x_{p(i)}).$$

According to the cases where i belongs to C, T, R and J, it is useful to examine which components of x_{T^+} interfere in x_i and $x_{p(i)}$. If neither i nor its parents belong to J, then $P_{i|p(i)}(x_i|x_{p(i)})$ can be factored during the summations $\sum_{x_J \in \Omega_J}$.

This factoring is valid if $i \in C$ because its parents are in M by the definition of M itself. Similarly, if $i \in T$ and at least one of its parents (say, i') belongs to C, then there is a possibility that the other parents of i also belong to M (because i is linked to i' in the moral graph); therefore, the parents do not belong to J. In contrast, if $i \in R$, it may occur that the parents of i are in J.
Therefore, we can write

$$P_M(x_M) = \sum_{x_J \in \Omega_J} \left(\prod_{i \in C \cup T} P_{i|p(i)}(x_i|x_{p(i)}) \prod_{i \in R \cup J} P_{i|p(i)}(x_i|x_{p(i)}) \right),$$

which means

$$P_M(x_M) = \left(\prod_{i \in C \cup T} P_{i|p(i)}(x_i|x_{p(i)}) \right) \times \left(\sum_{x_J \in \Omega_J} \prod_{i \in R \cup J} P_{i|p(i)}(x_i|x_{p(i)}) \right). \tag{17}$$

The computation of $P_F(x_F)$

$$P_F(x_F) = \sum_{x_C \in \Omega_C} P_M(x_C, x_F).$$

In expression (17), the factor $\sum_{x_J \in \Omega_J} \left(\prod_{i \in R \cup J} P_{i|p(i)}(x_i|x_{p(i)}) \right)$ depends only on x_F because if $i \in R \cup J$, its parents are in $F \cup J$, which means they cannot belong to C (by the definition of T, all the children of elements of C are in $C \cup T$ and therefore are not in $R \cup J$). Due to expression (17),

$$P_F(x_F) = \left(\prod_{i \in C \cup T} P_{i|p(i)}(x_i|x_{p(i)}) \right) \times \left(\sum_{x_J \in \Omega_J} \prod_{i \in R \cup J} P_{i|p(i)}(x_i|x_{p(i)}) \right). \tag{18}$$

Proof of the necessary condition of the Theorem

Using the contrapositive, we show that if we have a graph $\mathscr{G}$ that does not have the pathology property and that is a subset of C, then we have

$$\sum_{x_C \in \Omega_C} \prod_{i \in C \cup T} P_{i|p(i)}(x_i|x_{p(i)}) = P_{T|R}(x_T|x_R). \tag{19}$$

Assume (see Lemma 1) that $R \cup J$ is an initial part; thus, we can prove equality (19).

Because $P_{T|R}(x_T|x_R) = \frac{P_F(x_F)}{P_R(x_R)}$, we still have to compute

$$P_R(x_R) = \sum_{x_T \in \Omega_T} P_F(x_T, x_R)$$
$$= \sum_{x_T \in \Omega_T} \left(\sum_{x_C \in \Omega_C} \prod_{i \in C \cup T} P_{i|p(i)}(x_i|x_{p(i)}) \right) \times \left(\sum_{x_J \in \Omega_J} \prod_{i \in R \cup J} P_{i|p(i)}(x_i|x_{p(i)}) \right).$$

Because $R \cup J$ is an initial part, the factor $\prod_{i \in R \cup J} P_{i|p(i)}(x_i|x_{p(i)})$ depends exclusively on $x_{R \cup J}$, and after summing out $x_J \in \Omega_J$, the expression $\sum_{x_J \in \Omega_J} \prod_{i \in R \cup J} P_{i|p(i)}(x_i|x_{p(i)})$ depends only on x_R and clearly does not depend on x_T. Therefore,

$$P_R(x_R) = \left(\sum_{x_{C \cup T} \in \Omega_{C \cup T}} \prod_{i \in C \cup T} P_{i|p(i)}(x_i|x_{p(i)}) \right) \times \left(\sum_{x_J \in \Omega_J} \prod_{i \in R \cup J} P_{i|p(i)}(x_i|x_{p(i)}) \right)$$
$$= \sum_{x_J \in \Omega_J} \prod_{i \in R \cup J} P_{i|p(i)}(x_i|x_{p(i)}). \tag{20}$$

It is clear that, for each subset A of I, $\sum_{x_A \in \Omega_A} \prod_{i \in A} P_{i|p(i)}(x_i|x_{p(i)}) = 1$.

Equality (20) implies the following:

$$P_{T|R}(x_T|x_R) = \frac{\left(\sum_{x_C \in \Omega_C} \prod_{i \in C \cup T} P_{i|p(i)}(x_i|x_{p(i)}) \right) \left(\sum_{x_J \in \Omega_J} \prod_{i \in R \cup J} P_{i|p(i)}(x_i|x_{p(i)}) \right)}{\left(\sum_{x_J \in \Omega_J} \prod_{i \in R \cup J} P_{i|p(i)}(x_i|x_{p(i)}) \right)}$$
$$= \sum_{x_C \in \Omega_C} \prod_{i \in C \cup T} P_{i|p(i)}(x_i|x_{p(i)}).$$

Proof of the sufficient condition of the theorem

Given $(j_0, \ldots j_n)$, a similar sequence to the one produced in Lemma 2, we will construct a BN with Boolean variables X_i where for each i, $\Omega_i = \{0, 1\}$ such that

$$\exists\, (x_T, x_R) \in \{0, 1\}^{card(F)} \; P_{T|R}(x_T|x_R) \neq \sum_{x_C \in \{0,1\}^C} \prod_{i \in C \cup T} P_{i|p(i)}(x_i|x_{p(i)}) \; .$$

To do so, if $i \notin \{j_1, \ldots j_n\}$, we consider, for $P_{i|p(i)}$, the equiprobability that is independent of the values of the conditional variables, in other words,

$$\forall\, x_{p(i) \in \{0,1\}^{card(p(i))}} P_{i|p(i)}(0|x_{p(i)}) = P_{i|p(i)}(1|x_{p(i)}) = \frac{1}{2}.$$

Thus, for each j_k (with $1 \leq k \leq n$), $P_{x_{j_k}|p(x_{j_k})}$ will depend on the conditioning variable $x_{j_{k-1}}$ (and only on it if j_k has parents other than j_{k-1}). We write

$$P_{j_k|p(j_k)}(x_{j_k}|x_{p(j_k)}) = a_{k,x_{j_k}}^{x_{j_{k-1}}}$$

with $a_{k,0}^{x_{j_{k-1}}} + a_{k,1}^{x_{j_{k-1}}} = 1$.

Therefore, $P_{j_k|p(j_k)}$ is characterized by the two numbers $a_{k,0}^0$ and $a_{k,0}^1$, $a_{k,1}^0 = 1 - a_{k,0}^0$ and $a_{k,1}^1 = 1 - a_{k,0}^1$.

We know that it is always true that $P_F(x_F) = A(x_F)B(x_F)$, where

$$A(x_F) = \sum_{x_C \in \{0,1\}^C} \prod_{i \in C \cup T} P_{i|p(i)}(x_i|x_{p(i)})$$

and

$$B(x_F) = \sum_{x_J \in \{0,1\}^J} \prod_{i \in R \cup J} P_{i|p(i)}(x_i|x_{p(i)}) \ .$$

Here, none of the indices j_k (with $1 \leq k \leq n$) belong to $C \cup T$ (because $j_n \in R$ and if $1 \leq k \leq n-1$, then $j_k \in J$). Therefore, in the expression of $A(x_F)$, all the factors are equal to $\frac{1}{2}$, and thus, each term of the sum is equal to $(\frac{1}{2})^{card(C \cup T)}$. Because each of these terms is repeated $2^{card(C)}$ times,

$$A(x_F) = (\frac{1}{2})^{card(T)}.$$

We now compute $B(x_F)$. Every term $\prod_{i \in R \cup J} P_{i|p(i)}(x_i|x_{p(i)})$ involves all the factors $P_{j_k|p(j_k)}(x_{j_k}|x_{p(j_k)})$ (where $1 \leq k \leq n$) together with $card(R) + card(J) - n$ factors that are equal to $\frac{1}{2}$. In the summation $\sum_{x_J \in \{0,1\}^{card(J)}}$, there are terms that are equal to

$$(\frac{1}{2})^{card(R)+card(J)-n} \prod_{1 \leq k \leq n} a_{k,x_{j_k}}^{x_{j_{k-1}}}$$

and each of these terms is repeated $2^{card(J)-(n-1)}$ times (because $n-1$ elements of the sequence $(j_1, \ldots j_n)$ belong to J). Therefore, if we denote $y_k = x_{j_k}$, for all k such that $0 \leq k \leq n$, we have the following expression:

$$B(x_F) = (\frac{1}{2})^{card(R)-1} \sum_{(y_1,\ldots y_{n-1}) \in \{0,1\}^{n-1}} \prod_{1 \leq k \leq n} a_{k,y_k}^{y_{k-1}}$$

which depends only on the pair (y_0, y_n) because the only elements of the sequence $(j_0, \ldots j_n)$ that belong to F are j_0 (in T) and j_n (in R).

Finally, by decomposing x_F into (x_R, x_T), we produce

$$P_F(x_R, x_T) = (\frac{1}{2})^{card(T)+card(R)-1} g(y_0, y_n)$$

where

$$g(y_0, y_n) = \sum_{(y_1,\ldots y_{n-1})\in\{0,1\}^{n-1}} \prod_{1\le k\le n} a_{k,y_k}^{y_{k-1}} .$$

By computing

$$P_R(x_R) = \sum_{x_T\in\{0,1\}^{T)}} P_F(x_R, x_T) ,$$

we find that

$$P_R(x_R) = (\frac{1}{2})^{card(R)}(g(0, y_n) + g(1, y_n)).$$

Therefore,

$$P_{T|R}(x_R|x_R) = (\frac{1}{2})^{card(T)-1} \frac{g(y_0, y_n)}{g(0, y_n) + g(1, y_n)}.$$

However, in this example,

$$\sum_{x_C\in\{0,1\}^C} \prod_{i\in C\cup T} P_{i|p(i)}(x_i|x_{p(i)}) \quad = 2^{card(C)}(\frac{1}{2})^{card(C)+card(T)} = (\frac{1}{2})^{card(T)} .$$

Thus, the equality

$$P_{T|R}(x_R|x_R) = \sum_{x_C\in\{0,1\}^{C)}} \prod_{i\in C\cup T} P_{i|p(i)}(x_i|x_{p(i)})$$

holds if and only if $\frac{g(y_0,y_n)}{g(0,y_n)+g(1,y_n)} = \frac{1}{2}$, which means that $g(0, y_n) = g(1, y_n)$.

Returning to the definition of the function g, our search for an opposite example consists of finding a family of pairs $(a_{k,0}^0, a_{k,1}^1)_{1\le k\le n}$ such that the expression below depends only on y_0 and y_n:

$$\sum_{(y_1,\ldots y_{n-1})\in\{0,1\}^{n-1}} \prod_{1\le k\le n} a_{k,y_k}^{y_{k-1}}.$$

For y_n fixed (for example, $y_n = 0$), the above expression takes different values for $y_0 = 0$ and $y_0 = 1$.

This property can easily be interpreted in terms of the sequence of random variables $(Y_0, \ldots, Y_n) = (X_{j_0}, \ldots, X_{j_n})$

Indeed, this example is a Markov chain (not necessarily homogeneous) of boolean random variables such that, for each k (where $1 \leq k \leq n$),

$$P_{k|k-1}(y_k|y_{k-1}) = a_{k,y_k}^{y_{k-1}} .$$

Then,

$$P_{n|0}(y_n|y_0) = \sum_{(y_1,\ldots y_{n-1})\in\{0,1\}^{n-1}} \prod_{1\leq k\leq n} a_{k,y_k}^{y_{k-1}}.$$

Therefore, we dispose of the example we have been looking for because $P_{n|0}(0|0) \neq P_{n|0}(0|1)$ (the random variables Y_0 and Y_n are not independent). An obvious and necessary condition for this example is that, for all k in $\{1, \ldots, n\}$, the random variables Y_k and Y_{k-1} are not independent. This implies that $a_{k,0}^0 \neq a_{k,0}^1$. Indeed, if there were k such that Y_k and Y_{k-1} are independent, the Markov chain $(Y_0, \ldots, Y_n)$ would be divided into two independent sub-channels $(Y_0, \ldots, Y_{k-1})$ and $(Y_k, \ldots, Y_n)$.

To check that this condition is sufficient or, in other words, for Y_0 and Y_n to be independent, it is necessary that there exists at least one $k \in \{1, \ldots, n\}$ such that Y_k and Y_{k-1} are independent. This property results from a basic computation. If $n = 2$, then we have the following:

$$a_{1,0}^0 a_{2,0}^0 + a_{1,1}^0 a_{2,0}^1 = a_{1,0}^1 a_{2,0}^0 + a_{1,1}^1 a_{2,0}^1.$$

In other words,

$$a_{1,0}^0 a_{2,0}^0 + (1 - a_{1,0}^0) a_{2,0}^1 = a_{1,0}^1 a_{2,0}^0 + (1 - a_{1,0}^1) a_{2,0}^1,$$

which is equivalent to $a_{1,0}^0 a_{1,0}^1$ or $a_{2,0}^0 a_{2,0}^1$.

For $n \geq 3$, we establish by induction that, for Y_0 and Y_n to be independent, it is necessary that we dispose of at least one of the independences of Y_0 and Y_1 from one side or of Y_1 and Y_n from the other side. Then, for Y_1 and Y_n to be independent, it is necessary that we dispose of at least one of the independences of Y_1 and Y_2 from one side or of Y_2 and Y_n from the other side, etc. □

Example 4 Recall Example 2. With $C = \{b\}$, $T(C) = \{c, e\}$, and $R(C) = \{a, d\}$, we conclude, using Theorem 1, that $\sum_b P(b|a)P(c|b)P(d|c)P(e|d, b) \neq P(c, e|d, b) = P(T(C)|R(C))$, (presence of a pathology on the BN).

4 Non-pathological Situations and Bayesian Networks of Level Two

It may be useful to ensure that no pathology may occur for any subset S. Although this condition is elementary, it is apparently strong enough to imply that there is no couple (i, i') that satisfies the following condition: i is a parent of i', and furthermore, there exists a path $(i_0, ...i_n)$ $(n \geq 3)$ for which none of the elements i_j $(1 \leq j \leq n-2)$ is a parent of i'.

Let us now recall the definition of Bayesian networks of level two before presenting their relationship with non-pathological BNs.

Let $\mathscr{I}$ be a partition of I, and let us consider a directed acyclic graph $\mathscr{G}$ on $\mathscr{I}$. The vertices of the graph $\mathscr{G}$ are the atoms J of the partition $\mathscr{I}$, in other words, each vertex J of $\mathscr{I}$ is a subset of indices of j of I. We say that there is a link from J' to J'' (where J' and J'' are atoms of the partition $\mathscr{I}$) if $(J', J'') \in \mathscr{G}$. If $J \in \mathscr{I}$, we denote $p(J)$ as the set of parents of J, that is, the set of J' such that $(J', J) \in \mathscr{G}$.

Definition 2 The probability P_I is defined by the Bayesian Network of level two, BN2, on I, $(\mathscr{I}, G, (P_{J|p(J)})_{J\in\mathscr{I}})$, if we have the conditional probability $P_{J|p(J)}$ for each $J \in \mathscr{I}$; in other words, the probability of X_J is conditioned on $X_{p(J)}$ (which, if $p(J) = \emptyset$, is the marginal probability P_J), so that

$$P_I(x_I) = \prod_{J\in\mathscr{I}} P_{J|p(J)}(x_J|x_{p(J)}).$$

A BN is a special case of level 2, with the partitioning of I into single vertices.

In some cases, it can be useful to remark that $P_{J/p(J)}$ depends only on a subset K of $\cup_{J'\in p(J)} J'$; this is equivalent to the concept of bubble graphs from Shafer [12]. (For additional details on Bayesian networks of level two, see [17], [15], [16]).

Example 5 The probability distribution P_I associated with the Bayesian network of level 2 in Fig. 4 can be written as

$$P_I(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8) = P_1(x_1)P_2(x_2)P_{3,|1,2}(x_3|x_1, x_2)P_{4,5|2}(x_4, x_5|x_2)$$
$$P_{6/3,4,5}(x_6|x_3, x_4, x_5)P_{7,8/3}(x_7, x_8|x_3).$$

Let Q_S be the set of parts of S composed of singletons $\{\ell\}$, where ℓ covers the set L of elements of S with no parents in $S^+ - S$, and of parts $T(C)$, where C covers the $S-$conditional partition $\mathscr{C}$. Note that one of these two types of elements of Q_S may be absent.

It has been shown (see [17], [16]) that Q_S as defined constitutes a partition of S. This results from Eq. (15) and from Theorem 1 stating that if the BN does not present the pathology property relative to any subset $C \in \mathscr{C}$, in other words if the BN is non-pathological, then

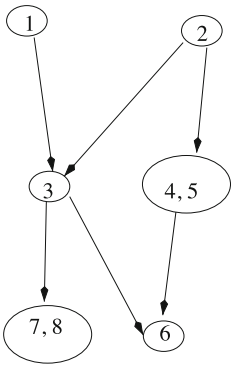


Fig. 4 Example of a Bayesian network of level two

$$P_S(x_S) = \left[\prod_{\ell \in L} P_{\ell|p(\ell)}(x_\ell | x_{p(\ell)})\right] \times \left[\prod_{C \in \mathscr{C}} P_{T(C)|R(C)}(x_{T(C)} | x_{R(C)})\right],$$

where $p(\ell)$ ($\ell \in L$) and $R(C)$ (where $C \in \mathscr{C}$) are subsets of S (which may be empty subsets).

Here, we see an expression of P_S that is similar to the probability distribution of a BN that can have none of the elements of S as vertices on its graph; however, the elements of the partition Q_S can be vertices of a Bayesian network of level 2 [16], as you can see in Example 6 below.

Example 6 Recall Example 2 with $S = \{a, c, d, e\}$. It has been shown in [16] that

$$\begin{aligned} P(S) &= \sum_b P(a)P(b|a)P(c|b)P(d|c)P(e|d,b) \\ &= P(a)\sum_b [P(b|a)P(c|b)P(d|c)P(e|d,b)] \\ &= P(a)P(c,d,e|a). \end{aligned}$$

We have a structure of a BN2 on S with two vertices $\{a\}$ and $\{c, d, e\}$. In contrast to Butz et al. computation in [2], the result is given as a product of probabilities and not as a simple potential in an extensive form.

However, to prove the above-mentioned identification, we should study in advance if Q_S is endowed with the structure of an oriented acyclic graph, denoted $\mathscr{G}'_S$, such that, for any ℓ, $p(\ell)$ belongs to the set of parents relative to $\mathscr{G}'_S$ of $\{\ell\}$ and, for any C, $R(C)$ is a subset of the set of parents relative to $\mathscr{G}'_S$ of $T(C)$. This preemptive study leads to the following definition of $\mathscr{G}'_S$:

- If $\ell \in L$ and $\ell' \in L$, then $(\{\ell\}, \{\ell'\}) \in \mathscr{G}'_S$ if $(\ell, \ell') \in \mathscr{G}$.
- If $\ell \in L$ and $C \in \mathscr{C}$, then $(\{\ell\}, T(C)) \in \mathscr{G}'_S$ if $\ell \in R(\mathscr{C})$.
- If $C \in \mathscr{C}$ and $C' \in \mathscr{C}$, then $(T(C), T(C')) \in \mathscr{G}'_S$ if $T(C) \cap R(C') \neq \varnothing$.

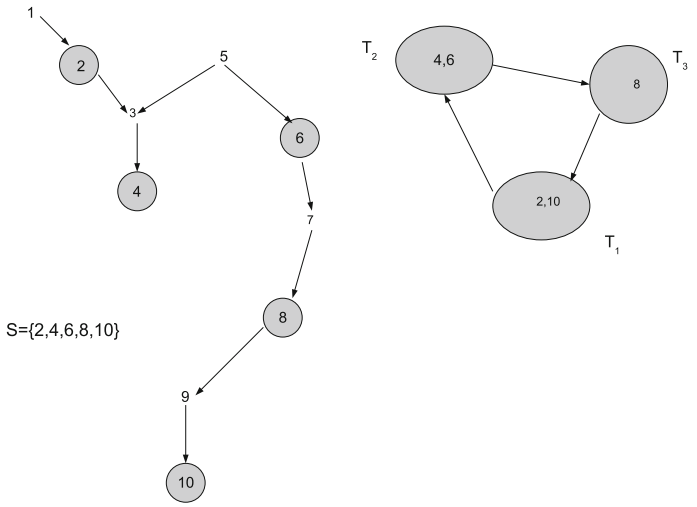


Fig. 5 Non-pathological graphs

However, we note that the oriented graph defined in this way on Q_S is not necessarily acyclic, as shown in the following example (Fig. 5), which is constructed from a non-pathological graph $\mathscr{G}$ with $S = \{2, 4, 6, 8, 10\}, L = \varnothing, C_1 = \{1, 9\}, C_2 = \{3, 5\}$, and $C_3 = \{7\}$.

In this case, we have the following: $T(C_1) = \{2, 10\}$, $T(C_2) = \{4, 6\}$, $T(C_3) = \{8\}$, $R(C_1) = \{8\}$, $R(C_2) = \{2\}$, and $R(C_3) = \{6\}$.

Although it is non-pathological, this graph presents a phenomenon with a "distant effect" (the arc from 1 to 9) similar to those effects that characterize the pathological graphs.

5 Conclusion

This work presented a method of interpreting factors used in a subset of BNs named non-pathological BNs. It also related non-pathological BNs to Bayesian networks of level two and presented results on the factorization of the marginal probability distribution of any subset of a BN as a product of conditional probability distributions; moreover, all intermediate computations can be written as conditional probability distributions. This has an advantage in keeping the probabilistic independence information available and thus technically usable; in addition, this information can be stored as a Bayesian network of level two and not in an extensive form as potential functions. Future work will include comparison of the pathological and topological

order conditions on more general BNs and the extension of this work to the case of evidence. We also propose to seek a characterization of the graphs $\mathscr{G}$ and the parts S such that the associated graph $\mathscr{G}'_S$ will be acyclic.

References

1. Butz, C.J., Yan, W.: The semantics of intermediate CPTs in variable elimination. In: Fifth European Workshop on Probabilistic Graphical Models (2010)
2. Butz, C.J., Yan, W., Madsen, A.L.: D-separation: strong completeness of semantics in Bayesian network inference. In: Twenty-sixth Canadian Conference on Artificial Intelligence (2013)
3. Butz, C.J., Yan, W., Madsen, A.L.: On semantics of inference in Bayesian networks. In: Symbolic and Quantitative Approaches to Reasoning with Uncertainty, Lecture Notes in Computer Science, vol. 7958, 73–84 (2013)
4. Castillo, E., Gutirrez, J.M., Hadi, A.S.: Expert Systems and Probabilistic Network Models. Springer, New York (1997)
5. Jensen, F.V.: An Introduction to Bayesian Networks. UCL Press, London (1996)
6. Jensen, F.V.: Bayesian Networks and Decision Graphs. Springer (2001)
7. Jensen, F.V., Lauritzen, S.L., Olesen, K.G.: Bayesian updating in causal probabilistic networks by local computations. Comput. Statis. Q. **4**, 269–282 (1990)
8. Koller, D., Friedman, N.: Probabilistic Graphical Models: Principles and Techniques. MIT press, Cambridge (2009)
9. Lauritzen, S.L., Spiegelhalter, D.J.: Local computation with probabilities on graphical structures and their application to expert systems. J. Roy. Statis. Soc. **50**, 157–194 (1988)
10. Pearl, J.: Fusion, propagation and structuring in belief networks. Artif. Intell. **29**, 241–288 (1986)
11. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers Inc., San Francisco, CA (1988)
12. Shafer, G.: Probabilistic expert system. In: CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 67. SIAM (1996)
13. Shenoy, P.P.: Binary join trees for computing marginals in the Shenoy-Shafer architecture. Int. J. Approx. Reason. V **17**(1), 1–25 (1997)
14. Shenoy, P.P., Shafer, G.: Axioms for probability and belief-function propagation. In: Shachter, R.D., Levitt, T.S., Lemmer, J.F., Kanal, L.N. (eds.) Uncertainty in Artificial Intelligence, vol. 4, pp. 169-198 (1990)
15. Smail, L.: D-separation and level two Bayesian networks. Artif. Intell. Rev. **31**(1–4), 87–99 (2005). doi:10.1007/s10462-009-9128-3
16. Smail, L.: Uniqueness of the level two bayesian network representing a probability distribution. Int. J. Math. Math. Sci. ID **845398**, (2011)
17. Smail, L., Raoult, JP.: Successive restrictions algorithm in Bayesian networks. LNCS 3646. Springer-Verlag Berlin Heidelberg, pp. 409–418 (2005)
18. Studeny, M.: Probabilistic Conditional Independent Structures. Springer (2005)

Optimal Drug Treatment in a Simple Pandemic Switched System Using Polynomial Approach

Abdessamad Tridane, Mohamed Ali Hajji and Eduardo Mojica-Nava

Abstract The aim of this work is to investigate the optimal control of the treatment in a simple pandemic model as a switched nonlinear system. We used a newly developed approach based on the theory of moments. This approach allows to transform a nonlinear, non-convex optimal control problem to an equivalent linear and convex one. To illustrate our finding, we used the example of influenza pandemic to compare the full treatment approach to our optimal moment and time switching solution.

Keywords Switched systems · Epidemic model · Optimal control · Theory of moment

AMS subject classifications [2010] Primary 93C10; Secondary 93D15, 34A38

1 Introduction

In order to avoid high mortalities and as a result of the severity of these outbreaks over years, humans have focused their efforts on finding the best strategies to control the spread of infectious diseases.

Due to poor planning, these efforts frequently fall short. For example, the supplies of drug treatments are often inadequate and inefficient, causing health facilities to run out of resources before meeting the needs [28].

A. Tridane (✉) · M.A. Hajji
United Arab Emirates University, Department of Mathematical Sciences,
College of Science, United Arab Emirates University, P.O. Box 15551, Al Ain,
United Arab Emirates
e-mail: a-tridane@uaeu.ac.ae

M.A. Hajji
e-mail: mahajji@uaeu.ac.ae

E. Mojica-Nava
Department of Electrical and Electronics Engineering,
Universidad Nacional de Colombia, Bogota, Colombia
e-mail: eamojican@unal.edu.co

T. Abualrub et al. (eds.), *Mathematics Across Contemporary Sciences*,
Springer Proceedings in Mathematics & Statistics 190,
DOI 10.1007/978-3-319-46310-0_14

For these reasons, the optimization of the existing control resources is a continuous concern in the public health. The optimal control theory has been a powerful approach to solve optimality problems in many disciplines. The majority of the techniques used in optimal control disease outbreak models (see e.g. [25]) are based on the Pontryagin maximum principle [20] and forward-backward numerical algorithms to solve the state and adjoint system of equations (for the use of numerical methods in optimal control of epidemiological models see [2, 16] and for all other types of optimal control problems see [24]).

One of the issues in using the standard optimal control approach is the suggested control might not take into consideration some realistic constraints. For example, the control agent could be a drug treatment that is not necessarily available at all times [15] or simply might run-out [1]. In this case, it is clear that assuming that control agent to be a continuous function is too optimistic of an assumption. In this situation, a switched control system would be a better approach to deal with a control problem of this nature.

Switched control systems are a class of hybrid systems that are composed of a number of subsystems which are defined by the switches [17]. These systems have extensively been used in recent years due to their applications in engineering and other disciplines [5, 29]. Hence, different studies have adapted the maximum principle to find the optimal control switched systems [9, 23, 26, 27]. However, one the biggest problems of these versions of maximal principle of switched systems is that they are numerically expensive since they involve the use of mixed integer programming [3, 4].

The recent work of Mojica-Nava et al. [19] introduced a new approach to finding the optimal switched control system. This approach is based on the use of the theory of moments for global polynomial optimization via semidefinite programming, which eases the numerical burden of using the previous approaches.

In this work, we use this new approach [19] in a simple classical model that reflects the switched aspect of the control in a pandemic model. A similar version of this model without switched control can be found in a recent work of Brauer [6, 7]. Our goal is to minimize the outflow from infected classes.

The paper is organized as follows: We introduce the problem in Sect. 2 then we transform the problem to an optimal control of switched system in Sect. 3. In Sect. 4, we present relaxation using moments approach where we transform the nonlinear, non-convex optimal control problem to a linear and convex one. Issues related to the implementation of the proposed method are presented in Sect. 5. Finally, we illustrate our results with simulation in Sect. 6. We draw conclusions and present discussion of our findings in Sect. 6.

2 Problem Statement

The aim of this work is to investigate the effect of treatment run-out supplies during a pandemic. For this reason, we consider the following simple pandemic treatment model

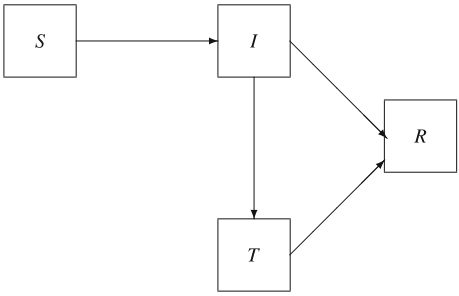


Fig. 1 Flow diagram of transitions between epidemiological classes of the STIR model

$$\begin{aligned}
\dot{S} &= -\beta \frac{S}{N}(I + \delta T), \\
\dot{I} &= \beta \frac{S}{N}(I + \delta T) - (\alpha + \gamma u)I, \\
\dot{T} &= \gamma u I - \eta T,
\end{aligned} \tag{1}$$

where the total population N is defined as $N = N(t_0) = S(t_0) + I(t_0)$. This SITR model is an extension of the standard model of Kermack and Mckendrick by adding a fraction of infectives to be treated [6, 7], where S, I and T represent the susceptible, the infected and treated individuals. The parameters in model (1) are defined as follows: β is the transmission rate from susceptible to infected host and α is the per capita loss rate of infected individuals through both mortality and recovery. We assume that individuals in the T class have infectivity reduced by a factor δ and γ is the rate of infectives that are treated. We also assume that the rate of removal from treated class is η.

The flow chart of our model is given in Fig. 1.

The basic reproduction number is calculated in [6] as

$$\mathscr{R}_0 = \frac{\beta}{\alpha + \gamma u} + \frac{\beta \delta \gamma u}{\eta(\alpha + \gamma u)} \tag{2}$$

and the final size is given by [7]

$$N - S_\infty = (\alpha + \gamma u) \int_0^\infty I(t)dt. \tag{3}$$

The parameter u in (1) is the treatment control which takes values 0 or 1, where $u = 1$ means treatment is underway and $u = 0$ means no treatment. Accordingly, system (1) can switch between two different subsystems (modes of operations), corresponding to $u = i,\ i = 0, 1$, as time progresses. Thus, we have a switched system.

3 The Switched System

By considering $x = (S, I, T)^\top$, we can rewrite our system (1) as

$$\dot{x}(t) = f_{\sigma(t)}(x(t)), \tag{4}$$

where $f_i : I\!R^3 \to I\!R^3$ is the ith vector field, $\sigma : [t_0, t_f] \to \mathcal{Q} = \{0, 1\}$ is the switching signal, a piecewise constant function of time, and $[t_0, t_f]$ is the time interval under consideration. The initial conditions are given by $x(0) = (S(0), I(0), T(0))^\top$. Every mode of operation of the system corresponds to a specific subsystem $\dot{x}(t) = f_i(x(t))$, for each $i \in \mathcal{Q}$, where $i = 0$ corresponds to $u = 0$ and $i = 1$ corresponds to $u = 1$.

Our goal is to study the number of possible switches of treatment that allows us to reduce the burden of the infection by reducing the size of the pandemic. Each subsystem $\dot{x}(t) = f_i(x(t))$, for $i \in \mathcal{Q}$, corresponds to a mode of the switching signal $\sigma(t)$ which is our control input. The values of the switching input signal must be chosen in a way to satisfy the given initial conditions and the desirable final conditions that represent specific desirable pandemic outcome. The optimal switching signal σ would represent a public health optimal strategy to control the pandemic within the limits of the available resources.

Before we continue our analysis, we assume the following [19]:

- There are no infinite switching accumulation points in time.
- The state does not have jump discontinuities.

Accordingly, we define the switched control of our system as a duplet of finite sequence of modes and a finite sequence of switching times $t_0 < t_1 < ... < t_f$.
Our optimal control cost function is defined, in Bolza form, as the functional

$$J = \int_{t_0}^{t_f} L_{\sigma(t)}(t, x(t))dt, \tag{5}$$

with the running cost $L_{\sigma(t)}(t, x(t))$ is given by

$$L_{\sigma(t)}(t, x(t)) = \alpha I(t) + \eta T(t), \tag{6}$$

where the term $(\alpha I(t) + \eta T(t))$ represents the outflow from infected classes at time t [12], and $\sigma(t) \in \{0, 1\}$.

The switched optimal control problem becomes

$$\min_{\sigma(t)} J(t_0, t_f, x(t), \sigma(t)) \tag{7}$$

subject to

$$\dot{x}(t) = f_{\sigma(t)}(x(t)), \tag{8}$$

where J is defined by (5) and $\sigma(t) \in \{0, 1\}$.

Following the approach in [19], we use Lagrange polynomials to transform the system (8) to a continuous non-switched control system. Therefore, we introduce a new *continuous* control variable $w \in \Omega = \{w \in \mathbb{R} \,|\, g(w) = 0\}$, where

$$g(w) = w(w-1). \tag{9}$$

Let the kth Lagrange polynomial, $l_k(w),\ k = 0, 1$, be defined by

$$l_0(w) = (1-w), \quad l_1(w) = w. \tag{10}$$

Then, according to Proposition 4 in [19], we can write system (8) as the following equivalent continuous system with polynomial dependence, $\mathscr{F}(x, w)$, in the new control variable $w \in \Omega$:

$$\dot{x} = \mathscr{F}(x, w) = f_0(x)l_0(w) + f_1(x)l_1(w). \tag{11}$$

Similarly, the running cost $L_{\sigma(t)}(t, x(t))$ is equivalently represented by the polynomial $\mathscr{L}(x, w)$ of degree 1 in w:

$$\mathscr{L}(x, w) = L_0(t, x(t))l_0(w) + L_1(t, x(t))l_1(w) \tag{12}$$

and the cost function J in (5) becomes

$$J(t_0, t_f, x(t), w) = \int_{t_0}^{t_f} \mathscr{L}(x, w)dt. \tag{13}$$

Finally, the polynomial equivalent optimal control problem (PEOCP) can be stated as

$$\min_{w \in \Omega} J(t_0, t_f, x(t), w) \tag{14}$$

subject to

$$\dot{x} = \sum_{k=0}^{1} f_k(x)l_k(w), \tag{15}$$

with $x(0) = x_0$, a given initial state. The polynomial constraint $w \in \Omega$ $(g(w) = 0)$ makes the problem nonconvex, as the feasible set Ω is non convex. To overcome the nonconvexity of the problem, the moments approach, described in the next section, is used to redefine the PEOCP in terms of moment variables which will render the optimization problem convex.

4 Relaxation Using Moments Approach

In this section, the moments approach is described for the relaxation of the PEOCP which transforms the nonconvex PECOP into convex semidefinite programs (SDPs). This approach is based on the concepts of moment and localizing matrices of probability measures supported in Ω.

4.1 Moment and Localizing Matrices

The concept of moment and localizing matrices of a probability measure is described in details in [13, 14]. For the convenience of the reader, we report only the important aspects.

Let $\mathscr{P}_r$ be the space of univariate polynomials of degree at most r in the variable $x \in \mathbb{R}$. If μ is a probability measure supported in some set $A \subset \mathbb{R}$, the ith moment of μ is defined as

$$m_i = \int_A x^i \mu(dx)$$

with $m_0 = 1$. If $p(x) \in \mathscr{P}_r$ of degree r, $p(x) = \sum_{i=0}^{r} p_i x^i$, then

$$\int_A p_r(x)\mu(dx) = \sum_{i=0}^{r} p_i m_i.$$

Now, if $m = \{m_j\}_{j=0}^{2r}$ is a sequence a moments of some probability measure μ, the moment matrix $M_r(m)$ is defined as the symmetric $(r+1) \times (r+1)$ matrix with (i, j) entries $M_r(m)(i, j) = m_{i+j}$, $0 \le i, j \le r$, i.e.,

$$M_r(m) = \begin{bmatrix} m_0 & m_1 & \cdots & m_r \\ m_1 & m_2 & \cdots & m_{r+1} \\ \vdots & \vdots & \cdots & \vdots \\ m_r & m_{r+1} & \cdots & m_{2r} \end{bmatrix}.$$

The localizing matrix relative to a polynomial $q(x)$ is defined as follows. Given a polynomial $q(x)$ of degree s, $q(x) = \sum_{i=0}^{s} q_i x^i$, the localizing matrix denoted by $M_r(q\, m)$ is defined as the symmetric matrix of size $(r+1) \times (r+1)$ with (i, j) entries, $0 \le i, j \le r$, given by

$$M_r(q\,m)(i,\,j) = \sum_{k=0}^{s} q_k m_{i+j+k}.$$

As an example, in the case of $r = 3$, the moment matrix $M_3(m)$ is

$$M_3(m) = \begin{bmatrix} m_0 & m_1 & m_2 & m_3 \\ m_1 & m_2 & m_3 & m_4 \\ m_2 & m_3 & m_4 & m_5 \\ m_3 & m_4 & m_5 & m_6 \end{bmatrix}.$$

If $q(x) = 2 - x^2, \{q_i\} = \{2, 0, -1\}$, the $(i,\,j)$ entries of the localizing matrix is given by $M_3(q\,m)(i,\,j) = 2m_{i+j} - m_{i+j+2}$, i.e.,

$$M_3(q\,m) = \begin{bmatrix} 2m_0 - m_2 & 2m_1 - m_3 & 2m_2 - m_4 & 2m_3 - m_5 \\ 2m_1 - m_3 & 2m_2 - m_4 & 2m_3 - m_5 & 2m_4 - m_6 \\ 2m_2 - m_4 & 2m_3 - m_5 & 2m_4 - m_6 & 2m_5 - m_7 \\ 2m_3 - m_5 & 2m_4 - m_6 & 2m_5 - m_7 & 2m_6 - m_8 \end{bmatrix}.$$

A key property of $M_r(m)$ and $M_r(q\,m)$ used in this paper is their positive semidefiniteness stated in the following proposition.

Proposition 1 *If $m = \{m_i\}$ is a sequence of moments of some probability measure μ supported in some set $A \subset \mathbb{R}$ and $q(x) = \sum_k q_k x^k$ is a polynomial with $q(x) \geq 0$, $\forall\, x \in A$, then the matrices $M_r(m)$ and $M_r(q\,m)$ are positive semidefinite.*

Proof Let $c = (c_0, c_1, \ldots, c_r) \in \mathbb{R}^{r+1}$. Let $p(x) = \sum_{i=0}^{r} c_i x^i$. Then

$$cM_r(m)c^\top = \sum_{i=0}^{r}\sum_{j=0}^{r} c_i c_j m_{i+j} = \sum_{i=0}^{r}\sum_{j=0}^{r} c_i c_j \int_A x^{i+j}\mu(dx) = \int_A (p(x))^2\mu(dx) \geq 0.$$

$$\begin{aligned} cM_r(q\,m)c^\top &= \sum_{i=0}^{r}\sum_{j=0}^{r}\sum_{k} c_i c_j q_k m_{i+j+k} \\ &= \sum_{i=0}^{r}\sum_{j=0}^{r}\sum_{k} c_i c_j q_k \int_A x^{i+j+k}\mu(dx) = \int_A q(x)(p(x))^2\mu(dx) \geq 0. \end{aligned}$$

which prove the positive semidefiniteness of $M_r(m)$ and $M_r(q\,m)$, since c is an arbitrary vector in $\mathbb{R}^{r+1}$.

4.2 Semidefinite Programs Using Moments Approach

It has been shown in [13], see also [14], that the minimisation problem (14) is equivalent to the minimisation problem

$$\min_{\mu \in P(\Omega)} \int_{\Omega} J \, \mu(dw), \tag{16}$$

that is,

$$\min_{w \in \Omega} J = \min_{\mu \in P(\Omega)} \int_{\Omega} J \mu(dw)$$

where $P(\Omega)$ is the space of probability measures supported in Ω. Since J is a polynomial of degree 1 in w (see (12) and (13)), we can rewrite the minimisation problem in terms of the moments of μ as

$$\min_{\mu \in P(\Omega)} \int_{\Omega} J \mu(dw) = \min_{m \in \mathscr{M}} \int_{t_0}^{t_f} \sum_{k=0}^{1} \sum_{i=0}^{1} L_k(t, x(t)) \alpha_{ki} m_i, \tag{17}$$

where $\alpha_{00} = 1, \alpha_{01} = -1, \alpha_{10} = 0, \alpha_{11} = 1$, are the coefficients of the Lagrange polynomials $l_0(w)$ and $l_1(w)$ in (10), and $\mathscr{M}$ is the space of moments defined by

$$\mathscr{M} = \{m = \{m_i\}, \ m_i = \int_{\Omega} w^i \mu(dw), \mu \in P(\Omega)\}.$$

The system state, Eq. (15), is rewritten in terms of the moments m_i as

$$\dot{x} = \sum_{k=0}^{1} \sum_{i=0}^{1} f_k(x) \alpha_{ki} m_i. \tag{18}$$

The constraint $m \in \mathscr{M}$ in (17) states that m is a vector of moments of some probability measure. This implies that

$$M_1(m) = \begin{bmatrix} m_0 & m_1 \\ m_1 & m_2 \end{bmatrix} \succeq 0.$$

The constraint on the control variable $w \in \Omega$, $g(w) = w(w-1) = w^2 - w = 0$, is written as two inequality constraints:

$$g_1(w) = g(w) = w^2 - w \geq 0, \tag{19}$$

$$g_2(w) = -g(w) = w - w^2 \geq 0. \tag{20}$$

The degree of both constraint functions g_1 and g_2 is even (=2), so following the results in [13], we consider the family of relaxed convex SDPs, with relaxation order $r \geq \max(\text{degree}(g_1)/2, \text{degree}(g_2)/2) = 1$:

$$\text{SDP}_r : \begin{cases} \min\limits_{m} \displaystyle\int_{t_0}^{t_f} \sum_{k=0}^{1}\sum_{i=0}^{1} L_k(t, x(t))\alpha_{ki} m_i \, dt, \\ \\ M_r(m) \succeq 0, \\ M_{r-1}(g_1\, m) \succeq 0, \\ M_{r-1}(g_2\, m) \succeq 0, \\ \dot{x} = \displaystyle\sum_{k=0}^{1}\sum_{i=0}^{1} f_k(x)\alpha_{ki} m_i. \end{cases} \tag{21}$$

It was shown also in [13] that min SDP_r is an increasing sequence of lower bounds for min J, and as $r \longrightarrow \infty$, min $\text{SDP}_r \uparrow \min J$.

For the lowest order of relaxation $r = 1$, we have the following SDP

$$\text{SDP}_1 : \begin{cases} \min\limits_{m} \displaystyle\int_{t_0}^{t_f} \sum_{k=0}^{1}\sum_{i=0}^{1} L_k(t, x(t))\alpha_{ki} m_i \, dt, \\ \\ M_1(m) \succeq 0, \\ M_0(g_1\, m) = m_2 - m_1 \succeq 0, \\ M_0(g_2\, m) = m_1 - m_2 \succeq 0, \\ \dot{x} = \displaystyle\sum_{k=0}^{1}\sum_{i=0}^{1} f_k(x)\alpha_{ki} m_i. \end{cases} \tag{22}$$

It is worth mentioning that one can use a higher relaxation order r but the number of moment variables will increase, which can make the problem numerically inefficient. However, it is found that in many situations the lowest order of relaxation can achieve the optimal value. In our simulation, we treat our problem with the lowest order of relaxation $r = 1$.

5 Numerical Implementation

In this section, we explain the numerical implementation steps used to solve (22). The SDP in (22) is a constrained minimisation problem over the moments $m(t)$ which are time dependent. We discretize the interval $[t_0, t_f]$ with nodal points t_i, $i = 0, 1, \ldots, N$, with $t_N = t_f$, using a uniform step h. Denote by $\mathbf{m}_i$ the vector of the ith moments, i.e., $\mathbf{m}_i = \{m_i(t_j)\}_{j=0,1,\ldots,N-1}$. Note that $\mathbf{m}_0 = [1,\ 1, \ldots,\ 1]$, since the zeroth moment $m_0 = 1$ for all t. Let the vector $\mathbf{m} = [\mathbf{m}_1\ \mathbf{m}_2]$ of length $2N$.

The integral defining the objective function and the state constraint differential equation in (22) are descritized using appropriate quadratures. A trapezoidal rule quadrature for the integral and a one-step forward discretization of the state equation give the following discrete version of (22):

$$\text{SDP}_1 : \begin{cases} \min\limits_{\mathbf{m}} h \sum\limits_{j=0}^{N-1} \sum\limits_{k=0}^{1} \sum\limits_{i=0}^{1} L_k(t_j, x(t_j))\alpha_{ki} m_i(t_j) \\ \\ M_1(\mathbf{m}) \succeq 0, \\ M_0(g_1\, \mathbf{m}) = \mathbf{m}_2 - \mathbf{m}_1 \succeq 0, \\ M_0(g_2\, \mathbf{m}) = \mathbf{m}_1 - \mathbf{m}_2 \succeq 0, \\ x(t_{j+1}) = x(t_j) + h \sum\limits_{k=0}^{1} \sum\limits_{i=0}^{1} f_k(x(t_j))\alpha_{ki} m_i(t_j). \end{cases} \tag{23}$$

where the minimisation is now over the vector $\mathbf{m} = [\mathbf{m}_1\ \mathbf{m}_2]$.

Problem (23) is solved using the built in Matlab function **fmincon**, which is a function designed for solving numerically nonlinear constrained minimization problems.

Once an optimal solution $\mathbf{m}^*(t_j) = (m_1^*(t_j), m_2^*(t_j))$ of (23) is reached, the switching signal $\sigma(t_j)$ is determined using a rank condition [19] as follows. If $\text{rank}(M_1(\mathbf{m}^*(t_j))) = \text{rank}(M_0(\mathbf{m}^*(t_j))) = 1$, then the optimal switching signal at t_j is $\sigma(t_j) = m_1(t_j)$, otherwise we use a sum up rounding procedure [21] as follows

$$\sigma(t_j) = \begin{cases} \lceil m_1(t_j) \rceil & \text{if } \int_{t_0}^{t_j} m_1(\tau) d\tau - h \sum\limits_{k=0}^{j-1} \sigma(t) k) \geq 0.5h, \\ \lfloor m_1(t_j) \rfloor & \text{otherwise,} \end{cases} \tag{24}$$

where $\lceil \cdot \rceil$ and $\lfloor \cdot \rfloor$ are the ceiling and floor functions, respectively.

6 Numerical Simulation

To illustrate our model, we need to simulate the results of our analysis. We choose the influenza pandemic as parameters of our model. The following parameters can be found in different papers that have studied the influenza pandemic. In our case, we used the parameters for models that studied the control strategy via vaccination and treatment [8, 10, 11]. The parameters are presented in Table 1.

The simulations illustrated in Fig. 2 describe the plot of the relaxed moment solution function $m_1^*(t_j)$ and the switching signal $\sigma(t_j)$. Figure 3 depicts the time series of the three compartments' populations considered in the model.

The optimal moment function and optimal switching signal showed that range of the switches corresponding to optimal solution is between $t = 0$ to $t = 26$ time

Table 1 Parameter estimation

Parameter	Description	Value	References
β	Transmission rate (days^{-1})	1.03–2.75	[8]
δ	Relative infectiousness of the asymptomatic class	0.5	[8, 10]
α	Mortality rate (days^{-1})	0.01	[11]
γ	Diagnostic rate (days^{-1})	0.5	[8]
η	Recovery rate for hospitalized class (days^{-1})	0.51	–
u	Control treatment on hospitalized individuals	0,1	–
$S(0)$	Initial number of susceptible individuals	174673	[8]
$I(0)$	Initial number of infectious individuals	132	[8]
$T(0)$	Initial number of infectious individuals	0	Assumed

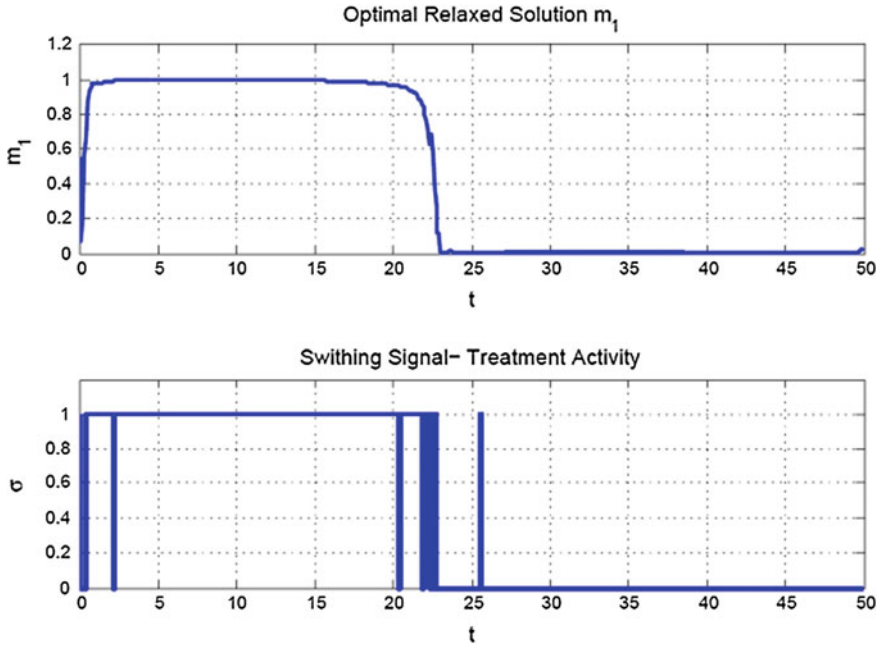


Fig. 2 The moment function $m_1(t)$ (*top*) and the optimal switched control of the treatment $\sigma(t)$ (*bottom*)

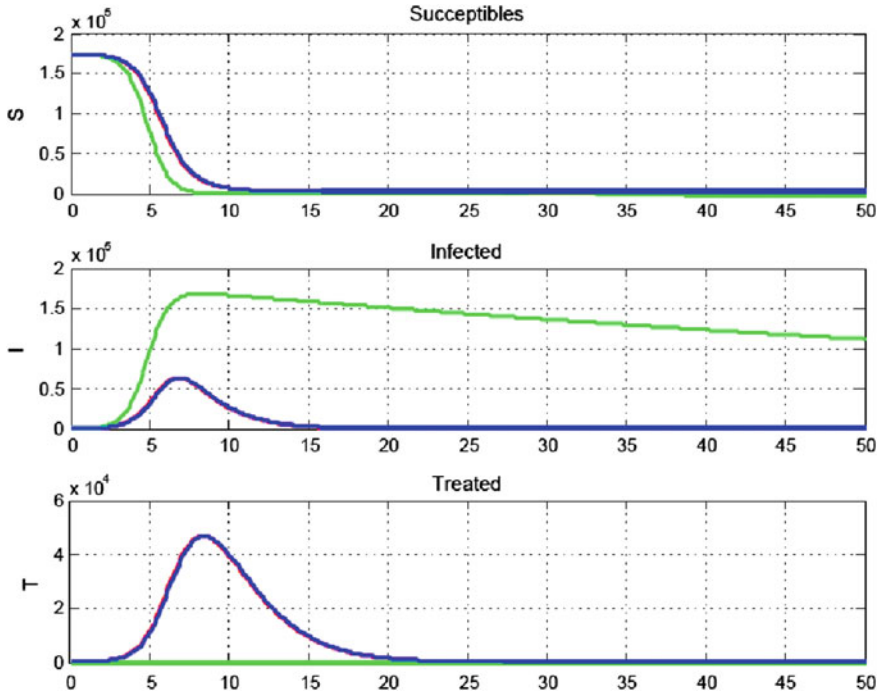


Fig. 3 The time series of the Susceptible, Infected and Treated. The plot displays these variable with no control $u = 0$ (*green*), optimal switched control (*blue*), and full control $u = 1$ (*red*)

units (which is days in the case of influenza) to treat the infected population. After that time there is no need for treatment. This finding is reflected on the time series in Fig. 3 where the peak size and the peak time of the pandemic curve of $u = 1$ (in red), which represent the full availability of the treatment at any given time, completely match the case of optimal switched control. This shows that we can achieve the same outcome of controlling the pandemic antiviral treatment by only treating (on and off) for a a limited period of time, hence, avoiding the consequences of the long term antiviral treatment which may exhaust drug stockpile and may develop drug resistance.

7 Conclusion

In this paper, we studied the optimal control problem of a SITR model. The control aimed to optimize the number of infected and treated population via only one control agent, i.e. the treatment. The method used for solving the optimal control problem of switched nonlinear systems was based on a polynomial approach developed by Mojica-Nava et al. [19]. The method based on transforming the problem into a

polynomial system which was transformed into a relaxed convex problem using the method of moments [13].

Our results showed that, by using this approach, we can achieve the same outcome of continuous treatment by only limiting treatment for period of time. This indicates that if the treatment is not available or run-out after a specific time, the outcome of the pandemic would be the same as if treatment is available at all times. It is important to mention that the suggested control switches are all in the early time of the pandemic, which line-up with the results in [1, 10]. Although the antiviral drug in a pandemic, like influenza, has been considered as the first line of the defence [1], the long term use of this drug could lead to the development of drug-resistance. This finding also suggests a solution to the long and extensive use of the antiviral drug by limiting its use (on and off) in the beginning of the pandemic and for a limited period of time.

The model suggested in this work is very simple and does not include other control strategies such as vaccination and isolation, that are used to protect the public health in the case of pandemics such as influenza. Our next step is to include these defence measures as switched control in more extended models that include all different levels of heterogeneities.

Acknowledgements The authors would like to thank the reviewer for the valuable comments and suggestions which help improve the quality of this work. The research of A.T and M.A.H is supported by the College of Sciences individual grant at United Arab Emirates University.

References

1. Arino, J., Bowman, C.S., Moghadas, S.M.: Antiviral resistance during pandemic influenza: implications for stockpiling and drug use. BMC Infect. Dis. **9**(8) (2009). doi:10.1186/1471-2334-9-8
2. Behncke, H.: Optimal control of deterministic epidemics. Optim. Control Appl. Methods **21**, 269–285 (2000)
3. Bemporad, A., Morari, M.: Control of systems integrating logic, dynamics, and constraints. Automatica **35**(3), 407–427 (1999)
4. Bemporad, A., Morari, M., Dua, V., Pistikopoulos, E.N.: The explicit solution of model predictive control via multiparametric quadratic programming. Proc. Am. Control Conf. Chicago **872–876** (2000)
5. Bengea, S., DeCarlo, R.: Optimal control of switching systems. Automatica **41**, 11–27 (2005)
6. Brauer, F.: General compartmental epidemic models. Chin. Ann. Math. **31B**(3), 289–304 (2010). doi:10.1007/s11401-009-0454-1
7. Brauer, F.: Age of infection epidemic models. MTBI May 2015
8. Chowell, G., Ammon, C.E., Hengartner, N.W., Hyman, J.M.: Transmission dynamics of the great influenza pandemic of 1918 in Geneva, Switzerland: assessing the effects of hypothetical interventions. J. Theor. Biol. **241**(2), 193–204 (2006)
9. Dmitruk, A.V., Kaganovich, A.M.: The hybrid maximum principle is a consequence of Pontryagin maximum principle. Syst. Control Lett. **57**(11), 964–970 (2008)
10. Feng, Z., Towers, S., Yang,Y.: Modeling the effects of vaccination and treatment on pandemic influenza. AAPS J. 13(3) (2011)
11. Gani, R., Hughes, H., Fleming, D., Grifin, T., Medlock, J., Leach, S.: Potential impact of antiviral use during influenza pandemic. Emerg. Infect. Dis. **11**, 1355–1362 (2005)

12. Jaberi-Doraki, M., Moghadas, S.M.: Optimal control of vaccination dynamics during an influenza epidemic. Math. Biosci. Eng. **11**, 5 (2014)
13. Lasserre, J.: Global optimization with polynomials and the problem of moments. SIAM J. Optim. **11**(3), 796–817 (2001)
14. Lasserre, J.: Semidefinite programming vs LP relaxations for polynomial programming. Math. Oper. Res. **27**(2), 347–360 (2002)
15. Lee, S., Chowell, G., Castillo-Chávez, C.: Optimal control for pandemic influenza: the role of limited antiviral treatment and isolation. J. Theor. Biol. **265**(2), 136–150 (2010)
16. Lenhart, S., Workman, J.T.: Optimal Control Applied to Biological Models. Chapman and Hall/CRC (2007)
17. Liberzon, D.: Switching in Systems and Control. Systems & Control: Foundations and Applications series. Birkhauser, Boston (2003)
18. Longini Jr., I.M., Halloran, M.E., Nizam, A., Yang, Y.: Containing pandemic influenza with antiviral agents. Am. J. Epidemiol. **159**(7), 623–633 (2004)
19. Mojica-Nava, E., Quijano, N., Rakoto-Ravalontsalama, N.: A polynomial approach for optimal control of switched nonlinear systems. Int. J. Rub. Nonlinear. Control **24**, 1797–1808 (2014)
20. Pontryagin, L.S., Boltyanskii, V.G., Gamkrelidze, R.V., Mishchenko, E.F.: The Mathematical Theory of Optimal Processes. Translated by K. N. Trirogoff. Classics of Soviet Mathematics. Original. Gordon & Breach Science Publishers, New York, NY (1961)
21. Sager, S.: Reformulations and algorithms for the optimization of switching decisions in nonlinear optimal control. J. Process Control **19**(8), 1238–1247 (2009)
22. Sunmi, L., Gerardo, C., Castillo-Chavez, C.: Optimal control for pandemic influenza: the role of limited antiviral treatment and isolation. J. Theoret. Biol. **265**, 136–150 (2010)
23. Sussmann, H.: A maximum principle for hybrid optimal control problems. In: Proceedings of the 38th IEEE Conference on Decision and Control, Phoenix, vol. 1, pp. 425-430 (1999)
24. Rao, A.V.: A survey of numerical methods for optimal control. Adv. Astron. Sci. **135**(1), 497–528 (2009)
25. Rachah, A., Torres, D.F.M.: Mathematical Modelling, Simulation, and Optimal Control of the 2014 Ebola Outbreak in West Africa. Discrete Dynamics in Nature and Society Volume 2015, Article ID 842792, 9 pages, 2015, http://dx.doi.org/10.1155/2015/842792
26. Riedinger, P., Daafouz, J., Iung, C.: Suboptimal switched controls in context of singular arcs. In: Proceedings of the 42nd IEEE Conference on Decision and Control, Hawaii, pp. 6254–6259 (2003)
27. Shaikh, M.S., Caines, P.E.: On the hybrid optimal control problem: theory and algorithms. IEEE Trans. Autom. Control **52**(9), 1587–1603 (2007)
28. Shillcutt, S., Morel, C., Goodman, C., Coleman, P., Bell, D., Whitty, C.J.M., Mills, A.: Cost effectiveness of malaria diagnostic methods in sub-saharan africa in an era of combination therapy. Bull. World Organ. **86**(2), 101–110 (2008)
29. Xu, X., Antsaklis, P.J.: Optimal control of switched systems based on parameterization of the switching instants. IEEE Trans. Auto. Cont. **49**(1) (2004)

On Carathéodory Quasilinear Functionals for BV Functions and Their Time Flows for a Dual H^1 Penalty Model for Image Restoration

Thomas Wunderli

Abstract We extend the theory of functionals defined on BV space by including certain Carathéodory functions $\varphi(x, \mathbf{p})$ for functionals of the form $\int_\Omega \varphi(x, Du)$, $u \in BV(\Omega)$, so that φ is only measurable in x without the usual continuity assumption in x, and prove lower semicontinuity in L^1 of $\int_\Omega \varphi(x, Du)$ as well as compactness with an extra with an L^1 condition on φ. We also consider the case of the dual H^1 penalty model with integral constraint introduced in Osher-Solé-Vese [38] for image restoration, with the more general energy term $\int_\Omega \varphi(x, Du)$, analyze the time flow of the dual H^1 model in BV, and derive an integral property for the flow in the case of one space dimension.

Keywords Bounded variation · Image restoration · Gradient flows · Dual of h^1 · Anisotropic diffusion

2000 Mathematics Subject Classification. Primary 35A05 · 35D05 · Secondary 49xx

1 Introduction

In this paper we present some results of gradient time flows in $L^2(\Omega)$ corresponding to minimization problems of functionals of the form

$$\mathcal{F}(u) := \int_\Omega \varphi(x, Du) + \lambda \int_\Omega |\nabla(\Delta^{-1})(I-u)|^2 \, dx$$

with dual H^1 penalty term $\lambda \int_\Omega |\nabla(\Delta^{-1})(I-u)|^2 \, dx$ defined for $u \in BV(\Omega) \cap L^2(\Omega)$, $\Omega \subset \mathbb{R}^n$ open and bounded and constant $\lambda > 0$. Here we assume the

T. Wunderli (✉)
Department of Mathematics and Statistics,
The American University of Sharjah, P.O. Box 26666, Sharjah, UAE
e-mail: twunderli@aus.edu

T. Abualrub et al. (eds.), *Mathematics Across Contemporary Sciences*,
Springer Proceedings in Mathematics & Statistics 190,
DOI 10.1007/978-3-319-46310-0_15

Carathéodory function $\varphi(x, \mathbf{p})$, $\varphi : \Omega \times \mathbb{R}^n \to [0, \infty)$, is for a.e. x both convex and has a linear growth assumption in $\mathbf{p}$, and also has an additional integrability assumption to insure compactness. We make no assumption of continuity in x. As described later in this section, the minimization problem was originally proposed for image restoration applications in Osher et al. [38] for the case of pure total variation term $\varphi(Du) = |Du|$ with dual H^1 penalty.

Existence, uniqueness, and qualitative properties for solutions for flows in L^1 and L^2 with pure total variation term and different boundary conditions were obtained in [9–13, 18] with no penalty term for the L^1 case, and simple L^2 penalty for the L^2 case. For the purpose of the study of entropy solutions, they also consider flows in L^1 with quasilinear term $\phi(x, Du)$ for $u \in BV$ where ϕ has a strong continuity assumption in x. For our case, in addition to the dual H^1 penalty, $\varphi(x, Du)$ includes certain Carathéodory functions that are only measurable in x with no continuity assumption in x. The flow considered in this paper is

$$\frac{\partial u}{\partial t} = div\left(\nabla_{\mathbf{p}}\varphi(x, Du)\right) - 2\lambda\Delta^{-1}(I - u) \text{ for } t > 0, \text{ on } \Omega$$

with constraint $\int_\Omega u\,dx = \int_\Omega I\,dx$, initial condition $u(0, x) = I(0)$, Neumann boundary condition $\frac{\partial u}{\partial \mathbf{n}} = 0$ on $\partial\Omega$, for open bounded $\Omega \subset \mathbb{R}^1$ or $\mathbb{R}^2$ with Lipschitz boundary, and $\varphi(x, Du)$ as mentioned above.

One of the objectives of image processing is to restore corrupted images while retaining important features of the image, such as edges. One of the first models for this purpose using total variation was the Rudin-Osher-Fatemi (ROF) model [40, 41]. The ROF model consists of finding a minimizer $u_m \in L^2(\Omega)$ of the functional

$$\mathcal{R}(u) := \int_\Omega |\nabla u| + \frac{\lambda}{2}\int_\Omega (u - I)^2\,dx \tag{1.1}$$

where $I : \Omega \to \mathbb{R}$, $\Omega \subset \mathbb{R}^n$ bounded and open, represents the noisy or corrupted image and u_m represents the restored or cleaned image. For these types of minimization models, the images are represented by functions $u : \Omega \to \mathbb{R}$, where $\Omega \subset \mathbb{R}^2$ is typically a rectangle, and $u(x)$ the image intensity at x. The first term on the right in the above functional is the total variation of u:

$$\begin{aligned} TV(u) &:= \int_\Omega |\nabla u| \\ &:= \sup\left\{\int_\Omega u\nabla\cdot\varphi\,dx \;:\; \varphi \in C_c^1(\Omega; \mathbb{R}^n), |\varphi(x)| \le 1 \text{ for all } x \in \Omega\right\}. \end{aligned}$$

The space of all such $u \in L^1(\Omega)$ with $TV(u) < \infty$ is known as the space of functions of bounded variation, or $BV(\Omega)$, with the norm $\|u\|_{BV} =: \|u\|_{L^1(\Omega)} + \int_\Omega |\nabla u|$. Any minimizer of $\mathcal{R}$ will be in $BV(\Omega)$. It is common to use the Lebesgue decomposition to write any $u \in BV$ as

$$\int_\Omega |Du| = \int_\Omega |\nabla u|\, dx + \int_\Omega |D^s u|$$

where we decompose the total variation measure Du,with $|Du| =: \int_\Omega |\nabla u|$ into the absolutely continuous part with respect to Lebesgue measure $\nabla u\, dx$ and the singular part $D^s u$ as

$$Du = \nabla u\, dx + D^s u.$$

In [15] the above integral for $u \in BV$ is extended to $\int_\Omega \varphi(x, Du)$ for functions $\varphi(x, \mathbf{p})$, $x \in \Omega$, $\mathbf{p} \in \mathbb{R}^n$, continuous on $\Omega \times \mathbb{R}^n$, and convex and of linear growth in $\mathbf{p}$ (see Theorem 2 in the next section). We also refer the reader to [27] for results concerning certain functionals of the form $\int_\Omega \sqrt{1 + |Du|} dx + \int_\Omega G(x, u)\, dx$.

The use of BV space with the TV term is that minimizers of $\mathcal{R}$ may still be discontinuous with jumps corresponding to edges, unlike images restricted to the Sobolev space $W^{1,1}$. The second term of $\mathcal{R}$ is the penalty which ensures that the restored image u does not deviate too far from the input image I. One way to solve this is to solve the gradient flow of the Euler-Lagrange equation

$$0 = div\left(\frac{\nabla u}{|\nabla u|}\right) - \lambda(u - I)$$

and let $t \to \infty$ for the solution $u(x, t)$. See also [45] for the time flow for applications to plasticity. The gradient flow is then

$$\frac{\partial u}{\partial t} = div\left(\frac{\nabla u}{|\nabla u|}\right) - \lambda(u - I) \text{ on } \Omega \times [0, \infty) \text{ with } \frac{\partial u}{\partial \mathbf{n}} = 0 \text{ on } \partial\Omega$$
$$u(x, 0) = I(x) \text{ on } \Omega.$$

We should also mention the use of primal dual methods, instead of the gradient time flow, for minimizing functionals such as (1.1). These are especially used for models with pure TV term $\int_\Omega |Du|$ due its non differentiability. See, for example, [23] and [31].

In general, the above model works very well for image denoising while retaining edges. Modifications of the ROF model have also been introduced in other works to provide better restoration of noisy images due to such unwanted effects such as the stair casing effect, which may occur in solving (1.1) numerically. See [1–6, 16, 21, 22, 28, 29, 44] for further discussion and models.

Certain details, such a oscillatory textures are not well preserved with the above L^2 norm penalty $\frac{\lambda}{2}\int_\Omega (u - I)^2\, dx$. In [35], Meyer introduced a new penalty designed to overcome this, by replacing the L^2 penalty with a weaker norm that can retain oscillatory textures. In [35], the new model problem is to find a minimizer of

$$\mathcal{M}(u) := \int_\Omega |\nabla u| + \lambda \|I - u\|_* .$$

The new penalty norm $\|f\|_*$ is defined there for all $f \in G$ by

$$\|f\|_* = \inf \left\{ \sqrt{g_1^2 + ... + g_n^2} : g = (g_1, ..., g_n),\ g_i \in L^\infty(\Omega) \text{ each } i, \text{ and } f = \operatorname{div} g \right\},$$

where G is the Banach space of all generalized functions f that can be written as $f = \operatorname{div} g$ on Ω for some $g = (g_1, ..., g_n)$, $g_i \in L^\infty(\Omega)$ each i, open $\Omega \subset \mathbb{R}^n$.

To simplify the Euler-Lagrange equation for the $n = 2$ case, the authors in [42] replaced the minimization of $\mathcal{M}$ with finding

$$\inf_{u,g_1,g_2} \left\{ G_p(u, g_1, g_2) = \int_\Omega |\nabla u| + \lambda \int_\Omega |I - (u + \partial_x g_1 + \partial_y g_2)|^2 \, dxdy \right.$$
$$\left. + \mu \left[\int_\Omega \left(\sqrt{g_1^2 + g_2^2} \right)^p dxdy \right]^{1/p} \right\}$$

where λ, μ are parameters and $p \to \infty$. Due to the three variable functions u, g_1, g_2 this yields three coupled equations as a result of the Euler-Lagrange equations.

This approach is further simplified in [38] by dropping the last term in the above functional, by writing $I - u = \operatorname{div} g$ for $g \in L^2(\Omega)^2$, and by formally using the Hodge decomposition of g:

$$g = \nabla P + q$$

where q is a divergence free vector field, thus giving $u - I = -\operatorname{div} g = -\Delta P$. The inverse Laplace operator Δ^{-1} is then defined by $P =: -\Delta^{-1}(u - I)$. In fact we have (see for example, [26])

Theorem 1 *Let $\Omega \subset \mathbb{R}^n$ be a bounded open region with Lipschitz boundary $\partial\Omega$ and $V_0 = \left\{ u \in H^1(\Omega) : \int_\Omega u \, dx = 0 \right\}$. If $v \in L^2(\Omega)$ with $\int_\Omega v \, dx = 0$, then the problem*

$$-\Delta P = v, \quad \frac{\partial P}{\partial n}|_{\partial\Omega} = 0,$$

has a unique solution P in V_0.

Consequently, the OSV model proposed in [38] is to instead find a minimizer of

$$\mathcal{E}(u) := \int_\Omega |\nabla u| + \lambda \int_\Omega |\nabla(\Delta^{-1})(I - u)|^2 \, dx = \int_\Omega |\nabla u| + \lambda \|I - u\|^2_{H^{-1}(\Omega)} \quad (1.2)$$

over the space $L^2(\Omega)$ with the constraint $\int_\Omega u \, dx = \int_\Omega I \, dx$. For the last term on the right side it is shown in [38] that for functions $v \in L^2(\Omega)$ with $\int_\Omega v \, dx = 0$, $\|v\|^2_{H^{-1}(\Omega)} = \int_\Omega |\nabla(\Delta^{-1})v|^2 \, dx$. The Euler-Lagrange equation for this is formally

$$0 = div\left(\frac{\nabla u}{|\nabla u|}\right) - 2\lambda\Delta^{-1}(I-u) \text{ on } \Omega \tag{1.3}$$
$$\frac{\partial u}{\partial \mathbf{n}}|_{\partial\Omega} = 0$$

with constraint $\int_\Omega u\,dx = \int_\Omega I\,dx$. This is solved there numerically on a rectangle $\Omega \subset \mathbb{R}^2$ by applying $-\Delta$ to both sides of (1.3) and solving the following time flow for $u(x, y, t)$

$$\frac{\partial u}{\partial t} = -\Delta \operatorname{div}\left(\frac{\nabla u}{|\nabla u|}\right) + 2\lambda(I-u)$$
$$0 = \frac{\partial u}{\partial \mathbf{n}}|_{\partial\Omega} = \frac{\partial \operatorname{div}\left(\frac{\nabla u}{|\nabla u|}\right)}{\partial \mathbf{n}}|_{\partial\Omega},\ u(x, y, 0) = I(x, y) \text{ on } \Omega,$$

$\int_\Omega u\,dx = \int_\Omega I\,dx$, and letting $t \to \infty$ to drive to the steady state solution of (1.3). Clearly, the first term on the right of the equation is not defined for all functions u in BV or even $W^{1,1}$. We thus need to define a weak solution to the time flow to (1.3).

We will expand the functional $\mathcal{E}$ to include a class of Carathéodory functions for the energy term $\varphi(x, \mathbf{p})$ that are convex and of linear growth in $\mathbf{p}$. By definition, a Carathéodory function, $\varphi : \Omega \times \mathbb{R}^n \to \mathbb{R}$, satisfies the following conditions:

(1) for each $\mathbf{p} \in \mathbb{R}^n$, $\varphi(\cdot, \mathbf{p}) : \Omega \to \mathbb{R}$ is a measurable function defined on Ω and

(2) for a.e. $x \in \Omega$, $\varphi(x, \cdot) : \mathbb{R}^n \to \mathbb{R}$ is continuous in the $\mathbf{p}$ variable.

The functional is now

$$\mathcal{F}(u) := \int_\Omega \varphi(x, Du) + \lambda \int_\Omega |\nabla(\Delta^{-1})(I-u)|^2\,dx$$
$$\text{such that } \int_\Omega u\,dx = \int_\Omega I\,dx.$$

For example we have the variable exponent case,

$$\varphi(x, \mathbf{p}) = \begin{cases} \frac{1}{q(x)}|\mathbf{p}|^{q(x)} & \text{if } |\mathbf{p}| \le 1 \\ |\mathbf{p}| - \frac{q(x)-1}{q(x)} & \text{if } |\mathbf{p}| > 1 \end{cases} \tag{1.4}$$

where $q(x) \in L^\infty(\Omega)$, $1 < \alpha \le q(x) \le 2$ a.e. See [36] and [24] for an application of a functional using the anisotropic diffusion term (1.4) with simple L^2 penalty term $\int_\Omega (u-I)^2\,dx$. We also refer the reader to [39] PDE problems with variable exponent. The time flow of the Euler-Lagrange equations for $\mathcal{F}$ becomes

$$\begin{aligned} \frac{\partial u}{\partial t} &= div\left(\nabla_{\mathbf{p}}\varphi(x,\nabla u)\right) - 2\lambda\Delta^{-1}(I-u) \qquad (1.5)\\ \frac{\partial u}{\partial \mathbf{n}} &= 0 \text{ on } \partial\Omega\\ u(x,0) &= I(x)\\ \int_\Omega u\,dx &= \int_\Omega I\,dx \text{ for all } t. \end{aligned}$$

The rest of the paper is organized as follows. We extend the definition of functionals $\int_\Omega \varphi(x, Du)$ defined for $u \in BV\,(\Omega)$ by including certain Carathéodory functions $\varphi(x, \mathbf{p})$ where we directly use the convex dual function φ^* of φ rather than the theory of convex functionals of measures as in [8, 14, 15]. We only assume measurability in x for φ whereas previous work uses a continuity condition in x to prove lower semicontinuity of $\int_\Omega \varphi(x, Du)$ in L^1. In addition we prove compactness in L^1 with an extra L^1 integrability condition on φ. This then allows for a greater class of functionals to be considered for minimization problems that use both the $\int_\Omega \varphi(x, Du)$ term for smoothing and the dual H^1 penalty for retaining oscillatory features of images. For example we may use a more robust selective smoothing term $\int_\Omega \varphi(x, Du)$ in place of the simple total variation term $\int_\Omega |Du|$ that is used in the OSV model. We thus consider the OSV dual H^1 penalty model from [38] with general energy term $\int_\Omega \varphi(x, Du)$ and the corresponding gradient time flow (1.5). We then use the semigroup method to prove existence, L^2 stability, and asymptotic convergence for the weak solution to the time flow (1.5). It should be noted that the semigroup method is used in [9–13], where, as previously mentioned, they proved existence of a strong solution of the total variation flow

$$\begin{aligned} \frac{\partial u}{\partial t} &= div\left(\frac{\nabla u}{|\nabla u|}\right)\\ u(x,0) &= I(x) \end{aligned}$$

with both Neumann or Dirichlet boundary conditions in L^1 and L^2. They also considered flows with a quasilinear term $\mathrm{div}(\nabla_{\mathbf{p}}\phi(x, \nabla u))$ with a modulus of continuity assumption for ϕ in the x variable. Since the flow for our case is in the Hilbert space L^2, we apply the theory of semigroups based on classical maximal monotone theory of Brezis [17]. Finally, we derive an integral property for solutions to the gradient flow for the case of space dimension $n = 1$ with pure TV term $\int_\Omega |Du|$. As we note in the Conclusion, it is hoped to extend this integral property, or possibly derive other properties, to our general case with the $\int_\Omega \varphi(x, Du)$ term.

2 The Stationary Problem and Important Results for BV Functions

We will first state some important theorems concerning functions in BV space. The following theorem is from [43]. Also see [7].

Theorem 2 *Let $\Omega \subset \mathbb{R}^n$ be bounded and open, $\varphi(x, \mathbf{p})$ be C^1 on $\Omega \times \mathbb{R}^n$, convex in $\mathbf{p}$, with linear growth for $|\mathbf{p}| \geq \beta > 0$, that is $c_1|\mathbf{p}| \leq \varphi(x, \mathbf{p}) \leq c_2(|\mathbf{p}| + 1)$ for $|\mathbf{p}| \geq \beta$ with constants $c_1, c_2, \beta > 0$, and where $\lim_{t\to\infty} \varphi(x, t\frac{\mathbf{p}}{|\mathbf{p}|})/t = \varphi^\infty(x)$. Then $\int_\Omega \varphi(x, Du)$ is lower semicontinuous in $L^1(\Omega)$.*

From [15] we also have a formula for $\int_\Omega \varphi(x, Du)$ for $u \in BV\ (\Omega)$ where φ is the C^1 function as stated in Theorem 2, in fact

$$\int_\Omega \varphi(x, Du) = \int_\Omega \varphi(x, \nabla u)\, dx + \int_\Omega \varphi^\infty(x)|D^s u|. \tag{2.1}$$

The approximation of BV functions by smooth functions for the anisotropic functional with (1.4) is given by:

Theorem 3 *If $\Omega \subset \mathbb{R}^n$ is bounded, open, $u \in BV(\Omega) \cap L^2(\Omega)$, and φ given by (1.4), then (1) there exists a sequence $\{u_j\} \subset C^\infty(\Omega) \cap H^1(\Omega)$ such that*

$$u_j \to u \text{ in } L^2(\Omega) \text{ and}$$
$$\lim_{j\to\infty} \int_\Omega \varphi(x, Du_j) = \int_\Omega \varphi(x, Du);$$

(2) if $\int_\Omega u\, dx = c$, we may take the sequence above to also satisfy $\int_\Omega u_j\, dx = c$; (3) if $u \in L^\infty(\Omega)$, and φ is independent of x, then we may also take the sequence to satisfy $\|u_j\|_{L^\infty} \leq C(\Omega)\, \|u\|_{L^\infty}$, and if Ω has Lipschitz boundary $\partial\Omega$, we may also take the sequence to satisfy $u_j \in C^\infty(\overline{\Omega})$.

Proof With simple modifications, the first part is proved as in [24] (in their case for u with trace value $Tu_{|\partial\Omega}$) using $\int_\Omega \varphi(x, Du) = \sup_{\phi\in\mathcal{V}} \{-\int_\Omega u div\phi + \varphi^*(x, \phi)\, dx\}$. In fact it is only assumed that $q(x) \in L^\infty(\Omega)$, $1 < \alpha \leq q(x) \leq 2$ a.e. For the second part we note that $u_j \to u$ in $L^2(\Omega)$ implies that $\int_\Omega u_j\, dx \to \int_\Omega u\, dx = c$. We then let $\widetilde{u}_j = u_j - \frac{1}{|\Omega|}\int_\Omega (u_j - c)dx$ giving

$$\widetilde{u}_j \to u \text{ in } L^2(\Omega) \text{ and}$$
$$\lim_{n\to\infty} \int_\Omega \varphi(x, D\widetilde{u}_j) = \int_\Omega \varphi(x, Du)$$
$$\int_\Omega \widetilde{u}_j \, dx = 0 \text{ for all } j.$$

For (3), note the remark in [25]. □

Remark 1 For the case of pure total variation $\varphi(\mathbf{p}) = |\mathbf{p}|$, from [32] Theorems 2 and 3 hold with the proof of (2) the same as above.

For similar approximation results with a different proof for functions φ defined on $\Omega \times \mathbb{R}^n$, with certain continuity conditions on both x and $\mathbf{p}$, see for example [12].

We now extend the definition of $\int_\Omega \varphi(x, Du)$ for a Carath éodory function which is continuous in $\mathbf{p}$ and no continuity assumption in x, by using the Legendre transform. For a convex function g on $\mathbb{R}^n$, the convex dual or Legendre transform $g^*$of g, is defined as $g^*(\mathbf{q}) = \sup_{\mathbf{p}\in\mathbb{R}^n} \{\mathbf{p}\cdot\mathbf{q} - g(\mathbf{p})\}$. If g is continuous then by the Fenchel-Moreau theorem [19] we in fact have $g^{**}(\mathbf{p}) = g(\mathbf{p}) = \sup_{\mathbf{q}\in\mathbb{R}^n} \{\mathbf{p}\cdot\mathbf{q} - g^*(\mathbf{q})\}$.

Proposition 1 *Let $\Omega \subset \mathbb{R}^n$ be open, $\varphi(x, \mathbf{p})$ a Carathéodory function on $\Omega \times \mathbb{R}^n$, continuous and convex in $\mathbf{p}$, of linear growth in $\mathbf{p}$ with $c_1|\mathbf{p}| - c_2 \le \varphi(x, \mathbf{p}) \le c_1(|\mathbf{p}| + 1)$ for $|\mathbf{p}| \ge \beta$, for constants $c_1 > 0$, $\beta, c_2 \ge 0$. Then (1) for a.e. x, $\varphi^*(x, \mathbf{q}) = \sup_{\{\mathbf{p}\in\mathbb{R}^n, |\mathbf{p}|\le\beta\}}\{\mathbf{q}\cdot\mathbf{p} - \varphi(x, \mathbf{p})\} = \max_{\{\mathbf{p}\in\mathbb{R}^n, |\mathbf{p}|\le\beta\}}\{\mathbf{q}\cdot\mathbf{p} - \varphi(x, \mathbf{p})\}$ and (2) $\varphi^*(x, \mathbf{q})$ is a Carathéodory function on $\Omega \times \{|\mathbf{q}| \le c_1\}$. Furthermore $\varphi^*(x, \mathbf{q}) = \infty$ for a.e. x, $|\mathbf{q}| > c_1$.*

Proof By the linear growth condition $\varphi(x, \mathbf{p}) \le c_1(|\mathbf{p}| + 1)$, we have $\varphi^*(x, \mathbf{q}) < \infty$ if and only if $|\mathbf{q}| \le c_1$ and this occurs for $|\mathbf{p}| \le \beta$ from the assumption $c_1|\mathbf{p}| - c_2 \le \varphi(x, \mathbf{p})$. The fact that the supremum is a maximum follows by continuity. This proves (1). To prove (2) we fix a.e. x and first assume that $\varphi(x, \mathbf{p})$ is strictly convex for $|\mathbf{p}| \le \beta$. The case where $\beta = 0$ gives $\varphi^*(x, \mathbf{q}) = \max_{\{\mathbf{p}\in\mathbb{R}^n, |\mathbf{p}|=0\}}\{\mathbf{q}\cdot\mathbf{p} - \varphi(x, \mathbf{p})\} = -\varphi(x, \mathbf{0})$ if $|\mathbf{q}| \le c_1$. Now assume $\beta > 0$. Then by strict convexity there is a unique $\mathbf{p}^*(\mathbf{q})$ with $|\mathbf{p}^*(\mathbf{q})| \le \beta$ so that $\varphi^*(x, \mathbf{q}) = \mathbf{q}\cdot\mathbf{p}^*(\mathbf{q}) - \varphi(x, \mathbf{p}^*(\mathbf{q}))$. To show that $\mathbf{p}^*$ is continuous we let $\mathbf{q}_n \to \mathbf{q}$. Thus there is a subsequence $\mathbf{q}_{n_k}$ such that $\mathbf{p}^*(\mathbf{q}_{n_k}) \to \mathbf{p}'$ for some $|\mathbf{p}'| \le \beta$. Hence for each $\mathbf{q}_{n_k}$, $\varphi^*(x, \mathbf{q}_{n_k}) = \mathbf{q}_{n_k}\cdot\mathbf{p}^*(\mathbf{q}_{n_k}) - \varphi(x, \mathbf{p}^*(\mathbf{q}_{n_k})) \ge \mathbf{q}_{n_k}\cdot\mathbf{p} - \varphi(x, \mathbf{p})$ for all $|\mathbf{p}| \le \beta$. Thus for each $|\mathbf{p}| \le \beta$ $\mathbf{q}\cdot\mathbf{p} - \varphi(x, \mathbf{p}) \le \lim_{k\to\infty} \varphi^*(x, \mathbf{q}_{n_k}) = \lim_{k\to\infty}(\mathbf{q}_{n_k}\cdot\mathbf{p}^*(\mathbf{q}_{n_k}) - \varphi(x, \mathbf{p}^*(\mathbf{q}_{n_k})) = \mathbf{q}\cdot\mathbf{p}' - \varphi(x, \mathbf{p}')$. Therefore $\mathbf{p}' = \mathbf{p}^*(\mathbf{q})$. To show that the full sequence $\mathbf{p}^*(\mathbf{q}_n)$ converges to $\mathbf{p}^*(\mathbf{q})$ we assume that there is another subsequence $\mathbf{q}_{n_i}$ and $\varepsilon > 0$ such that $\mathbf{q}_{n_i} \to \mathbf{q}$ but $\left|\mathbf{p}^*(\mathbf{q}_{n_i}) - \mathbf{p}^*(\mathbf{q})\right| \ge \varepsilon$ for all n_i. We extract a further subsequence $\mathbf{q}_{n_{i_j}}$ with $\mathbf{q}_{n_{i_j}} \to \mathbf{q}$ and $\mathbf{p}^*(\mathbf{q}_{n_{i_j}}) \to \mathbf{p}''$. Repeating the above argument we have $\mathbf{p}'' = \mathbf{p}^*(\mathbf{q})$ but $\left|\mathbf{p}'' - \mathbf{p}^*(\mathbf{q})\right| \ge \varepsilon$, a contradiction. Since $\mathbf{p}^*(\mathbf{q})$ is continuous, so is $\varphi^*(x, \mathbf{q})$. Without the strict convex assumption on $\varphi(x, \mathbf{q})$ we consider $\varphi_\varepsilon(x, \mathbf{p}) := \varphi(x, \mathbf{p}) + \varepsilon|\mathbf{p}|^2$ for $|\mathbf{p}| \le \beta$. As $\varepsilon \ge \varphi^*(x, \mathbf{p}) - \varphi^*_\varepsilon(x, \mathbf{p})$ and

$\varphi^*(x, \mathbf{p}) \geq \varphi^*_\varepsilon(x, \mathbf{p})$ we have $\varepsilon \geq |\varphi^*(x, \mathbf{p}) - \varphi^*_\varepsilon(x, \mathbf{p})|$ and thus $\varphi^*_\varepsilon \to \varphi^*$ uniformly on $|\mathbf{p}| \leq \beta$ as $\varepsilon \to 0$. Since $\varphi_\varepsilon(x, \mathbf{p})$ is strictly convex for $|\mathbf{p}| \leq \beta$, $\varphi^*_\varepsilon(x, \mathbf{p})$ is continuous in $\mathbf{p}$, hence it follows that $\varphi^*(x, \mathbf{p})$ is continuous for $|\mathbf{p}| \leq \beta$. Finally, φ^* being for fixed $\mathbf{q}$ the pointwise maximum of measurable functions in x, is measurable in x. Item (2) is proved. □

This proposition then allows us to define the following:

Definition 1 For open $\Omega \subset \mathbb{R}^n$ and $\varphi(x, \mathbf{p})$ a Carathéodory function on $\Omega \times \mathbb{R}^n$, continuous and convex in $\mathbf{p}$, of linear growth in $\mathbf{p}$ with $c_1|\mathbf{p}| - c_2 \leq \varphi(x, \mathbf{p}) \leq c_1(|\mathbf{p}| + 1)$ for $|\mathbf{p}| \geq \beta$, for constants $c_1 > 0$, $\beta, c_2 \geq 0$. Define

$$\int_\Omega \varphi(x, Du) = \sup_{\phi \in \mathcal{V}} \left\{ - \int_\Omega u div\phi + \varphi^*(x, \phi(x))\, dx \right\}$$

where $\varphi^*(x, \mathbf{q}) = \sup_{\{\mathbf{p} \in \mathbb{R}^n, |\mathbf{p}| \leq \beta\}} \{\mathbf{q} \cdot \mathbf{p} - \varphi(x, \mathbf{p})\}$ for each $\mathbf{q} \in \mathbb{R}^n$ with $|\mathbf{q}| \leq c_1$ and

$$\mathcal{V} = \left\{ \phi \in C^1_c(\Omega, \mathbb{R}^n) : |\phi(x)| \leq c_1 \text{ for all } x \in \Omega \right\}.$$

Note that the supremum is only taken for $\phi \in \mathcal{V}$ since from the proposition $\varphi^*(x, \mathbf{q}) = \infty$ if $|\mathbf{q}| > c_1$.

We remark that this is the definition used in [24] for the specific case of the anisotropic functional $\int_\Omega \varphi(x, Du)$ where φ, given by (1.4), satisfies the conditions of Definition 1, and φ^* is directly calculated. Also for the total variation case $\varphi(\mathbf{p}) = |\mathbf{p}|$ we have $c_1 = 1$ and φ^* is the usual

$$\varphi^*(\mathbf{q}) = \begin{cases} 0 & \text{if } |\mathbf{q}| \leq 1 \\ \infty & \text{otherwise} \end{cases}$$

From Definition 1, lower semicontinuity in $L^1(\Omega)$ follows immediately as in [32].

Theorem 4 *If Ω and φ satisfy the conditions of Definition 1, $\int_\Omega \varphi(x, Du)$ is lower semicontinuous in $L^1(\Omega)$.*

Proof Let $u_n \to u$ in $L^1(\Omega)$. Then for fixed $\phi \in \mathcal{V}$ we have $-\int_\Omega u div\phi + \varphi^*(x, \phi)\, dx = \lim_{n\to\infty} \left(- \int_\Omega u_n div\phi + \varphi^*(x, \phi)\, dx \right) \leq \liminf_{n\to\infty} \int_\Omega \varphi(x, Du_n)$. Taking the supremum on the left gives $\int_\Omega \varphi(x, Du) \leq \liminf_{n\to\infty} \int_\Omega \varphi(x, Du_n)$. □

With an added L^1 condition on φ we have

Theorem 5 *If $\Omega \subset \mathbb{R}^n$ is open and bounded, φ satisfies the conditions of Definition 1 and in addition $\int_\Omega \sup_{|\mathbf{p}| \leq \beta} |\varphi(x, \mathbf{p})|\, dx \leq c_3$ for some $c_3 > 0$, then $\int_\Omega \varphi(x, Du) < \infty$ if and only if $u \in BV(\Omega)$. In fact we have $c_1 \int_\Omega |Du| \leq \int_\Omega \varphi(x, Du) + C(c_1, c_3, \beta, \Omega)$ and $\int_\Omega \varphi(x, Du) \leq c_1 \int_\Omega |Du| + C(c_1, c_3, \beta, \Omega)$ for some constant $C(c_1, c_3, \beta, \Omega) \geq 0$.*

Proof From the definition of φ^* we have $\varphi^*(x, \phi(x)) \leq |\phi(x)|\beta + \sup_{|\mathbf{p}|\leq\beta} |\varphi(x, \mathbf{p})|$ and thus

$$\begin{aligned} c_1 \int_\Omega |Du| &= \sup_{\phi\in\mathcal{V}} \left\{ -\int_\Omega u div\phi \, dx \right\} \\ &\leq \sup_{\phi\in\mathcal{V}} \left\{ -\int_\Omega u div\phi + \varphi^*(x, \phi) \, dx \right\} \\ &\quad + \sup_{\phi\in\mathcal{V}} \left| \int_\Omega \varphi^*(x, \phi) \, dx \right| \\ &\leq \int_\Omega \varphi(x, Du) + C(c_1, c_3, \beta, \Omega) \end{aligned}$$

where $C(c_1, c_3, \beta, \Omega) \geq 0$; and also

$$\begin{aligned} \int_\Omega \varphi(x, Du) &= \sup_{\phi\in\mathcal{V}} \left\{ -\int_\Omega u div\phi + \varphi^*(x, \phi) \, dx \right\} \\ &\leq c_1 \int_\Omega |Du| + C(c_1, c_3, \beta, \Omega). \end{aligned}$$

□

We then have the compactness theorem:

Theorem 6 *Let φ satisfy the conditions of Theorem 5. Let u_j be a sequence in $BV(\Omega)$ with $\int_\Omega \varphi(x, Du_j)$ bounded, where $\Omega \subset \mathbb{R}^n$ is bounded with Lipschitz boundary $\partial\Omega$. Then there is a subsequence of u_j, also denoted by u_j, and $u \in L^p(\Omega)$ such that $u_j \to u$ strongly in $L^p(\Omega)$ for all $1 \leq p < n/(n-1)$ and weakly in $L^{n/(n-1)}(\Omega)$.*

Proof From Theorem 5, u_j is a sequence bounded in $BV(\Omega)$. The theorem then follows from Giusti [32]. □

Remark 2 We assumed that $c_1|\mathbf{p}| - c_2 \leq \varphi(x, \mathbf{p}) \leq c_1(|\mathbf{p}| + 1)$ for $|\mathbf{p}| \geq \beta$ for ease of proof. However, we may replace this with the more general linear growth condition $k_1|\mathbf{p}| - c \leq \varphi(x, \mathbf{p}) \leq k_2(|\mathbf{p}| + 1)$ for $|\mathbf{p}| \geq \beta$ for $k_2 > k_1 > 0$, $\beta, c \geq 0$, with the same convex and Carathéodory condition on φ. In this case we still have $\varphi^*(x, \mathbf{q}) < \infty$ if and only if $|\mathbf{q}| \leq k_2$. If $\varphi^*(x, \mathbf{q})$ achieves is supremum on a bounded set $|\mathbf{p}| \leq K$ where K is independent of $\mathbf{q}$, then Proposition 1, Definition 1, and Theorems 4–7 hold with the respective L^1 integral condition on φ.

We return to the minimization problem from [37] using the OSV model. We extend this model to include any φ as stated in Theorem 5. This assumption will hold in the sequel unless stated otherwise. As stated in the introduction the minimization model is

$$\min_{u\in BV\cap V_I} \mathcal{F}(u) := \int_\Omega \varphi\,(x, Du) + \lambda \left\|u - I\right\|^2_{H^{-1}(\Omega)} = \tag{2.2}$$
$$\int_\Omega \varphi\,(x, Du) + \lambda \int_\Omega \left|\nabla(\Delta^{-1})(I-u)\right|^2 dx$$

where $V_I =: \left\{u \in L^2\,(\Omega) \mid \int_\Omega u\,dx = \int_\Omega I\,dx\right\}$, $\Omega \subset \mathbb{R}^n$.

Theorem 7 *For $n = 1$ or 2, the functional F is convex, lower semicontinuous and thus the stationary problem (2.2) has a unique solution.*

Proof This is proved in [38] for the original problem of minimizing (1.2). We just note that for φ we still have lower semicontinuity and compactness from Theorems 4 and 5. Existence and uniqueness then follows from standard theory. □

For the rest of the paper we assume $n = 1$ or 2.

3 Time Flow of the Weak Solution

We return to the gradient time flow corresponding to the stationary problem (2.2) using the Euler-Lagrange equation of (2.2), $div\,(\varphi_P(x, \nabla u)) - 2\lambda\Delta^{-1}(u - I) = 0$. Without loss of generality we will assume $\varphi(x, \mathbf{0}) = 0$.

Definition 2 The time flow of (2.2) is defined by

$$\frac{\partial u}{\partial t} = div\left(\nabla_{\mathbf{p}}\varphi(x, \nabla u)\right) - 2\lambda\Delta^{-1}(I - u) \tag{3.1}$$
$$\frac{\partial u}{\partial \mathbf{n}} = 0 \text{ on } \partial\Omega \tag{3.2}$$
$$u(x, 0) = I(x) \tag{3.3}$$
$$\int_\Omega u\,dx = \int_\Omega I\,dx \text{ for all } t. \tag{3.4}$$

where $\Omega \subset \mathbb{R}^n$ is an open bounded region with Lipschitz boundary $\partial\Omega$, and $I \in L^2(\Omega) \cap BV\,(\Omega)$.

Since u is assumed to be only in BV, this must be defined as a weak solution as will be given below. In the sequel, Ω satisfies the conditions stated in Definition 2. Following, for example, [25, 45] we motivate the definition of a weak solution to (3.1)–(3.4) by assuming sufficient smoothness of u and v satisfying the constraint

$$\int_\Omega u\,dx = \int_\Omega v\,dx = \int_\Omega I\,dx$$

for a.e. t, multiplying (3.1) by $v-u$, integrating by parts, using convexity of φ, namely $\varphi(x,\mathbf{p})-\varphi(x,\mathbf{q})\geq\nabla\varphi_P(x,\mathbf{q})\cdot(\mathbf{p}-\mathbf{q})$, noting that

$$\begin{aligned}\int_\Omega \Delta^{-1}(I-u)(v-u)\,dx &= -\int_\Omega \Delta^{-1}(I-u)\Delta\Delta^{-1}(v-u)\,dx\\ &= \int_\Omega \nabla\Delta^{-1}(I-u)\cdot\nabla\Delta^{-1}(v-I+I-u)\,dx,\end{aligned}$$

and finally expanding and using Young's inequality to get for a.e.t

$$\begin{aligned}&\int_\Omega u_t(v-u)\,dx+\int_\Omega\varphi(x,\nabla v)+\lambda\int_\Omega|\nabla(\Delta^{-1})(I-v)|^2\,dx \qquad (3.5)\\ &\geq\int_\Omega\varphi(x,\nabla u)+\lambda\int_\Omega|\nabla(\Delta^{-1})(I-u)|^2\,dx.\end{aligned}$$

By Theorem 3 we see that (3.5) holds for u, v in $BV(\Omega)$ satisfying the above constraint for a.e. t. We therefore define a *weak solution* $u\in L^2((0,T);L^2(\Omega)\cap V_I)\cap L^1((0,T);BV(\Omega))$, $u_t\in L^2((0,T),L^2(\Omega))$ of (3.1)–(3.4) to satisfy (3.5) for all

$$v\in L^2((0,T);L^2(\Omega)\cap V_I)\cap L^1((0,T);BV(\Omega))$$

where

$$V_I=\left\{v\in L^2(\Omega):\ \int_\Omega v\,dx=\int_\Omega I\,dx\right\}.$$

In what follows, let H_0 be the Hilbert space

$$H_0=\left\{v\in L^2(\Omega):\ \int_\Omega v\,dx=0\right\}.$$

Theorem 8 *Let φ satisfy the conditions of Theorem 5 and $I\in L^2(\Omega)\cap BV(\Omega)$. There exists a unique weak solution $u(t)$ to (3.1)–(3.4). That is, for a.e. $t>0$, $u(t)\in L^2(\Omega)$ with $u(t)-I\in BV(\Omega)\cap H_0$, $u_t\in L^\infty((0,\infty);H_0)$*

$$\begin{aligned}&\int_\Omega u_t(v-u)\,dx+\int_\Omega\varphi(x,Dv)+\lambda\int_\Omega|\nabla(\Delta^{-1})(I-v)|^2\,dx \qquad (3.6)\\ &\geq\int_\Omega\varphi(x,Du)+\lambda\int_\Omega|\nabla(\Delta^{-1})(I-u)|^2\,dx\end{aligned}$$

for each $v-I\in BV(\Omega)\cap H_0$. Hence for the case with constraint $\int_\Omega I\,dx=0$ we have for a.e.$t>0$

$$\frac{1}{2}\int_\Omega u^2\,dx + \int_0^t\int_\Omega \varphi(x, Du)\,ds + \lambda\int_0^t\int_\Omega |\nabla(\Delta^{-1})(I-u)|^2\,dx\,ds \tag{3.7}$$
$$\leq \lambda\int_0^t\int_\Omega |\nabla\Delta^{-1}I|^2\,dxds$$

and

$$\frac{1}{2}\int_\Omega (u-I)^2\,dx + \int_0^t\int_\Omega \varphi(x, Du)\,ds \tag{3.8}$$
$$+\lambda\int_0^t\int_\Omega |\nabla(\Delta^{-1})(I-u)|^2\,dxds \leq \int_0^t\int_\Omega \varphi(x, DI)\,ds;$$

and for the general case $\int_\Omega I\,dx = c$, *we have (3.8) for a.e.* $t > 0$

$$\frac{1}{2}\int_\Omega (u-\frac{c}{|\Omega|})^2\,dx + \int_0^t\int_\Omega \varphi(x, Du)\,ds + \lambda\int_0^t\int_\Omega |\nabla(\Delta^{-1})(I-u)|^2\,dxds \tag{3.9}$$
$$\leq \lambda\int_0^t\int_\Omega |\nabla(\Delta^{-1})(I-\frac{c}{|\Omega|})|^2\,dxds.$$

Also for initial conditions $I_1, I_2 \in L^2(\Omega)\cap BV(\Omega)$ *with corresponding solutions* u_1, u_2,

$$\|u_1-u_2\|_{L^2(\Omega)} \leq \|I_1-I_2\|_{L^2(\Omega)}$$

for a.e. $t > 0$. *Finally, The solution u to (3.1)–(3.4) converges weakly in* $L^2(\Omega)$ *and strongly in* $L^1(\Omega)$ *to the minimizer of* u_∞ *of 2.2 as* $t \to \infty$.

Proof We first assume $\int_\Omega I\,dx = 0$. The functional

$$F(u) =: \begin{cases} \int_\Omega \varphi(x, Du) + \lambda\int_\Omega |\nabla(\Delta^{-1})(I-u)|^2\,dx & \text{if } u \in BV(\Omega)\cap H_0 \\ \infty & \text{if } u \in H_0\backslash BV(\Omega) \end{cases}$$

on H_0 is proper, convex, and lower semicontinuous from Theorem 4. Consequently from the theory from maximal monotone operators and semigroups [17], the subdifferential $\partial F(u)$ is a maximal monotone operator with a unique, absolutely continuous solution $u(t) \in [0,\infty) \to H_0$, $u(0) = I$, $u_t \in L^\infty((0,\infty); H_0)$, to

$$-u_t \in \partial F(u(t)).$$

Thus by the definition of ∂F, the first inequality (3.6) holds. Also from [17]

$$\|u_1-u_2\|_{L^2(\Omega)} \leq \|I_1-I_2\|_{L^2(\Omega)}$$

for solutions u_1, u_2 with corresponding initial conditions $I_1, I_2 \in L^2(\Omega)\cap BV(\Omega)$. The inequalities (3.7) and (3.8) are obtained by letting $v = 0$ and $v = I$ respectively and integrating with respect to t. For the general constraint $\int_\Omega I\,dx = c$, we replace

u in (3.6) with $\widetilde{u} = u - \frac{c}{|\Omega|}$ so that $\int_\Omega \widetilde{u}\, dx = 0$. Letting $v = \frac{c}{|\Omega|}$ gives (3.9), noting $\varphi(x, \mathbf{0}) = 0$.

We now consider the asymptotic limit of the solution $u(t)$ as $t \to \infty$. Let u be the solution to (3.1)–(3.3). Since $-\frac{du}{dt} \in \partial F(u)$ the theorem from [20] proves that $u(t) \rightharpoonup u_\infty$ in $L^2(\Omega)$ weakly as $t \to \infty$. To prove strong convergence in $L^1(\Omega)$ we use Theorem A.33 in [12], which implies that, after adjusting by a constant if necessary, $\int_\Omega \varphi(x, Du(t)) + \lambda \int_\Omega |\nabla(\Delta^{-1})(I - u(t))|^2\, dx$ is a decreasing function of t with

$$\int_\Omega \varphi(x, Du(t)) + \lambda \int_\Omega |\nabla(\Delta^{-1})(I - u(t))|^2\, dx \leq \int_\Omega \varphi(x, DI).$$

From Poincaré's inequality for BV functions, $\int_\Omega |u - u_\Omega| dx \leq C(\Omega) \int_\Omega |Du|$ where $u_\Omega := \dfrac{\int_\Omega u dx}{|\Omega|} = 0$ for a.e. t. Thus by Theorem 5 $u(t)$ is bounded in $BV(\Omega)$ and by compactness (Theorem 6) and uniqueness of limits, there exists a subsequence $u(t_n) \to u_\infty$ in $L^1(\Omega)$ as $t_n \to \infty$. Hence $u(t) \to u_\infty$ in $L^1(\Omega)$ as $t \to \infty$. Again adjusting by a constant, we again have $u(t) \to u_\infty$ in $L^1(\Omega)$ for the general case of $\int_\Omega u dx = c$. □

For φ satisfying the conditions of Theorem 2, formula (2.1) holds. Now replacing v in (3.6) with $\eta(v - u) + u$ for $\eta > 0$ dividing by η and letting $\eta \to 0^+$ we obtain as in [33, 34] for any $v \in BV(\Omega) \cap H_0(\Omega)$ with $D^s v << |D^s u|$

$$\int_\Omega (v - u) u_t\, dx \leq -\int_\Omega \nabla_{\mathbf{p}} \varphi(x, \nabla u) \cdot (\nabla v - \nabla u)\, dx + \int_\Omega \varphi^\infty(x) \frac{D^s u}{|D^s u|} \cdot (D^s v - D^s u) - \int_\Omega \left[2\lambda \Delta^{-1}(I - u)\right](v - u)\, dx$$

where $\frac{D^s u}{|D^s u|}$ denotes the Radon-Nikodym derivative of $D^s u$ with respect to $|D^s u|$. Note that $\left|\frac{D^s u}{|D^s u|}\right| = 1$, $|D^s u|$-a.e. Repeating for $\eta < 0$ we have equality:

$$\int_\Omega (v - u) u_t\, dx = -\int_\Omega \nabla_{\mathbf{p}} \varphi(x, \nabla u) \cdot (\nabla v - \nabla u)\, dx + \int_\Omega \varphi^\infty(x) \frac{D^s u}{|D^s u|} \cdot (D^s v - D^s u) - \int_\Omega \left[2\lambda \Delta^{-1}(I - u)\right](v - u)\, dx \tag{3.10}$$

for a.e. $t \geq 0$, for all $v \in BV(\Omega) \cap H_0$ with $D^s v << |D^s u|$. Now letting $v = u + \phi$ for any $\phi \in C_0^\infty(\Omega) \cap H_0$

$$\frac{\partial u}{\partial t} = div\, (\varphi_P(x, \nabla u)) - 2\lambda \Delta^{-1}(I - u)\ \mathcal{D}'(\Omega) \cap H_0$$

as $D^s \phi = 0$. This gives

Corollary 1 *For the case* $\varphi \in C^1(\Omega \times \mathbb{R}^n)$ *satisfying the conditions of Theorem 2, the weak solution* $u(t)$ *to (3.1)–(3.4), satisfies for a.e.* $t \geq 0$,

$$\frac{\partial u}{\partial t} = div\,(\nabla \varphi_P(x, \nabla u)) - 2\lambda \Delta^{-1}(I-u)\,\mathcal{D}'(\Omega) \cap H_0$$

(in the distributional sense), and for fixed a.e. $t \geq 0$, *(3.10) holds for all* $v \in BV(\Omega) \cap H_0$ *with* $D^s v << |D^s u|$.

In the following theorem we note a property of the weak solution u to (3.1)–(3.3), inspired by a result for the stationary case in [38]. Additionally we extend this to an integral result for the case of $n = 1$.

Theorem 9 *Let* $w =: -2\lambda\Delta^{-1}(I-u)$. *If* u *is a weak solution to (3.1)-(3.3) for* $n = 1$ *and* $\varphi(\mathbf{p}) = \mathbf{p}$ *then there exists a* $g \in L^\infty(\Omega)$ *with* $\|g\|_\infty \leq 1$ *such that* $g' = w - u_t =: -2\lambda\Delta^{-1}(I-u) - u_t$.

Proof By assumption we have for a.e. $t \in [0, T]$

$$-u_t \in \partial J(u) + 2\lambda\Delta^{-1}(I-u)$$

where

$$J(u) =: \int_\Omega |\nabla u|.$$

Thus

$$-u_t - 2\lambda\Delta^{-1}(I-u) \in \partial J(u)$$

and hence by duality (see [30])

$$u \in \partial J^*(-2\lambda\Delta^{-1}(I-u) - u_t)$$

for a.e. t. where

$$J^*(u) = \sup_{u \in L^2(\Omega)} \int_\Omega (uv - J(u))dx = \begin{cases} 0 \text{ if } u \in K \\ +\infty \text{ otherwise} \end{cases}$$

and

$$K =: \left\{ div\,\mathbf{g} |\ \mathbf{g} \in (L^2(\Omega))^2 \text{ and } \|\mathbf{g}\|_\infty \leq 1 \right\}.$$

Therefore

$$\begin{aligned} 0 \in & -2\lambda u + 2\lambda \partial J^*(-2\lambda\Delta^{-1}(I-u) - u_t) \\ & = 2\lambda(I-u) - 2\lambda I + 2\lambda \partial J^*(-2\lambda\Delta^{-1}(I-u) - u_t). \end{aligned}$$

Hence for $w =: -2\lambda\Delta^{-1}(I-u)$

$$\begin{aligned} 0 &\in \partial \frac{\left\|\nabla(w+2\lambda\Delta^{-1}I)\right\|_{L^2}^2}{2} + 2\lambda\partial J^*(w-u_t) \\ &= \partial \frac{\left\|\nabla(w+2\lambda\Delta^{-1}I)\right\|_{L^2}^2}{2} + 2\lambda\partial \overline{J}(w) \end{aligned}$$

where $\overline{J}(v) =: J^*(v-u_t)$ and ∂ denotes the subdifferential. Thus w is in fact a minimizer of

$$G(\widehat{w}) =: \frac{\left\|\nabla(\widehat{w}+2\lambda\Delta^{-1}I)\right\|_{L^2}^2}{2} + 2\lambda\overline{J}(\widehat{w})$$

over all $\widehat{w} \in H^1(\Omega) \cap V_0$. For $n=1$, Ω an open interval, by choosing a $\widehat{w} \in H^1(\Omega) \cap V_0$ with $\|\widehat{w}-u_t\|_{L^2} \le |\Omega|^{-1/2}$ we can find g on Ω with $g' = \widehat{w} - u_t$ and $\|g\|_\infty \le 1$, namely $g(x) = \int_a^x (\widehat{w}-u_t)\,dx$, some $a \in \Omega$. Thus the functional $\overline{J}$ (and hence G) is proper, that is, $\overline{J}$ is finite for some $\widehat{w}$. Therefore $\widehat{w} \in K$ and the theorem is proved. □

Corollary 2 *If u is a weak solution to (3.1)–(3.3) for $n=1$, $\varphi(\mathbf{p}) = |\mathbf{p}|$, with Ω an open interval, then for each subinterval $[z, z'] \subset \Omega$,*

$$\operatorname{ess\,sup}_{t\ge 0} \left| \int_z^{z'} \lambda\Delta^{-1}(I-u) + \frac{1}{2}u_t\,dx \right| \le 1.$$

Proof On each subinterval $[z, z']$ of Ω we have for a.e. $t \ge 0$

$$\int_z^{z'} g'\,dx = \int_z^{z'} (-2\lambda\Delta^{-1}(I-u) - u_t)\,dx.$$

Hence as $\|g\|_\infty \le 1$,

$$\left| \int_z^{z'} (-2\lambda\Delta^{-1}(I-u) - u_t)\,dx \right| \le 2$$

for a.e. $t \ge 0$. □

4 Conclusion

We have defined $\int_\Omega \varphi(x, Du)$ for a class Carath éodory functions $\varphi(x, \mathbf{p})$ that are convex and of linear growth in $\mathbf{p}$, with the use of the convex dual φ^* of φ. With this definition, lower semicontinuity in L^1 immediately follows without any continuity assumption in x as was assumed in previous work. We then used these results to

prove the existence of the flow in $BV \cap L^2$ of the dual H^1 penalty image restoration model, with our general energy term $\int_\Omega \varphi(x, Du)$, rather than just $\int_\Omega |Du|$ as in [44].

For further study, we note that for functions $u \in W^{1,1}(\Omega)$, integration by parts and the Fenchel-Moreau theorem gives

$$\begin{aligned} -\int_\Omega u \operatorname{div} \phi + \varphi^*(x, \phi(x))\, dx &= \int_\Omega \nabla u \cdot \phi - \varphi^*(x, \phi(x))\, dx \\ &\leq \int_\Omega \sup_{\phi \in \mathcal{V}} \left\{ \nabla u \cdot \phi - \varphi^*(x, \phi(x)) \right\} dx \\ &= \int_\Omega \varphi^{**}(x, \nabla u)\, dx = \int_\Omega \varphi(x, \nabla u) dx. \end{aligned}$$

Thus $\int_\Omega \varphi(x, Du) \leq \int_\Omega \varphi(x, \nabla u) dx$. To show the reverse inequality, we require a sequence of functions ϕ_j in $C_c^1(\Omega)$ such that

$$\sup_j \int_\Omega \nabla u \cdot \phi_j - \varphi^*(x, \phi_j(x))\, dx \geq \int_\Omega \sup_{\phi \in \mathcal{V}} \left\{ \nabla u \cdot \phi - \varphi^*(x, \phi(x)) \right\} dx.$$

For $\varphi \in C^1(\Omega \times \mathbb{R}^n)$ we may use the implicit function theorem as was done in [43], whereas in this case we only have φ^* measurable in x. Using Proposition 1, it is hoped we can extend Theorem 3 to our class of φ as was done for the anisotropic model in [24], as well as extend formula 2.1 and hence Corollary 1, if φ is C^1 in $\mathbf{p}$. We may also consider extensions of Theorem 9 and Corollary 2 for this class of φ, noting the use of the dual J^*.

References

1. Allard, W.K.: Total variation regularization for image denoising I. SIAM J. Math. Anal. **39**(4), 1150–1190
2. Allard, W.K.: Total variation regularization for image denoising II. SIAM J. Imaging Sci. **1**(4), 400–417
3. Allard, W.K.: Total variation regularization for image denoising III. SIAM J. Imaging Sci. **2**(2), 532–568
4. Alliney, S.: Digital filters as absolute norm regularizers. IEEE Trans. Signal Process. **40**(6), 1548–1562 (1992)
5. Alliney, S.: Recursive median filters of increasing order: a variational approach. IEEE Trans. Signal Process. **44**(6), 1346–1354 (1996)
6. Alliney, S.: A property of the minimum vectors of a regularizing functional defined by means of the absolute norm. IEEE Trans. Signal Process. **45**(4), 913–917 (1997)
7. Ambrosio, L., Buttazzo, G., Fonseca, G.: Lower semicontinuity problems in Sobolev spaces with respect to a measure. J. Math. Pures Appl. **75**, 211–224 (1996)
8. Ambrosio, L., Mortola, S., Tortorelli, V.M.: Functionals with linear growth defined on vector valued BV functions. J. Math. Pures Appl. **70**, 269–323 (1991)
9. Andreu, F., Ballester, C., Caselles, V., Mazón, J.M.: The Dirichlet problem for total variation flow. J. Funct. Anal. **180**, 347–403 (2001)

10. Andreu, F., Ballester, C., Caselles, V., Mazón, J.M.: Minimizing total variation flow. Differ. Integral Eq. **14**(3), 321–360 (2001)
11. Andreu, F., Caselles, V., Díaz, J.I., Mazón, J.M.: Some qualitative properties for the total variation flow. J. Funct. Anal. **188**(2), 516–547 (2002)
12. Andreu-Vaillo, F., Caselles, V., Mazón, J.M.: Parabolic quasilinear equations minimizing linear growth functionals. Progress in Mathematics (Boston, Mass.), vol. 223. Basel, Birkhuser (ISNB 3-7643-6691-2/HBK). xiv, 340 p. (2004)
13. Andreu, F., Mazón, J.M., Moll, J.S., Caselles, V.: The minimizing total variation flow with measure initial conditions. Commun. Contemp. Math. **6**(3), 431–494 (2004)
14. Anzellotti, G.: The Euler equation for functionals with linear growth. Trans. Amer. Math. Soc. **290**, 483–500 (1985)
15. Anzellotti, G., Giaquinta, M.: Convex functionals and partial regularity. Arch. Rat. Mech. Anal. **102**, 243–272 (1988)
16. Blomgren, P., Chan, T., Mulet, P., Wong, C.K.: Total variation image restoration: Numerical methods and extensions. In: Proceedings of the 1997 IEEE International Conference on Image Processing, vol. 3. pp. 384–387
17. Brezis, H.: Opérateurs Maximaux Monotones. North-Holland, Amsterdam (1993)
18. Bellettini, G., Caselles, V., Novaga, M.: Total variation flow in $\mathbb{R}^n$,. J. Differ. Eq. **184**, 475–525 (2002)
19. Borwein, J.M., Lewis, A.S.: Convex Analysis and Nonlinear Optimization: Theory and Examples (2 edn.), pp. 76–79. Springer (2006)
20. Brück, R.E.: Asymptotic convergence of nonlinear contraction semigroups in Hilbert space. J. Funct. Anal. **18**, 15–26 (1975)
21. Chan, T., Esedoglu, S.: Aspects of total variation regularized L1 function approximation. SIAM J. Appl. Math. **65**(5), 1817–1837 (2005)
22. Chambolle, A., Lions, P.L.: Image recovery via total variation minimization and related problems. Numerische Mathematik **76**, 167–188 (1997)
23. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vis. **40**, 120–145 (2011)
24. Chen, Y., Levine, S., Rao, M.: Variable exponent, linear growth functionals in image restoration. SIAM J. Appl. Math **66**(4), 1383–1406 (2006)
25. Chen, Y., Wunderli, T.: Adaptive total variation for image restoration in BV space. J. Math. Anal. Appl. **272**, 117–137 (2002)
26. Dautray, R., Lions, J.-L.: Mathematical Analysis and Numerical Methods for Science and Technology, vol. 2. Springer-Verlag, Berlin, Functional and Variational Methods (1988)
27. Degiovanni, M., Marzocchi, M., Radulescu, V.: Multiple solutions of hemivariational inequalities with area-type term. Calc. Var. PDE **10**, 355–387 (2000)
28. Demengel, F., Temam, R.: Convex functions of a measure and applications. Indiana Univ. Math. J. **33**(5), 673–709 (1984)
29. Dobson, D., Santosa, F.: Recovery of blocky images from noisy and blurred data. SIAM J. Appl. Math. **56**, 1181–1198 (1996)
30. Ekeland, I., Temam, R.: Convex Analysis and Variational Problems. North Holland, Amsterdam (1976)
31. Esser, E., Zhang, X., Chan, T.F.: A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. SIAM J. Imaging Sci. **3**(4), 1015–1046 (2010)
32. Giusti, F.: Minimal Surfaces and Functions of Bounded Variation, Monographs Math, vol. 80. Birkhauser, Basel (1984)
33. Hardt, R., Kinderlehrer, D.: Elastic plastic deformation. Appl. Math. Optim. **10**, 203–246 (1983)
34. Hardt, R., Kinderlehrer, D.: Variational problems with linear growth. PDE's Cal. Var. vol. 2, pp. 633–659. Birkhauser (1989)
35. Meyer, Y.: Oscillating Patterns in Image Processing and Nonlinear Evolution Equations. Univ. Lecture Ser. 22, AMS, Providence, RI (2002)

36. Osher, S., Esedoglu, S.: Decomposition of images by the anisotropic Rudin-Osher-Fatemi model. Commun. Pure Appl. Math. **57**(12), 1609–1626 (2004)
37. Osher, S., Scherzer, O.: G-Norm properties of bounded variation regularization. Commun. Math. Sci. **2**(2), 237–254 (2004)
38. Osher, S., Solé, A., Vese, L.: Image decomposition and restoration using total variation minimization and the H^{-1} norm. Multiscale Model. Simul. **1**(3), 349–370
39. Radulescu, V., Repovs, D.: Partial Differential Equations with Variable Exponent: Vartiational Methods and Qualitative Analysis. CRC Press, Tayor and Francis, Boca Raton (2015)
40. Rudin, L.I., Osher, S.: Total variation based image restoration with free local constraints. Proc. ICIP IEEE Int. Conf. Image Process. Austin, TX, pp. 31–35 (1994)
41. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D. **60**, 259–268 (1992)
42. Vese, L., Osher, S.: Modeling textures with total variation minimization and oscillating patterns in image processing. J. Sci. Comput. **19**(1-3) (2003)
43. Wunderli, T.: On time flows of minimizers of general convex functionals of linear growth with variable exponent in BV space and stability of pseudosolutions. J. Math. Anal. Appl. **364**(2), 591–598 (2010)
44. Vese, L.: A study in the BV space of a denoising-deblurring variational problem. Appl. Math. Optim. **44**, 131–161 (2001)
45. Zhou, X.: An evolution problem for plastic antiplanar shear. Appl. Math. Optim. **25**, 263–285 (1992)