Fall 2016

Matrix Results and Techniques in Quantum Information Science and Related Topics

Diane Christine Pelejo
College of William and Mary - Arts & Sciences, dcpelejo@gmail.com

Follow this and additional works at: https://scholarworks.wm.edu/etd
Part of the Physical Sciences and Mathematics Commons

Recommended Citation
http://doi.org/10.21220/S2CQ13

This Dissertation is brought to you for free and open access by the Theses, Dissertations, & Master Projects at W&M ScholarWorks. It has been accepted for inclusion in Dissertations, Theses, and Masters Projects by an authorized administrator of W&M ScholarWorks. For more information, please contact scholarworks@wm.edu.
Matrix Results and Techniques in Quantum Information Science and Related Topics

Diane Christine Pelejo
Rodriguez, Rizal, Philippines

Master of Science, University of the Philippines, 2011
Bachelor of Science, University of the Philippines, 2009

A Dissertation presented to the Graduate Faculty
of the College of William and Mary in Candidacy for the Degree of
Doctor of Philosophy

Department of Applied Science

The College of William and Mary
Jan 2017
APPROVAL PAGE

This Dissertation is submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Diane Christine Pelejo

Approved by the Committee, November, 2016

Committee Chair
Professor Chi Kwong Li, Mathematics
The College of William and Mary

Professor Gregory D. Smith, Applied Science
The College of William and Mary

Professor Leah B. Shaw, Mathematics
The College of William and Mary

Professor Junping Shi, Mathematics
The College of William and Mary

Professor Yin Tung Poon, Mathematics
Iowa State University
ABSTRACT

In this dissertation, we present several matrix-related problems and results motivated by quantum information theory. Some background material of quantum information science will be discussed in chapter 1, while chapter 7 gives a summary of results and concluding remarks.

In chapter 2, we look at $2^n \times 2^n$ unitary matrices, which describe operations on a closed $n$-qubit system. We define a set of simple quantum gates, called controlled single-qubit gates, and their associated operational cost. We then present a recurrence scheme to decompose a general $2^n \times 2^n$ unitary matrix to the product of no more than $2^n - 1(2^n - 1)$ single qubit gates with small number of controls.

In chapter 3, we address the problem of finding a specific element $\Phi$ among a given set of quantum channels $\mathcal{S}$ that will produce the optimal value of a scalar function $D(\rho_1, \Phi(\rho_2))$, on two fixed quantum states $\rho_1$ and $\rho_2$. Some of the functions we considered for $D(\cdot, \cdot)$ are the trace distance, quantum fidelity and quantum relative entropy. We discuss the optimal solution when $\mathcal{S}$ is the set of unitary quantum channels, the set of mixed unitary channels, the set of unital quantum channels, and the set of all quantum channels.

In chapter 4, we focus on the spectral properties of qubit-qudit bipartite states with a maximally mixed qudit subsystem. More specifically, given positive numbers $a_1 \geq \ldots \geq a_{2n} \geq 0$, we want to determine if there exists a $2^n \times 2^n$ density matrix $\rho$ having eigenvalues $a_1, \ldots, a_{2n}$ and satisfying $\text{tr}_1(\rho) = \frac{1}{n}I_n$. This problem is a special case of the more general quantum marginal problem. We give the minimal necessary and sufficient conditions on $a_1, \ldots, a_{2n}$ for $n \leq 6$ and state some observations on general values of $n$.

In chapter 5, we discuss projection methods and illustrate their usefulness in: (a) constructing a quantum channel, if it exists, such that $\Phi(\rho^{(1)}) = \sigma^{(1)}, \ldots, \Phi(\rho^{(k)}) = \sigma^{(k)}$ for given $\rho^{(1)}, \ldots, \rho^{(k)} \in \mathcal{D}_n$ and $\sigma^{(1)}, \ldots, \sigma^{(k)} \in \mathcal{D}_m$, (b) constructing a multipartite state $\rho$ having a prescribed set of reduced states $\rho_1, \ldots, \rho_r$ on $r$ of its subsystems, (c) constructing a multipartite state $\rho$ having prescribed reduced states and additional properties such as having prescribed eigenvalues, prescribed rank or low von Neuman entropy; and (d) determining if a square matrix $A$ can be written as a product of two positive semidefinite contractions.

In chapter 6, we examine the shape of the Minkowski product of convex subsets $K_1$ and $K_2$ of $\mathbb{C}$ given by $K_1K_2 = \{ab : a \in K_1, b \in K_2\}$, which has applications in the study of the product numerical range and quantum error-correction. In [81], it was conjectured that $K_1K_2$ is star-shaped when $K_1$ and $K_2$ are convex. We give counterexamples to show that this conjecture does not hold in general but we show that the set $K_1K_2$ is star-shaped if $K_1$ is a line segment or a circular disk.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acknowledgments</td>
<td>iv</td>
</tr>
<tr>
<td>Dedication</td>
<td>v</td>
</tr>
<tr>
<td>List of Tables</td>
<td>vi</td>
</tr>
<tr>
<td>List of Figures</td>
<td>vii</td>
</tr>
</tbody>
</table>

## CHAPTER

### Notations

1. Preliminary Information
   1.1 State Space and Observables
   1.2 Composite Systems
   1.3 Evolution of a System
   1.4 Quantum Channels
   1.5 Scalar Functions on Quantum States

2. Decomposition of Quantum Gates
   2.1 Introduction
   2.2 Two-qubit and Three-qubit cases
   2.3 General Scheme
   2.4 Total Number of Controls and Comparison to a Previous Study
   2.5 Concluding Remarks and Future Research

3. Optimal Bounds on Functions of Quantum States under Quantum Channels
   3.1 Introduction
   3.2 Schur Convex Functions
3.3 Fidelity, relative entropy, and other functions .................. 50
3.4 Proof of Theorem 3.3.3 ........................................... 60
3.5 Concluding remarks and further research .......................... 66
4 Bipartite Qubit-Qudit States with Maximally Mixed Reduced State .... 68
  4.1 Introduction .................................................... 68
  4.2 Some Necessary Eigenvalue Inequalities .......................... 70
  4.3 Low Dimension Solutions ...................................... 76
  4.4 Further Remarks .............................................. 83
5 Projection Methods .................................................. 87
  5.1 Introduction ................................................... 87
  5.2 Quantum Channel Construction ................................ 90
    5.2.1 Projection Operators .................................... 91
    5.2.2 Numerical Experiments ................................. 97
  5.3 Quantum States with Prescribed Reduced States and Prescribed Eigenvalues ....................................... 104
    5.3.1 Projection Operators .................................... 107
    5.3.2 Numerical Experiments ................................. 109
  5.4 Bipartite States with Prescribed Reduced States and Rank .......... 112
    5.4.1 Constructions of a Low Rank Solution .................. 114
    5.4.2 Numerical Experiments ................................. 124
  5.5 Bipartite States with Prescribed Reduced States and Low Entropy ... 126
  5.6 Product of Two Positive Contractions ........................... 130
    5.6.1 Characterizations ....................................... 131
    5.6.2 Alternating projections and numerical examples .......... 144
  5.7 Conclusion ................................................... 150
ACKNOWLEDGMENTS

First and foremost, I would like to thank my adviser Dr. Chi-Kwong Li for his patience and generosity in guiding me through my doctoral studies. Thank you for all the words of encouragement, all the mathematical knowledge and techniques you have imparted to me and most of all thank you for all the opportunities you have opened up for me to improve myself as a mathematician.

I would also like to express my gratitude to all my research collaborators, Dr. Kuo-Zhong Wang, Dr. Henry Wolkowicz, Dr. Yuen-Lam Voronin, Dr. Dmitriy Drusvyatskiy, Dr. Xuefeng Duan and Dr. Yiu-Tung Poon, for the fruitful discussions that lead to the results presented in this dissertation.

Thank you to the College of William and Mary and the Office of Graduate Studies and Research for providing resources for students like me to succeed.

The Reves Center for International Studies has been of excellent assistance when it comes to matters regarding my status as an international student.

I owe the Applied Science and the Mathematics departments my deepest gratitude for the assistantships that they offered so that I can support myself financially while pursuing my PhD. Thank you to the APSC and Mathematics department heads Dr. Christopher Del Negro and Dr. George Rublein, respectively. Special thanks also to the administrators Ms. Rosie Fox, Ms. Lydia Whitaker, Ms. Lianne Ashburne and Ms. Davina Santos.

I would also like to thank my college alma mater, the University of the Philippines Diliman, and my UPD professors Dr. Agnes Paras, Dr. Marian Roque, Dr. Jose Balmaceda, Dr. Issa Masangkay, Dr. Carlene Arceo, Dr. Noli Reyes, Dr. Fidel Nemenzo and Dr. Julius Basilia for the training in Mathematics that prepared me for my doctoral studies. Special thanks to Dr. Dennis Merino for being a mentor and for recommending W&M to me when I was looking for a graduate program in the USA.

Finally I would like to thank my family and my friends for their emotional support and for serving as my inspiration during my journey. Thank you Dr. Tina Picardo for being my first friend in the USA and for introducing me to your wonderful family. And thank you to my partner Mr. Ryan Redmon who has been my rock for the past 3 years.
I dedicate this dissertation to my parents Christian Palileo and David Pelejo and to my brothers Dearborn Tria, Ian Dave Pelejo and Harvey Dexter Pelejo.
# LIST OF TABLES

2.1 Scheme table for decomposing 2-qubit quantum gates ........................................... 23  
2.2 Scheme table for decomposing 3-qubit quantum gates ........................................... 24  
2.3 Partial scheme table for annihilating the lower left block of a 4-qubit quantum gate .......................................................... 36  
2.4 Comparison of total cost of decomposing \( n \)-qubit quantum gates into a product of controlled gates using the scheme presented in this chapter \((T_1(n))\) and that of \([90] \ (T_2(n))\). .......................................................... 41  
5.1 Using DR algorithm; for solving huge problems .................................................. 99  
5.2 Using DR algorithm; with \([m \ n \ k \ mn \ toler \ iterlimit] = [30 \ 30 \ 16 \ 900 \ 1e-14 \ 3500]\); max/min/mean iter and number rank steps for finding max-rank of \(P\). The 3500 here means 9 decimals accuracy attained for last step. 101  
5.3 Using MAP algorithm; with \([m \ n \ k \ mn \ toler \ iterlimit] = [30 \ 30 \ 16 \ 900 \ 1e-14 \ 3500]\); max/min/mean iter and number rank steps for finding max-rank of \(P\). The 3500 mean-itors means max iterlimit reached; low accuracy attained. 102  
5.4 Using MAP algorithm with facial reduction for decreasing the rank ................. 103  
5.5 Using DR algorithm for rank constrained problems with ranks \(r_s\) to \(r_f\) ........ 103  
5.6 Using DR algorithm for rank constrained problem instance one in Table 5.5 with \(m = n = 12, k = 9, r = 15\) and starting constrained rank 20 till final successful constrained rank 7; feasibility failed for constrained rank 6 with iteration limit 3,500. .......................................................... 104  
5.7 Low rank solutions obtained using Algorithms 5.4.3, 5.4.5, and 5.4.8 ............... 125  
5.8 Low rank solution from Algorithm 5.4.1 using the solutions from Algorithms 5.4.3 and 5.4.5 as starting point. .......................................................... 125  
5.9 Low rank solutions obtained using Proposition 5.4.3 and Algorithms 5.4.5, and 5.4.8 .......................................................... 126  
5.10 Low rank solutions obtained Algorithm 5.4.1 utilizing the solutions from Proposition 5.4.3 and Algorithms 5.4.5, and 5.4.8 as starting point. .................. 126  
5.11 Low rank solutions obtained using Proposition 5.4.3 and Algorithms 5.4.5, and 5.4.8 as starting point. .......................................................... 127  
5.12 Low rank solutions obtained Algorithm 5.4.1 utilizing the solutions from Proposition 5.4.3 and Algorithms 5.4.5, and 5.4.8 as starting point. ............... 127
# LIST OF FIGURES

2.1 Circuit diagrams for controlled 2-qubit gates. ........................................ 21
2.2 $n$ versus $\log_{10}(T_2(n) - T_1(n))$ graph ........................................... 41
4.1 LR skew-tableaux of shape $\bar{s}(R)/\bar{s}(P)$ and content $\bar{s}(Q)$ for inequalities (4.38)-(4.43). ................................................................. 84
4.2 LR skew-tableaux of shape $\bar{s}(R)/\bar{s}(P)$ and content $\bar{s}(Q)$ for inequalities (4.44)-(4.55). ................................................................. 86
6.1 Three cases of the Minkowski product of two lines described in Theorem 6.2.4. ................................................................. 160
6.2 Plot of $S_2 = L_1L_1$, where $L_1 = \text{Co}(e^{i\frac{\pi}{4}}, e^{-i\frac{\pi}{4}})$. ................................................................. 165
6.3 Sets described in Example 6.3.1. ................................................................. 166
6.4 The set $P = K_1K_1$ in Example 6.3.2 does not contain the segment $\text{Co}(1, \alpha_{\frac{\pi}{2}})$. ................................................................. 167
6.5 The following figures illustrate the canonical representations of a line segment $K_1 = \text{Co}(\hat{a}, \hat{b})$ and a convex set $K_2$ described in Theorem 6.4.3 . . . . 171
6.7 ................................................................. 173
6.8 ................................................................. 174
6.9 ................................................................. 175
6.10 The product set $(\text{Co}(1, 2e^{i\frac{11\pi}{12}}) \cup \text{Co}(1, 2e^{-i\frac{11\pi}{12}})) \cdot D(1, \frac{1}{2})$ is not simply connected. ................................................................. 178
Notations

The following notations will be used throughout this thesis.

\( \mathbb{Z}, \mathbb{R}, \mathbb{C} \) the sets of integers, real numbers and complex numbers

\( \text{Re}(z), \text{Im}(z), \bar{z} \) real part, imaginary part and conjugate of \( z \in \mathbb{C} \)

\( i \) the imaginary number \( \sqrt{-1} \)

\( \mathbb{R}^n \) \( n \)-dimensional real vector space

\( \mathbb{C}^n \) \( n \)-dimensional complex vector space

\( |x\rangle \) a vector in a Hilbert space

\( \langle x| \) the conjugate transpose of the vector \( |x\rangle \)

\( |0\rangle, |1\rangle, \ldots, |n-1\rangle \) standard basis vectors of \( \mathbb{C}^n \)

\( \langle u|v \rangle \) inner product of \( |u\rangle \) and \( |v\rangle \)

\( |u\rangle \otimes |v\rangle \) tensor product of \( |u\rangle \) and \( |v\rangle \), also denoted \( |u\rangle|v\rangle \) or \( |uv\rangle \)

\( \mathbb{R}^{m \times n} \) set of \( m \times n \) matrices with real entries

\( \mathbb{C}^{m \times n} \) set of \( m \times n \) matrices with complex entries

\( A^T, \overline{A}, A^* \) transpose, conjugate, conjugate transpose of matrix \( A \)

\( E_{ij} \) the matrix whose \( (i,j)^{th} \) entry is 1 and all other entries zero

\( I_n \) or \( I \) the \( n \times n \) identity matrix

\( \text{diag}(d_1, \ldots, d_n) \) a diagonal matrix with diagonal entries \( d_1, \ldots, d_n \)

\( \text{eig}^\downarrow(X) \) \( n \)-tuple of eigenvalues of \( X \), arranged in nonincreasing order

\( \text{eig}^\uparrow(X) \) \( n \)-tuple of eigenvalues of \( X \), arranged in nondecreasing order

\( \Lambda^\downarrow(X), \Lambda^\uparrow(X) \) diagonal matrix whose diagonal entries are \( \text{eig}^\downarrow(X), \text{eig}^\uparrow(X) \)
\( \text{tr}(A) \) the trace of \( A \)
\( \text{det}(A) \) the determinant of \( A \)
\( \text{rank}(A) \) the rank of \( A \)
\( \lambda_i(A) \) the \( i^{th} \) largest eigenvalue of \( A \)
\( s_i(A) \) the \( i^{th} \) largest singular value of \( A \)
\( U_n \) set of \( n \times n \) unitary matrices
\( H_n \) set of \( n \times n \) hermitian matrices
\( PSD_n \) set of positive semidefinite matrices
\( D_n \) set of density matrices
\( A \preceq B \) \( A \) is majorized by \( B \), where \( A, B \in H_n \)
\( A \oplus B \) the direct sum of matrices \( A \) and \( B \)
\( A \otimes B \) the tensor product of matrices \( A \) and \( B \)
\( \Omega_n \) the set \( \{ (a_1, \ldots, a_n) \in \mathbb{R}^n \mid a_1 \geq \cdots \geq a_n \geq 0, \sum_{j=1}^{n} a_j = 1 \} \)
\( \partial S \) the boundary of the set \( S \)
\( \text{Co}(S) \) convex hull of the set \( S \)
\( S^c \) complement of the set \( S \)
\( S_1 \cup S_2, S_1 \cap S_2, S_1 \setminus S_2 \) union, intersection and difference of two sets \( S_1 \) and \( S_2 \)
\( \mathcal{C}(\Phi) \) Choi matrix representation of the linear map \( \Phi \)
\( H(\rho) \) von Neumann entropy of positive semidefinite matrix \( \rho \)
\( H(\rho||\sigma) \) relative entropy of two positive semidefinite matrices \( \rho, \sigma \)
\( F(\rho,\sigma) \) fidelity between two positive semidefinite matrices \( \rho, \sigma \)
\( \delta_{st} \) Kronecker delta function
Quantum information science has been a source of many research topics in the past 30 years [16]. Matrix and operator theory has a significant role in the development of this field. In particular, Hilbert spaces, Hermitian operators, positive semidefinite operators, unitary transforms and trace-preserving completely-positive maps are some of the mathematical tools used by every textbook to lay out the foundations of quantum information.

In this section, we present these concepts, based on the Copenhagen interpretation of quantum mechanics, that are relevant to the problems discussed in this dissertation. We also define terms and establish notations.

1.1 State Space and Observables

The first postulate of quantum mechanics states that any isolated physical system $X$ is associated to a *Hilbert space* —a complex vector space with inner product—called the *state space* of the system. The isolated system is completely described by a unit vector $|\psi\rangle$ in the state space [79]. This unit vector is called the *state vector* or *state* of the system. Note that throughout this dissertation, we will focus on systems with finite-dimensional
state spaces and we will denote the $n$–dimensional complex space by $\mathbb{C}^n$. We will also denote the conjugate transpose of the vector $|x\rangle$ by $\langle x|$.

One of the fundamental differences between a quantum mechanical system and a classical system is that the state of a classical system can only be one of $n$ state vectors, while a quantum mechanical system can be in a superposition of those $n$ states. Several quantum algorithms that are proven to be more efficient than classical ones rely on this feature of quantum systems \[26, 86\].

As the name suggests, the state of a system contains all information regarding its physical properties. However, due to the well-known uncertainty principle introduced by physicist Werner Heisenberg, it is not possible for an observer to measure all of these properties with absolute certainty in the outcome of all variables measured. To describe this phenomenon mathematically, we consider a physical quantity or observable that has $n$ classical outcomes and associate to it an $n \times n$ Hermitian matrix $A$ with spectral decomposition $A = \sum_{j=1}^{n} \lambda_j |x_j\rangle\langle x_j|$. The eigenvectors $|x_1\rangle, \ldots, |x_n\rangle$ represent the $n$ classical states, while the eigenvalues $\lambda_1, \ldots, \lambda_n$ of $A$ represent the possible outcomes of measuring the observable that $A$ represents. If the state $|\psi\rangle$ of the system is an eigenvector of $A$, say $|\psi\rangle = e^{i\theta} |x_j\rangle$, then the measurement outcome is $\lambda_j$. If we do the same measurement on several copies of $|\psi\rangle$, the outcome will always be the same. However if $|\psi\rangle$ is not an eigenvector of $A$, say $|\psi\rangle = c_1 |x_1\rangle + c_2 |x_2\rangle + \cdots + c_n |x_n\rangle$ is a superposition of $|x_1\rangle, \ldots, |x_n\rangle$, then the measurement outcome is $|x_1\rangle$ with probability $|c_1|^2$, $|x_2\rangle$ with probability $|c_2|^2$ and so on. Moreover, after performing the measurement of the system, the state of the system immediately collapses from $|\psi\rangle$ to the observed eigenvector $|x_j\rangle$.

The simplest quantum mechanical system is a qubit which can be associated to the
two-dimensional complex Hilbert space $\mathbb{C}^2$. The standard basis for $\mathbb{C}^2$ consists of

$$|0\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{and} \quad |1\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

The state of a qubit can be represented by $a|0\rangle + b|1\rangle$, where $a, b \in \mathbb{C}$ such that $|a|^2 + |b|^2 = 1$. There are several physical systems that realize qubits. Some examples are given by the polarization of a photon (vertical or horizontal), the spin of an electron (spin up or spin down) and the energy state of an electron orbiting a single atom (ground state or excited state). On the other hand, a system associated to a more general complex Hilbert space $\mathbb{C}^d$ is referred to as a qudit system.

Suppose that we have a quantum system associated to $\mathbb{C}^n$ whose state is only known to be equal to $|\psi_j\rangle$ with probability $p_j$ for $j = 1, \ldots, n$. If $n = 1$, then we say that the system is in a pure state. Otherwise, the system is said to be mixed and its state is described by a density matrix of the form

$$\rho = \sum_{j=1}^{n} p_j |\psi_j\rangle \langle \psi_j|,$$  \hspace{1cm} (1.2)

for some orthonormal set $\{|\psi_1\rangle, \ldots, |\psi_n\rangle\}$ and $p_1, \ldots, p_n \in \mathbb{R}_{+,0}$ such that $\sum_{j=1}^{n} p_j = 1$. In particular, if $p_1 = \cdots = p_n = \frac{1}{n}$, then we say that the system is maximally mixed. Note that when the system is in a pure state $|\psi\rangle$, we can still represent the state by the rank one density matrix $\rho = |\psi\rangle \langle \psi|$. 

Throughout this text, we will denote the set of $m$-by-$n$ complex matrices by $\mathbb{C}^{m \times n}$, the set of $n$-by-$n$ Hermitian matrices by $\mathcal{H}_n$, the set of $n \times n$ positive semidefinite matrices by $\text{PSD}_n$ and the set of $n \times n$ density matrices by $\mathcal{D}_n$. The $n$-by-$n$ identity matrix will be written as $I_n$ or simply $I$. We also let $\{|0\rangle, \ldots, |n-1\rangle\}$ be the standard basis vectors.
for $\mathbb{C}^n$. For any $A \in \mathcal{H}_n$, we will denote the $n$–tuple of eigenvalues of $A$, arranged in nonincreasing order (respectively, nondecreasing order), by $eig^\downarrow(A)$ (resp., $eig^\uparrow(A)$).

In the next section, we will look at systems that are made of multiple subsystems and present the mathematical tools used to describe relations between the subsystems and the overall state of the system.

### 1.2 Composite Systems

The interaction of two or more quantum systems produce interesting quantum effects [84]. Perhaps the most intriguing feature of quantum mechanics is the concept of quantum entanglement wherein the combined state of two or more systems is not directly described by the individual states of its component systems and vice versa. Mathematically, given two quantum systems $X_1$ and $X_2$ with respective state spaces $\mathbb{C}^m$ and $\mathbb{C}^n$, we can consider their combined system $X = (X_1, X_2)$. The bipartite system $X$ is associated to the space $\mathbb{C}^{mn}$. To describe relations between the global state and the state of its subsystems, we define the tensor product operation on matrices as follows.

Suppose $A \in \mathbb{C}^{m \times n}$ and $B \in \mathbb{C}^{k \times l}$, then the tensor product of $A$ and $B$ is the matrix $A \otimes B \in \mathbb{C}^{mk \times nl}$ such that

$$
\begin{array}{c}
\text{if } A = \begin{bmatrix}
a_{11} & \cdots & a_{1n} \\
\vdots & \ddots & \vdots \\
a_{m1} & \cdots & a_{mn}
\end{bmatrix}, \text{ then } A \otimes B = \\
\begin{bmatrix}
a_{11}B & \cdots & a_{1n}B \\
\vdots & \ddots & \vdots \\
a_{m1}B & \cdots & a_{mn}B
\end{bmatrix}
\end{array}
$$

(1.3)

In particular, if $|\psi_1\rangle \in \mathbb{C}^m = \mathbb{C}^{m \times 1}$ and $|\psi_2\rangle \in \mathbb{C}^n = \mathbb{C}^{n \times 1}$, we will sometimes denote $|\psi_1\rangle \otimes |\psi_2\rangle$ by $|\psi_1\rangle|\psi_2\rangle$ or $|\psi_1\psi_2\rangle$.

For independent quantum systems $X_1$ and $X_2$, the state of the bipartite system $X = (X_1, X_2)$ can be described as a tensor product $|x_1\rangle \otimes |x_2\rangle$ of the state vectors $|x_1\rangle$ of $X_1$ and
$|x_2\rangle$ of $X_2$. Note however that not all elements of $\mathbb{C}^{mn}$ can be written as a tensor product $|x_1\rangle \otimes |x_2\rangle$ [85]. State vectors that cannot be written as a tensor product represent (pure) states of entangled systems.

In the density matrix formulation, we say that two systems $X_1$ and $X_2$ are independent if the state $\rho$ of their combined system $X = (X_1, X_2)$ is the tensor product of their states $\rho_{X_1}$ and $\rho_{X_2}$. That is, $\rho = \rho_{X_1} \otimes \rho_{X_2}$. More generally, if $\rho = \sum_{j=1}^{k} p_j \sigma_j \otimes \gamma_j$ for some probability vector $p = [p_j]$ and density matrices $\sigma_j \in \mathcal{D}_m$ and $\gamma_j \in \mathcal{D}_n$, then $\rho$ is said to be separable. A density matrix that is not separable represents an entangled (mixed) state.

We can obtain information on the state of a subsystem by performing an operation called partial trace on the state of the whole system.

For an ordered pair of integers $(m,n)$ and a matrix $A \in \mathbb{C}^{mn \times mn}$, we define the first and second partial trace of $A$, with respect to $(m,n)$, as follows

$$
\text{tr}_1(\rho) = \sum_{s=1}^{m} \left( |x_s\rangle \otimes I_n \right) A \left( |x_s\rangle \otimes I_n \right) \quad \text{and} \quad \text{tr}_2(\rho) = \sum_{t=1}^{n} \left( I_m \otimes |y_t\rangle \right) \rho \left( I_m \otimes |y_t\rangle \right)
$$

where $|x_1\rangle, \ldots, |x_m\rangle$ forms an orthonormal basis of $\mathbb{C}^m$ and $|y_1\rangle, \ldots, |y_n\rangle$ forms an orthonormal basis of $\mathbb{C}^n$.

From this definition, it is clear that if $\rho = \zeta \otimes \sigma$ for some $\zeta \in \mathcal{D}_m$ and $\sigma \in \mathcal{D}_n$, as is the case for independent bipartite systems, then $\text{tr}_1(\rho) = \sigma$ and $\text{tr}_2(\rho) = \zeta$. In general, if $\rho = [\rho_{st}]_{1 \leq s,t \leq m}$ for $\rho_{st} \in M_n$, then

$$
\text{tr}_1(A) = A_{11} + \ldots + A_{mm} \quad \text{and} \quad \text{tr}_2(A) = \left( \text{tr}(A_{ij}) \right)_{1 \leq i,j \leq m}
$$

Let $X_1$ be associated to $\mathbb{C}^m$ and $X_2 \in \mathbb{C}^n$. Suppose that the bipartite system $X = (X_1, X_2)$ is in the state $\rho_X$. Then the respective reduced state $\rho_{X_1}$ and $\rho_{X_2}$ of $X_1$ and $X_2$ are given
by
\[
\text{tr}_2(\rho_X) = \rho_{X_1} \quad \text{and} \quad \text{tr}_1(\rho_X) = \rho_{X_2}
\] (1.6)

We say that \( \rho_X \) is an extension of \( \rho_{X_1} \), and also of \( \rho_{X_2} \), while the two latter density matrices are called reduced states (or marginal states) of the former. Given \( \zeta \in \mathcal{D}_m \) and \( \sigma \in \mathcal{D}_n \), there may be several extensions \( \rho \in \mathcal{D}_{mn} \) for \( \zeta \) and \( \sigma \). In fact, the set
\[
\{ \rho \in \mathcal{D}_{mn} \mid \text{tr}_1(\rho) = \sigma \text{ and } \text{tr}_2(\rho) = \zeta \}
\] (1.7)
is a compact convex set. Moreover, the set of eigenvalues of elements of set (1.7) is a convex polytope. In Chapter 4, we study the minimal set of inequalities that \((a_1, \ldots, a_{2n}) \in \Omega_{2n}\) must satisfy for there to exist a density matrices \( \rho \in \mathcal{D}_{2n} \) satisfying \( \text{tr}_1(\rho) = \frac{1}{n}I_n \) and \( \text{eig}^+(\rho) = (a_1, \ldots, a_{2n}) \).

One can extend the definition of a partial trace map to states of a multipartite system \( X = (X_1, \ldots, X_k) \). Suppose the state space of the subsystem \( X_j \) is \( \mathbb{C}^{n_j} \). Let \( J = \{j_1, \ldots, j_r\} \subseteq \{1, \ldots, k\} \) and let \( J^c = \{1, \ldots, k\} \setminus J \). We define the partial trace map
\[
\text{tr}_{J^c} : \mathcal{H}_{n_1 \cdots n_k} \rightarrow \mathcal{H}_{n_{j_1} \cdots n_{j_r}}
\] as the linear map satisfying
\[
\text{tr}_{J^c}(A_1 \otimes \cdots \otimes A_k) = \left( \prod_{s \in J^c} \text{tr}(A_s) \right) A_{j_1} \otimes \cdots \otimes A_{j_r}
\] (1.8)
for any \( A_j \in \mathcal{H}_{n_j} \). Then the reduced state \( \rho_{X_J} \) of the subsystem \( X_J = (X_{j_1}, \ldots, X_{j_r}) \) is given by \( \rho_{X_J} = \text{tr}_{J^c}(\rho_X) \). The Matlab script \texttt{parttrace.m} in Appendix A.1 can be used to compute the reduced state of a subsystem given the global state of the multipartite system.

We can define the set
\[
\{ \rho \in \mathcal{D}_{n_1 \cdots n_k} \mid \text{tr}_{J_t}(\rho) = \rho_{X_{J_t}} \text{ for } t = 1, \ldots, \ell \}
\] (1.9)
for given subsets $J_1, \ldots, J_\ell \subseteq \{1, \ldots, k\}$, and density matrices $\rho_{X_{J_t}}$ of size $\prod_{s \in J_t} n_s$ for $t = 1, \ldots, \ell$. Note that if there exists $t_1$ and $t_2$ such that $J_{t_1} \cup J_{t_2} \neq \emptyset$, then the set 1.9 may or may not be empty. The problem of determining if a given set of density matrices are compatible as reduced states of a global state is a special case of the quantum marginal problem [84]. In Chapter 5, we discuss a numerical method called alternating projection to find a solution to such a problem.

### 1.3 Evolution of a System

The second postulate of quantum mechanics states that the evolution of a closed quantum system is described by the Schrödinger equation

$$i\hbar \frac{d|\psi\rangle}{dt} = \hat{H}|\psi\rangle$$

where $\hat{H}$ is a Hermitian matrix called the *Hamiltonian* of the system. In discrete time, this says that the state $|\psi'\rangle$ of the system at time $t'$ is related to an earlier state $|\psi\rangle$ at time $t$ via a unitary transformation $U(t,t')$, that is

$$|\psi'\rangle = U(t,t')|\psi\rangle$$

or in terms of the density matrix formulation,

$$\rho' = U(t_1, t_2)\rho U(t_1, t_2)^*.$$  \hspace{1cm} (1.12)

In matrix theory, a unitary transformation is represented by a *unitary* matrix $U$. That is $UU^* = I$. We will denote the set of unitary matrices by $U_n$.

In computer science, circuits are made of logic gates that are applied sequentially to
perform a particular task. We can view a task as a function on the register of the computer. Given a set of available functions (logic gates), one wishes to express a general function as a composition of these available functions in the circuit. Analogously, in quantum information science, we are interested in building a quantum circuit using quantum logic gates as building blocks to perform desired quantum operations. Since unitary matrices describe operations on a closed quantum system, we wish to express a general unitary matrix as a product of simple quantum gates.

Unitary matrices of the form

\[ I_{n_1} \otimes \cdots \otimes I_{n_{j-1}} \otimes U_j \otimes I_{n_{j+1}} \otimes \cdots \otimes I_{n_k} \]  

are ideal quantum operations on multipartite systems \((X_1, \ldots, X_k)\) with state spaces \(\mathbb{C}^{n_1}, \ldots, \mathbb{C}^{n_k}\). To see this, consider the effect of (1.13) on the vector \(|\psi\rangle = |\psi_1\rangle \otimes \cdots \otimes |\psi_k\rangle \in \mathbb{C}^{n_1 \cdots n_k}\). The result is

\[ |\psi'\rangle = |\psi_1\rangle \otimes |\psi_{j-1}\rangle \otimes |\psi'_j\rangle \otimes |\psi_{j+1}\rangle \cdots \otimes |\psi_k\rangle \]  

wherein the \(j^{th}\) component has been altered while the other components have not. Unitary matrices of the form (1.13) are called local quantum gates or free quantum gates because we do not need knowledge of the other component states to perform an operation on the \(j^{th}\) state. However, if we want a transformation that changes the state of the \(j^{th}\) system only when the other component systems are known to be in particular states, then such operations will be more costly. Such operations are referred to as controlled quantum gates. In Chapter 2, we define controlled-quantum gates and their associated cost and describe a scheme to decompose a general \(n\)-qubit unitary matrix — that is \(U \in \mathcal{U}_2^n\) — into a product of controlled gates with the aim of reducing the cost from another scheme.
found in the literature.

1.4 Quantum Channels

In the preceding section, we introduced unitary transformations that describe operations on closed quantum systems. In this section, we consider a general quantum operation \( \Phi : \mathbb{C}^{n \times n} \rightarrow \mathbb{C}^{m \times m} \) that maps a quantum state to another quantum state. Such a map must send a density matrix to another density matrix. If we assume \( \Phi \) is linear, this implies that \( \Phi \) must be trace-preserving, that is

\[
\text{tr}(\Phi(X)) = \text{tr}(X) \quad \text{for all } X \in \mathbb{C}^{n \times n}
\]  

(1.15)

and must also preserve positive semidefiniteness. In fact, quantum operations must be completely positive so that the tensor of two such maps also preserve positive semidefiniteness. We will define what this means in the following.

Let \( \Phi : \mathbb{C}^{m \times m} \rightarrow \mathbb{C}^{n \times n} \) and \( \Psi : \mathbb{C}^{r \times r} \rightarrow \mathbb{C}^{s \times s} \) be two linear maps. We can define a new linear map \( \Psi \otimes \Phi : \mathbb{C}^{rm \times rm} \rightarrow \mathbb{C}^{sn \times sn} \) satisfying \( \Psi \otimes \Phi(X \otimes Y) = \Psi(X) \otimes \Phi(Y) \) for any \( X \in \mathbb{C}^{r \times r} \) and \( Y \in \mathbb{C}^{m \times m} \). Denote the identity map on \( \mathbb{C}^{n \times n} \) by \( 1_n \). We say that the map \( \Phi \) is completely positive if for any \( k \in \mathbb{Z}^+ \), the map \( 1_k \otimes \Phi \) satisfies

\[
1_k \otimes \Phi(X) \in \text{PSD}_{nk} \quad \text{whenever } X \in \text{PSD}_{mk}
\]  

(1.16)

If \( \Phi \) is both trace-preserving and completely positive, then \( \Phi \) is called a quantum channel. Note that if \( \Phi_1 \) and \( \Phi_2 \) are two quantum channels, the map \( \Phi_1 \otimes \Phi_2 \) is also a quantum channel. Quantum channels represent a more general set of quantum operations that can be observed in open quantum systems. These are quantum systems that interact with
other quantum systems. Such interaction causes *decoherence* or loss of information due to the quantum noise brought about by the interaction of the system with its environment.

A quantum channel \( \Phi : \mathbb{C}^{n \times n} \rightarrow \mathbb{C}^{m \times m} \) has a convenient operator-sum representation due to Kraus [57] given by

\[
\Phi(\rho) = \sum_{j=1}^{k} F_j \rho F_j^* \tag{1.17}
\]

for some \( F_j \in \mathbb{C}^{m \times n} \) for all \( j \) and \( \sum_{j=1}^{k} F_j^* F_j = I_n \). The operators \( F_1, \ldots, F_k \) are called *error operators*. These error operators are significant in the study of quantum error-correction.

Another useful representation of \( \Phi \) is the *Stinespring representation* [88] which states that there exist a linear isometry \( P \in \mathbb{C}^{mp \times n} \), that is \( P^* P = I_n \) such that for all \( \rho \in \mathbb{C}^{n \times n} \),

\[
\Phi(\rho) = \text{tr}_2(P \rho P^*) \tag{1.18}
\]

where the partial traces are with respect to \((m, p)\).

Finally, \( \Phi \) is associated to a unique positive semidefinite \( mn \)-by-\( mn \) matrix

\[
\mathcal{C}(\Phi) = [\Phi(E_{ij})]_{i,j=1}^{m} = \begin{bmatrix}
\Phi(E_{11}) & \cdots & \Phi(E_{1n}) \\
\vdots & \ddots & \vdots \\
\Phi(E_{n1}) & \cdots & \Phi(E_{nn})
\end{bmatrix} \tag{1.19}
\]

called the *Choi matrix of* \( \Phi \) [20]. The trace preserving property of \( \Phi \) ensures that

\[
\text{tr}_2(\mathcal{C}(\Phi)) = \left[\text{tr}(\Phi(E_{ij}))\right]_{i,j=1}^{m} = I_m. \tag{1.20}
\]

One of the problems we will consider in Chapter 5 is the quantum channel interpolation problem. That is, given \( \rho_1, \ldots, \rho_k \in \mathcal{D}_m \) and \( \sigma_1, \ldots, \sigma_k \in \mathcal{D}_n \), we wish to determine if there is a quantum channel \( \Phi \) such that \( \Phi(\rho_i) = \sigma_i \).
1.5 Scalar Functions on Quantum States

There are several functions defined on quantum states that are of interest in quantum information science. These functions reveal properties of quantum systems or relations between quantum systems. We discuss some of these functions that will appear in this dissertation.

Schatten \( p \)-norm

For any \( X \in \mathbb{C}^{m \times n} \) and any \( p \geq 1 \), the Schatten–\( p \) norm of \( X \) is defined and denoted by

\[
||X||_p = \begin{cases} 
\left[ \text{tr} \left( (A^*A)^{p/2} \right) \right]^{\frac{1}{p}} & \text{if } p < \infty \\
\max \{ \sqrt{\langle x|A^*Ax \rangle} \mid \langle x|x \rangle = 1 \} & \text{if } p = \infty
\end{cases}
\]

If \( X \) has singular values \( s(X) = (s_1, \ldots, s_k) \), then \( ||X||_p \) is just the \( \ell_p \)-norm of \( s(X) \), i.e.,

\[
||X||_p = \begin{cases} 
||s(X)||_{\ell_p} = \left( \sum_{j=1}^{k} s_j^p \right)^{\frac{1}{p}} & \text{if } p < \infty \\
s_1 & \text{if } p = \infty
\end{cases}
\]

When \( p = 1 \), we get the trace norm, while \( p = 2 \) gives the Frobenius/Hilbert-Schmidt norm and when \( p = \infty \), we get the spectral norm of \( X \). Note that \( ||\cdot||_p \) is invariant under partial isometry, that is

\[
||U^*XV||_p = ||X||
\]

for any \( U, V \) with appropriate sizes such that \( U^*U = I \) and \( V^*V = I \). In Chapter 3, we will discuss the optimal values of \( ||\rho_1 - \Phi(\rho_2)||_p \) when \( \rho_1 \) and \( \rho_2 \) are fixed quantum states and the optimum is taken over all quantum channels \( \Phi \) contained in a given set.
The von Neumann Entropy

In classical information theory, the Shannon entropy of a probability vector \( p = (p_1, \ldots, p_n) \) given by \(- \sum_{j=1}^{n} p_j \log p_j\) can be viewed as a measure of the amount of uncertainty in a random experiment described by \( p \), or equivalently, the amount of information gained by learning the result of the experiment [92].

The quantum analog of the Shannon entropy is the von Neumann entropy. The von Neumann entropy of a state \( \rho \in D_n \), whose eigenvalues are \( a_1, \ldots, a_n \), is defined to be

\[
H(\rho) = -\text{tr}(\rho \log \rho) = -\sum_{j=1}^{n} a_j \log a_j, \tag{1.24}
\]

where the logarithm is in base 2 and we take \( 0 \log 0 = 0 \) by convention. For any \( \rho \in D_n \),

\[
0 \leq H(\rho) \leq \log n = H \left( \frac{1}{n} I_n \right), \tag{1.25}
\]

and \( H(\rho) = 0 \) if and only if \( \text{rank}(\rho) = 1 \), i.e. \( \rho \) is a pure state. Intuitively, there is less uncertainty when \( \rho \) is a pure state and maximum uncertainty when \( \rho \) is maximally mixed.

It is known that for any bipartite state \( \rho \in D_{mn} \),

\[
H(\rho) \leq H(\text{tr}_1(\rho)) + H(\text{tr}_2(\rho)), \tag{1.26}
\]

where equality in the first equation is satisfied when \( \rho = \text{tr}_1(\rho) \otimes \text{tr}_2(\rho) \). This is referred to as the subadditivity of \( H(\cdot) \). In addition to this, \( H(\cdot) \) is also strongly subadditive, that is, for any bipartite state \( \rho \in D_{mn} \) and any tripartite state \( \sigma \in D_{mnr} \),

\[
H(\sigma) + H(\text{tr}_1(\sigma)) \leq H(\text{tr}_1(\sigma)) + H(\text{tr}_3(\sigma)). \tag{1.27}
\]
It is clear from (1.26) that the maximum value of $H(\rho)$ over all elements $\rho$ in the set described in (1.7) is given by $H(\sigma \otimes \zeta)$. However, the minimum value is not easy to compute. In Chapter 5.5, we employ a numerical method to address the problem of finding the minimum value of $H(\rho)$ over all elements of (1.7).

**The Quantum Relative Entropy**

Given $\rho, \sigma \in \mathcal{D}_n$, we define the *quantum relative entropy* of $\rho$ with $\sigma$ to be

$$H(\rho \| \sigma) = \begin{cases} 
\text{tr}(\rho \log \rho - \rho \log \sigma) & \text{if range}(\rho) \subseteq \text{range}(\sigma) \\
\infty & \text{otherwise} 
\end{cases}$$

(1.28)

This quantity is nonnegative for any $\rho, \sigma \in \mathcal{D}_n$. It is also jointly convex in its two inputs, i.e. for any $0 \leq \lambda \leq 1$,

$$H(\lambda \rho_0 + (1 - \lambda)\rho_1 \| \lambda \sigma_0 + (1 - \lambda)\sigma_1) \leq \lambda H(\rho_0 \| \sigma_0) + (1 - \lambda)H(\rho_1 \| \sigma_1).$$

(1.29)

And lastly, it is monotone under any quantum channel $\Phi$. That is,

$$H(\Phi(\rho) \| \Phi(\sigma)) \leq H(\rho \| \sigma).$$

(1.30)

**The Quantum Fidelity Function**

Given two quantum states $\rho_1, \rho_2 \in \mathcal{D}_n$, we define the fidelity between $\rho_1$ and $\rho_2$ by

$$F(\rho_1, \rho_2) = \text{tr}\sqrt{\sqrt{\rho_1} \rho_2 \sqrt{\rho_1}} = \|\sqrt{\rho_1} \sqrt{\rho_2}\|_1.$$

(1.31)
If $\rho_1$ and $\rho_2$ are pure states, say $\rho_1 = |x\rangle\langle x|$ and $\rho_2 = |y\rangle\langle y|$, then $F(\rho_1, \rho_2) = |\langle x|y\rangle|$. For general states $\rho_1, \rho_2$, Uhlmann’s theorem states that

$$F(\rho_1, \rho_2) = \max\{F(\sigma_1, \sigma_2) \mid \text{tr}(\sigma_1) = \rho_1, \text{tr}(\sigma_2) = \rho_2 \text{ and rank}(\sigma_1) = \text{rank}(\sigma_2) = 1\}. \tag{1.32}$$

Recall that a density matrix $\sigma$ is represents a pure state if and only if rank($\sigma$) = 1. An extension of $\rho$ that is a pure state $\sigma$ is called a purification of $\rho$. Thus, we can interpret $F(\rho_1, \rho_2)$ as a measure of how close a purification of $\rho_1$ resembles a purification of $\rho_2$.

In Chapter 3, we will discuss the optimal values of a class of functions $D(\rho_1, \Phi(\rho_2))$ over elements $\Phi$ of a set of quantum channels. This will include the optimal values of $F(\rho_1, \Phi(\rho_2))$ and $H(\rho_1||\Phi(\rho_2))$.

**Majorization**

In this section, we define Schur-convexity and Schur-concavity, which are useful properties of some functions of quantum states. First, we need to define the concept of majorization. Majorization is an important tool in matrix theory that is used to prove several inequalities on certain classes of functions. One may consult the excellent monograph [76] for more information and applications.

Let $a$ and $b$ be two collections of $n$ real numbers, say $a = \{a_1, \ldots, a_n\}$ and $b = \{b_1, \ldots, b_n\}$ such that $a_1 \geq a_2 \geq \cdots \geq a_n$ and $b_1 \geq b_2 \geq \cdots \geq b_n$. We say that $a$ is majorized by $b$, written $a \prec b$, if for all

$$\sum_{j=1}^{n} a_j = \sum_{j=1}^{n} b_j \quad \text{and} \quad \sum_{j=1}^{k} a_j \leq \sum_{j=1}^{k} b_j \text{ for all } k = 1, \ldots, n - 1. \tag{1.33}$$

Let $A$ and $B$ be two $n \times n$ Hermitian matrices. We say that $A$ is majorized by $B$, written $A \prec B$, if $\text{eig}^\downarrow(A) \prec \text{eig}^\downarrow(B)$. 

16
We now define Schur-convexity and concavity. A function $f : \mathbb{R}^n \to \mathbb{R}$ is Schur-convex if $f(x) \leq f(y)$ whenever $x \prec y$. It is strictly Schur convex if $f(x) < f(y)$ whenever $x \prec y$ and $x \neq y$. Similarly, $f$ is Schur-concave if $f(x) \geq f(y)$ whenever $x \prec y$. It is strictly Schur concave if $f(x) > f(y)$ whenever $x \prec y$ and $x \neq y$.

The $\ell_p$ norms $f(x) = ||x||_p$, where $p \geq 1$, are Schur-convex. The Shannon entropy $f(x) = -\sum_j x_j \log x_j$ is Schur-concave. In [73], it was shown that for fixed nonnegative numbers $p_1 \geq \cdots \geq p_n$ such that $p_1 + \cdots + p_n = 1$, the function $f(x) = \sum_j \sqrt{p_j x_j^j}$ is Schur-concave. Here $x_j^\uparrow$ denote the $j^{th}$ smallest component of $x$. Similarly, the function $f(x) = -\sum_j p_j \log x_j^\downarrow$ can be shown to be Schur-convex.

One extends the definition of Schur-convexity/concavity to functions of the form $F : \mathcal{H}_n \to \mathbb{R}$ satisfying $F(\cdot) = f(eig(\cdot))$ for some function $f : \mathbb{R}^n \to \mathbb{R}$. $F$ is said to be Schur-convex (respectively, Schur-concave) if $f$ is. For example, any unitary similarity invariant norm is Schur-convex [76] while the von Neumann Entropy $H(\cdot)$ is Schur-concave [92].
CHAPTER 2

Decomposition of Quantum Gates*

2.1 Introduction

The foundation of quantum computation [79] involves the encoding of computational tasks into the temporal evolution of a quantum system. A register of qubits, identical two-state quantum systems, is employed, and quantum algorithms can be described by unitary transformations and projective measurements acting on the state vector of the register. In this context, unitary matrices are called quantum gates. Mathematically, a two-state quantum system has vector states $|\psi\rangle$ in $\mathbb{C}^2$, known as qubits. The two vectors in the standard basis $\{|0\rangle, |1\rangle\}$ for $\mathbb{C}^2$ correspond to two physically measurable quantum states. An $n$-qubit system containing registers of $n$-qubits has vector states in the Euclidean space $\mathbb{C}^2 \otimes \cdots \otimes \mathbb{C}^2 = (\mathbb{C}^2)^\otimes n$ with basis vectors

$$
|j_n \cdots j_1\rangle = |j_n\rangle \otimes \cdots \otimes |j_1\rangle, \quad j_1, \ldots, j_n \in \{0, 1\}
$$

*The material in this chapter is contained in the paper [62], which is a joint work of C.K. Li and the author.
corresponding to the $2^n$ physically measurable states.

For a single qubit, one can use quantum gates corresponding to unitary transformations to manipulate the qubit. For an $n$-qubit system with large $n$, it is challenging and expensive to implement quantum gates. One often has to decompose a general quantum gate into the product of simple/elementary unitary gates which can be readily created physically. For a discussion on decomposing a unitary matrix into sets of elementary quantum gates, see, for example, [24], [27], [44], [87], and their references. By elementary linear algebra, it is known that every $N \times N$ unitary matrix can be written as the product of no more than $N(N-1)/2$ 2-level unitary matrices (Given’s transforms), i.e., unitary matrices obtained from the identity matrix by changing a $2 \times 2$ principal submatrix.

For example, if $U \in \mathcal{U}_4$, then there are unitary matrices of the form

$$U_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \star & \star \\ 0 & 0 & \star & \star \end{pmatrix}, \quad U_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \star & \star & 0 \\ 0 & \star & \star & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad U_3 = \begin{pmatrix} \star & \star & 0 & 0 \\ \star & \star & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

so that $U_1U$ has a zero $(4,1)$ entry, $U_2U_1U$ has zero entries at the $(4,1)$ and $(3,1)$ positions, and $U_3U_2U_1U$ has zero entries at the $(4,1), (3,1), (2,1)$ positions, and $(1,1)$ entry equal to one. Because $U_3U_2U_1U$ is unitary, it will be of the form $[1] \oplus \tilde{U}$ with $\tilde{U} \in \mathcal{U}_3$. We can then find unitary matrices of the form

$$U_4 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \star & \star \\ 0 & 0 & \star & \star \end{pmatrix}, \quad U_5 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \star & \star & 0 \\ 0 & \star & \star & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad U_6 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \star & \star \\ 0 & 0 & \star & \star \end{pmatrix}$$

so that $U_5U_4U_3U_2U_1U$ has the form $I_2 \oplus V$ with $V \in \mathcal{U}_2$ and $U_6 \ldots, U_1U = I_4$. It follows that $U = U_1^* \cdots U_6^*$.

In the context of quantum information science, not all 2-level unitary matrices are easy to implement. In this context, one considers matrices of sizes $N = 2^n$ labeled by binary
sequences $j_n \cdots j_1 \in \{0,1\}^n$ corresponding to the measurable quantum state $|j_n \cdots j_1\rangle$.

Then certain two level unitary matrices correspond to quantum operations acting on the $s^{th}$ qubit provided the other qubits $|j_n\rangle, \ldots, |j_{s+1}\rangle, |j_{s-1}\rangle, \ldots, |j_1\rangle$ assume specified values in $\{|0\rangle, |1\rangle\}$. These are known as the fully controlled qubit gates. For example, when $n = 2$, we label the rows and columns of matrices by $00, 01, 10, 11$. There are four types of fully-controlled 2-qubit gates:

$$(0V): \begin{pmatrix} v_{11} & v_{12} & 0 & 0 \\ v_{21} & v_{22} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad (1V): \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & v_{11} & v_{12} \\ 0 & 0 & v_{21} & v_{22} \end{pmatrix}$$

$$(V0): \begin{pmatrix} v_{11} & 0 & v_{12} & 0 \\ 0 & 1 & 0 & 0 \\ v_{21} & 0 & v_{22} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad (V1): \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & v_{11} & 0 & v_{12} \\ 0 & 0 & 1 & 0 \\ 0 & v_{21} & 0 & v_{22} \end{pmatrix}$$

with $V = \begin{pmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{pmatrix} \in \mathcal{U}_2$. In particular, a $(0V)$-gate corresponds to the unitary operator

$$a|00\rangle + b|01\rangle + c|10\rangle + d|11\rangle \mapsto |0\rangle V(a|0\rangle + b|1\rangle) + |1\rangle (c|0\rangle + d|1\rangle),$$

which will only change the part of the vector state with the first qubit equal to $|0\rangle$.

Similarly, a $(1V)$-gate corresponds to the unitary operator

$$a|00\rangle + b|01\rangle + c|10\rangle + d|11\rangle \mapsto |0\rangle (a|0\rangle + b|1\rangle) + |1\rangle V(c|0\rangle + d|1\rangle),$$

which will only change the part of the vector state with the first qubit equal to $|1\rangle$. The $(V0)$-gate and $(V1)$-gate have the same physical interpretation. One can associate the 4 types of controlled qubit gates with the circuit diagrams in Figure 2.1.
FIG. 2.1: Circuit diagrams for controlled 2-qubit gates.

For $n = 3$, we have fully-controlled qubit gates of the types:

$$(00V), (01V), (10V), (11V), (0V0), (0V1), (1V0), (1V1), (V00), (V01), (V10), (V11).$$

One easily extends this idea and notation to define fully-controlled gates acting on $n$-qubits.

In [90] (see also [72]), it was shown that one can decompose a quantum gate into the product of 2-level matrices corresponding to fully-controlled qubit gates. While fully-controlled qubit gates are relatively simple, it is still not easy to implement because the qubit gate $V$ can only act on the target bit after verifying that the other $(n - 1)$-qubits satisfy the controlled bits. As mentioned in [90], in practice it is desirable to replace fully controlled qubit gates by qubit gates with as few controls as possible. For example, when $n = 2$, the following types of unitary gates with no controls

$$(*)V$$

$$V = \begin{pmatrix} v_{11} & 0 & 0 & 0 \\ 0 & v_{12} & 0 & 0 \\ 0 & 0 & v_{11} & v_{12} \\ 0 & 0 & 0 & v_{22} \end{pmatrix}, \quad V* = \begin{pmatrix} v_{11} & 0 & 0 & 0 \\ 0 & v_{12} & 0 & 0 \\ 0 & 0 & v_{11} & v_{12} \\ 0 & 0 & 0 & v_{22} \end{pmatrix}$$

are easier to implement. Note that a $(0V)$-gate is applied on the left of a matrix $A \in \mathbb{C}^{4 \times 4}$, only rows 00 and 01 are affected. Similarly, a $(1V)$-gate will only affect the 10 and 11 gate of $A$. However, a $(*)$-gate and $(V*)$-gate will affect all rows of $A$.

In general, we can consider a $(c_n c_{n-1} \cdots c_1)$-unitary gate with $c_n, \ldots, c_1 \in \{0, 1, *, V\}$, where only one of the terms is $V$, and the number of terms in $\{0, 1\}$ is the total number of controls. For example, a $(11 * 0V1)$-unitary gate acting on 6-qubit states has 4 controls, and the target qubit is the fifth one. Our goal is to address the following problem.
Problem 2.1.1. Given $U \in \mathcal{U}_{2^n}$, write $U = U_1 \cdots U_N$ such that $\sum_{j=1}^{N} \# \text{control}(U_j)$ is as small as possible.

In [90], a recurrence scheme was proposed to decompose a unitary gate as the product of controlled qubit gates with small number of controls. The purpose of this chapter is to present another simple recurrence scheme, which provide an alternative choice for implementation. Moreover, the ideas and techniques in the construction may be helpful for further research in this and related problems.

This chapter is organized as follows. In Section 2.2, we will illustrate our scheme for the 2-qubit and 3-qubit case, and discuss how it can be extended. In Section 2.3, we present the general scheme with detailed description of the implementation steps and explanation of their validity. In Section 2.4, we obtain formulas for the number of $k$-controlled single qubit gates in the decomposition and compare our results to those in scheme in [90]. Concluding remarks and future research directions are mentioned in Section 2.5.

2.2 Two-qubit and Three-qubit cases

For an $n$-qubit unitary gate $U \in \mathcal{U}_N$ with $N = 2^n$, we will describe a recurrence scheme for generating controlled single qubit unitary gates $U_1, \ldots, U_r$ with $r \leq N(N - 1)/2$ such that $U_r \cdots U_1 U = I_N$. Consequently, $U = U_1^\dagger \cdots U_r^\dagger$.

Our scheme is done as follows. Assume we have the reduction scheme for the $(n - 1)$-qubit case.

Step 1 Partition $U \in \mathcal{U}_N$ into a $2 \times 2$ block matrix with each block lying in $\mathbb{C}^{N/2 \times N/2}$.

Step 2 Use the scheme of the $(n - 1)$-qubit case to help reduce $U$ to the form $I_{N/2} \oplus \tilde{U}$ with $\tilde{U} \in \mathcal{U}_{N/2}$. 
Step 3 Apply the scheme for the \((n - 1)\)-qubit case, with some modification, to transform \(\tilde{U}\) to \(I_{N/2}\).

In Step 2, we need to eliminate the nonzero off-diagonal entries of \(U\) for the first \(N/2\) columns. We will do these elimination column by column starting from column 1, then moving to column 2 and so on, making sure that the entries annihilated by previous steps will remain zero. For column \(1 \leq j \leq N/2\), we first eliminate the off-diagonal entries \((j + 1, j), \ldots, (N/2 + 1, j)\) using the scheme in the \((n - 1)\)-qubit case. Then eliminate entries \((N/2 + 1, j), \ldots (N, j)\) using a recurrent scheme based on the annihilation of entries \((N/2 + 1, 1), \ldots (N, 1)\) of column 1. It is therefore important to clearly explain the scheme to annihilate the lower half of the first column.

First, we will specify the scheme for two-qubit gates and three-qubit gates.

The two-qubit gate.

In the following tables, we indicate the order of the entries to be eliminated in our scheme, and also the \((c_2c_1)\)-gates used to do the elimination.

<table>
<thead>
<tr>
<th>Column 1</th>
<th>entries</th>
<th>(2,1)</th>
<th>(4,1)</th>
<th>(3,1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>gates</td>
<td>(*V)</td>
<td>(1V)</td>
<td>(V*)</td>
</tr>
<tr>
<td>Column 2</td>
<td>entries</td>
<td>(3,2)</td>
<td>(4,2)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>gates</td>
<td>(1V)</td>
<td>(V1)</td>
<td></td>
</tr>
<tr>
<td>Column 3</td>
<td>entries</td>
<td>(4,3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>gates</td>
<td>(1V)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TABLE 2.1: Scheme table for decomposing 2–qubit quantum gates

Here we first eliminate the \((2,1)\) entry as in the 1-qubit case. In a similar manner, annihilate the \((4,1)\) entry, treating it as the second entry of the lower left half of the first column. To keep the \((2,1)\) entry zero, we use a gate with a 1 – control in the leftmost bit.
Finally we annihilate the $(3,1)$ entry with the help of the $(1,1)$ entry. In this case, we can use a control-free gate to do so. At this point, the current form of the matrix is $[1] \oplus U'$, where $U' \in \mathcal{U}_3$.

Then we move to the second column. We adapt the procedure of eliminating the $(4,1)$ and $(3,1)$ entries to eliminate the $(3,2)$ and $(4,2)$ entries. The gates used must not change the zero entries in the first column. After this, the matrix takes the form $I_2 \oplus U_1$ with $U_1 \in \mathcal{U}_2$. We can deal with the matrix $U_1$ as in the 1-qubit case using a $(1V)$-gate so that the first two rows will not be affected.

**The three qubit case.**

We execute the reduction scheme for three qubit gates as described in the following table, implementing the indicated controlled gates from left to right and then from the first column to the next.

<table>
<thead>
<tr>
<th>Column 1</th>
<th>entries</th>
<th>$(2,1)$</th>
<th>$(4,1)$</th>
<th>$(3,1)$</th>
<th>$(6,1)$</th>
<th>$(8,1)$</th>
<th>$(7,1)$</th>
<th>$(5,1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>gates</td>
<td>$(**V)$</td>
<td>$(*1V)$</td>
<td>$(<em>V</em>)$</td>
<td>$(1*V)$</td>
<td>$(1V*)$</td>
<td>$(V**)$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Column 2</th>
<th>entries</th>
<th>$(3,2)$</th>
<th>$(4,2)$</th>
<th>$(5,2)$</th>
<th>$(7,2)$</th>
<th>$(8,2)$</th>
<th>$(6,2)$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>gates</td>
<td>$(*)V$</td>
<td>$(*V1)$</td>
<td>$(1*V)$</td>
<td>$(1V*)$</td>
<td>$(V*1)$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Column 3</th>
<th>entries</th>
<th>$(4,3)$</th>
<th>$(8,3)$</th>
<th>$(6,3)$</th>
<th>$(5,3)$</th>
<th>$(7,3)$</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>gates</td>
<td>$(*)V$</td>
<td>$(1*V)$</td>
<td>$(10V)$</td>
<td>$(1V*)$</td>
<td>$(V1*)$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Column 4</th>
<th>entries</th>
<th>$(7,4)$</th>
<th>$(5,4)$</th>
<th>$(6,4)$</th>
<th>$(8,4)$</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>gates</td>
<td>$(1*V)$</td>
<td>$(10V)$</td>
<td>$(1V*)$</td>
<td>$(V11)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Column 5</th>
<th>entries</th>
<th>$(6,5)$</th>
<th>$(8,5)$</th>
<th>$(7,5)$</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>gates</td>
<td>$(1*V)$</td>
<td>$(11V)$</td>
<td>$(1V*)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Column 6</th>
<th>entries</th>
<th>$(7,6)$</th>
<th>$(8,6)$</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>gates</td>
<td>$(11V)$</td>
<td>$(1V1)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Column 7</th>
<th>entries</th>
<th>$(8,7)$</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>gates</td>
<td>$(11V)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**TABLE 2.2: Scheme table for decomposing 3-qubit quantum gates**

In this case, we have 3 types of unitary gates with no control, 12 types of unitary gates
with 1 control (0 or 1) and 1 target qubit and 12 types of unitary gates with 2 controls and 1 target qubit.

**Remarks 2.2.1.** Here we give some remarks about the reduction of a 3-qubit unitary gate to help illustrate our recurrence scheme and how it can be extended. The comments are numbered according to the major steps 1–3 of our scheme described in the beginning of this section.

(S1) We partition the $8 \times 8$ unitary matrix into a 2-by-2 block matrix so that each block is $4 \times 4$.

(S2) We consider Column 1, 2, 3, 4,

**For Column 1**, the elimination of $(2, 1), (4, 1), (3, 1)$ entries will be done as in the $4 \times 4$ (2-qubit) case by changing the 2-qubit $(c_2c_1)$-gates to $(\ast c_2c_1)$-gates in these steps.

We then annihilate the $(6, 1), (8, 1)$ and $(7, 1)$ entries the same way we annihilated the $(2, 1), (4, 1)$ and $(3, 1)$ entries by treating the lower half as a $4 \times 4$ matrix. However, we have to ensure that the $(1, 1)$ entry will not interact with the zero entries at the $(2, 1), (3, 1), (4, 1)$ positions in these steps. So, we adapt the 2-qubit $(c_2c_1)$-gates to $(c_3c_2c_1)$-gates, we will use the following rule:

$$
\text{let } c_3 = 1 \text{ if } (c_2c_1) \text{ is } (\ast V) \text{ or } (V\ast); \text{ otherwise, let } c_3 = \ast.
$$

So, a $(1 \ast V)$-gate can be used to annihilate the $(6, 1)$ entry, a $(\ast 1V)$-gate can be used to annihilate the $(8, 1)$ entry and a $(1V\ast)$-gate to annihilate the $(7, 1)$ entry. Finally, we can apply a $(V\ast\ast)$-gate to eliminate the $(5, 1)$ entry using the $(1, 1)$ entry.

Note that the $(c_3c_2c_1)$-gates used in the Column 1 satisfy $c_3, c_2, c_1 \in \{\ast, 1, V\}$ with $c_1 \neq 1$. This property will hold for the general case.
Once all off-diagonal entries in Column 1 are annihilated, we obtain a matrix of the form $[1] \oplus U'$, where $U' \in M_7$. We can proceed to Column 2.

For Column 2, we can annihilate the $(3, 2)$ and $(4, 2)$ entries using the scheme for annihilating the second column in the $4 \times 4$ case by changing the 2-qubit $(c_2c_1)$-gates to $(c_2c_1)$-gates in these steps.

Next, we adapt the scheme of annihilating the $(6, 1), (8, 1), (7, 1), (5, 1)$ entries to annihilate the lower half entries of the second column. Note that it is imperative that the $(6, 2)$ entry be the last entry to be annihilated since it is the only entry in the lower half of the column that can be annihilated using the $(2, 2)$ entry. In view of this, we will change the order of annihilation of the entries to:

$$(5, 2), (7, 2), (8, 2), (6, 2).$$

If we identify $(1, 2, \ldots, 8)$ with the binary sequences $(000, 001, \ldots, 111)$, then

$$(6, 8, 7, 5) \text{ corresponds to } (101, 111, 110, 100), \text{ and } (5, 7, 8, 6) \text{ corresponds to } (100, 110, 111, 101).$$

The conversion can be easily realized by

$$(100, 110, 111, 101) = (101, 111, 110, 100) \oplus (001, 001, 001, 001) = (101 \oplus 001, 111 \oplus 001, 110 \oplus 001, 100 \oplus 001),$$

where $i_3i_2i_1 \oplus j_3j_2j_1$ is an entry-wise addition such that $0 \oplus 0 = 1 \oplus 1 = 0$ and $0 \oplus 1 = 1 \oplus 0 = 1$. Note that we will use a similar conversion for columns 3 and 4.

We also need to modify the $(c_3c_2c_1)$-gates used to annihilate the $(6, 1), (8, 1), (7, 1)$ entries to annihilate the $(5, 2), (7, 2), (8, 2)$ entries. To accommodate the change in the order
of annihilation, one must modify any control found in $c_1$. We also have to prevent the $(1,1)$ entries interacting with the $(2,1), (3,1), (4,1)$ entries, and also prevent the $(2,2)$ entries interacting with the $(3,2)$ and $(4,2)$ entries. This can be done by making sure that at least one of $c_2$ and $c_3$ is equal to 1. Thus, we modify $(c_3c_2c_1)$ by the following rules:

$$
	ext{change } c_3 \text{ to } 1 \text{ if none of } c_2, c_3 \text{ is } 1; \text{ change } c_1 \text{ to } 0 \text{ if } c_1 = 1.
$$

However, one sees that applying these rules will not change the $(c_3c_2c_1)$-gates in view of the fact that $c_1 \neq 1$. Hence we can use exactly the same set of $(c_3c_2c_1)$-gates to eliminate the $(5,2), (7,2), (8,2)$ entries of Column 2.\footnote{As we will see, the same phenomenon will hold for columns 3 and 4, and also for the general case.} Thus, we will use $(1^*V), (*1V), (1V*)$ gates to annihilate the $(5,2), (7,2)$ and $(8,2)$ entries, respectively.

To annihilate the $(6,2)$ entry, we need to utilize the nonzero $(2,2)$ entry. These two entries correspond to rows 101 and 001. This means that the target bit of the gate we need is the third bit (leftmost). Because we do not want to change the form of the upper half of the first column, we need to make sure that the gate is not satisfied by 000 but is satisfied by 001 and 101. Thus, we use a $(V * 1)$-gate. Once this is done, the matrix is now reduced to the form $I_2 \oplus V''$ where $V'' \in U_6$.

For Column 3, the $(4,3)$ entry is annihilated using the scheme for the third column of the $4 \times 4$ case.

Similar to the case in Column 2, we can adapt the scheme of eliminating the $(6,1), (8,1), (7,1), (5,1)$ entries to annihilate the $(8,3), (6,3), (5,3), (7,3)$ entries. The conversion $(6,8,7,5)$ to $(8,6,5,7)$ is done by performing

$$(111,101,100,110) = (101,111,110,100) \oplus (010,010,010,010)$$

$\dagger$
using the binary number correspondence of the indices.

We also need to modify the \((c_3c_2c_1)\)-gates used to annihilate the \((6,1),(8,1),(7,1)\) entries to annihilate the \((8,3),(6,3),(7,3)\) entries. In these steps, we have to prevent the \((1,1)\) entries interacting with the \((2,1),(3,1),(4,1)\) entries, the \((2,2)\) entries interacting with the \((3,2),(4,2)\) entries, and the \((3,3)\) entry interacting with the \((4,3)\) entry. One can do this by adjusting the \(c_3\) and \(c_2\) values in the \((c_3c_2c_1)\)-gates used for the annihilation of the \((6,1),(8,1),(7,1),(5,1)\) entries by the following rules:

\[
\text{change } c_3 \text{ to } 1 \text{ if } c_3 \text{ is not } 1; \text{ change } c_2 \text{ to } 0 \text{ if } c_2 = 1.
\]

Since \(c_3\) is 1, for \(i = 1, 2, 3, 4\), the \((i,i)\) entry will not interact with other \((k,i)\) entries for \(1 \leq k \leq 4\) and \(k \neq i\). Note that a \((c_3c_2c_1)\)-gate corresponds to a unitary matrix \(\tilde{V} \in M_8\).

Changing a control bit in the position of \(c_2\) corresponds to changing \(\tilde{V}\) by a permutation similarity \(P^\dagger \tilde{V} P\), where \(P\) corresponds to the change of the basis \(\{|000\}, \ldots, |111\}\) to \(\{|010\}, \ldots, |101\}\), here we change \(|j_2j_2j_1\rangle\) to \(|j_3(j_2 \oplus 1)j_1\rangle\). Thus, the modified \((c_3c_2c_1)\)-gates can be used for Column 3. We will give a general description of this procedure in the next section. Here, we obtain the \((1*V),(10V),(1V*)\) gates, which can be used to annihilate the \((8,3),(6,3),(5,3)\) entries.

Finally, to annihilate the \((7,3)\) entry, we use the \((3,3)\) entry. Hence, the target bit of the gate we need is the leftmost bit. To avoid changing the form of the first and second columns, we need to use controls that are not satisfied by 000 and 001 but is satisfied by 010 and 110. Thus, we use the gate \((V1*)\).

**For Column 4**, we need not do anything about the first four entries at this point.

We will adapt the scheme for the \((6,1),(8,1),(7,1),(5,1)\) entries to annihilate the \((7,4),(5,4),(6,4),(8,4)\) entries. The conversion \((6,8,7,5)\) to \((7,5,6,8)\) is done by per-
forming

\[(110, 100, 101, 111) = (101, 111, 110, 100) \oplus (011, 011, 011, 011)\]

using the binary number correspondence of the numbers.

We adjust the \((c_3c_2c_1)\)-gates used for the \((6, 1), (8, 1), (7, 1)\) entries to annihilate the \((7, 4), (5, 4), (6, 4)\) entries as follows,

\[
\text{change } c_3 \text{ to 1 if } c_3 \text{ is not 1; for } j = 1, 2, \text{ change } c_j \text{ to 0 if } c_j = 1. 
\]

Note that column 4 is associated to the binary sequence 011.\(^1\) We will obtain the \((1 \star V), (10V), (1V \star)\) gates, which can be used to annihilate the \((7, 4), (5, 4), (6, 4)\) entries.\(^2\) Finally use a \((V11)\)-gate to annihilate the \((8, 4)\) entry using the \((4, 4)\) entry while avoiding any change in the form of the first three columns.

(S3) Note that after Column 4 is dealt with, the matrix takes the form \(I_4 \oplus V'\) where \(V' \in M_4\).

We can then use the scheme for the 2-qubit case to transform \(V'\) to \(I_4\). However, to avoid changing the form of the first four columns, we need to extend the \((c_2c_1)\)-gates used in the \(4 \times 4\) case to \((1c_2c_1)\)-gates for the remaining steps. This explains the tables for columns 5 to 7.

2.3 General Scheme

In this section, we present the general recurrence scheme for the annihilation of the off-diagonal entries of an \(n\)-qubit unitary gate by adapting the reduction scheme of the \((n - 1)\)-qubit case. We will carry out Steps 1 – 3 described at the beginning of Section

\(^1\)As we will see in the next section, we always adjust the gates according to the the binary sequence associated to the column index.

\(^2\)Note also that the \((c_3c_2c_1)\)-gates are the same as those used in Column 3 before the final step. We will also explain this in the next section.
2. As illustrated in the 3-qubit case and explained in Remark 2.1, Step 2 of the scheme requires some careful attention. For each column $\ell = 1, \ldots, N/2$ with $N = 2^n$, we can always annihilate the off-diagonal entries in the upper half of column $\ell$ using the scheme for annihilating the first column for an $(n-1)$-qubit unitary gate. One only needs to change a $(c_{n-1} \cdots c_2 c_1)$-gate to a $(\ast c_{n-1} \cdots c_1)$-gate.

For the lower half of column $\ell$, we have to refine Step 2 to the following steps.

Step 2.1 For column 1, use the reduction scheme for an $(n-1)$-qubit to eliminate the off-diagonal entries in the upper half of the column by changing the $(c_{n-1} \cdots c_1)$-gates used in the $(n-1)$-qubit gate case to $(\ast, c_{n-1}, \ldots, c_n)$-gates in these steps.

Next, we apply the same scheme to eliminate the entries in the lower half except for the $(n/2 + 1, 1)$ entry, which will be eliminated last. This is done by changing the $(c_{n-1} \cdots c_1)$-gates in the $(n-1)$-qubit case to $(c_n \cdots c_1)$-gates, where

$$c_n = \begin{cases} 
1 & \text{none of } c_{n-1}, \ldots, c_1 \text{ equals } 1, \\
\ast & \text{otherwise.}
\end{cases} \quad (2.2)$$

The $(c_n \cdots c_1)$-gate constructed in this way will ensure that the $(1, 1)$ entry will not interact with $(2, 1), \ldots, (N/2, 1)$ entries when we annihilate the $(N/2 + j, 1)$ entry for $j = 2, \ldots, N/2$ because $1 \in \{c_n, \ldots, c_1\}$. Finally, apply a $(V \ast \cdots \ast)$-gate to annihilate the $(N/2 + 1, 1)$ entry.

An easy inductive argument will verify that the $(c_n \cdots c_1)$-gates used in Column 1 satisfy $c_n, \ldots, c_1 \in \{\ast, 1, V\}$ with $c_1 \neq 1$.

The annihilation steps of Column 1 can be summarized in the following.
Procedure 2.1

Suppose in the \((n-1)\)-qubit case, the off-diagonal entries in the first column are eliminated in the order of

\[(b_1,1), \ldots, (b_{N/2-1},1) \text{ by } C_1 - \text{gate}, \ldots, C_{N/2-1} - \text{gate}.
\]

Eliminate the entries in the upper half of the Column 1 in the order of

\[(b_1,1), \ldots, (b_{N/2},1) \text{ by } (\ast C_1) - \text{gate}, \ldots, (\ast C_{N/2-1}) - \text{gate}.
\]

For \(C = (c_{n-1} \cdots c_1)\) let \(G(C) = (c_n c_{n-1} \cdots c_1)\) with \(c_n\) satisfying (2.2).

Eliminate the entries in the lower half of the column in the order of

\[(d_1,1), \ldots, (d_{N/2-1},1) \text{ by } G(C_1) - \text{gate}, \ldots, G(C_{N/2-1}) - \text{gate},
\]

where \(d_i = b_i + N/2\) for \(i = 1, \ldots, N/2 - 1\), and eliminate the \((N/2 + 1,1)\) entry by a \((V\ast \cdots \ast) - \text{gate}.
\]

Step 2.2 For column \(\ell\) with \(2 \leq \ell \leq N/2\), we can use the same scheme as that of the \((n-1)\)-qubit case to eliminate the off-diagonal entries in the upper half. Then we can adapt the scheme for eliminating the entries in the lower half of Column 1 to other columns. To this end, we need to modify

(a) the order of the elimination of the entries in the lower half so that the last entry in the lower half will be eliminated by the \((\ell, \ell)\) entry.

(b) the control gates used to do the elimination so that

(b.i) they will not affect the zero entries obtained in the previous steps; and

(b.ii) they will annihilate the entries in the order prescribed in (a).

To achieve (a) and (b), identify \(k \in \{1, \ldots, 2^n\}\) with the binary sequence \(\tilde{k}_n \cdots \tilde{k}_1 \in\)
\{0 \cdots 0, \ldots, 1 \cdots 1\} \) so that
\[
k = \sum_{j=1}^{n} \tilde{k}_j 2^{j-1} + 1.
\]

For (a), if we annihilate the entries in the lower half of Column 1 in the order of \((d_1, 1), \ldots, (d_{N/2}, 1)\), then we will annihilate the entries in the lower half of column \(\ell\) in the order of \((d_1 \oplus \ell, \ell), \ldots, (d_{N/2} \oplus \ell, \ell)\), where the binary sequence of \(d_j \oplus \ell\) is obtained by entry-wise addition \(\oplus\) (without carried digits) of the two binary sequences of \(d_j\) and \(\ell\) such that \(0 \oplus 0 = 1 \oplus 1 = 0\) and \(0 \oplus 1 = 1 \oplus 0 = 1\). Note that \(d_{N/2} = N/2 + 1\), and hence \(d_{N/2} \oplus \ell = N/2 + \ell\), so that \((N/2 + \ell, \ell)\) is the last entry in the lower half of Column \(\ell\) to be eliminated.

For (b), suppose \(2^{m-1} < \ell \leq 2^m\) with \(m \in \{1, \ldots, n - 1\}\) and \(\ell = \sum_{j=1}^{n} \tilde{\ell}_j 2^{j-1} + 1\). We adjust the \((c_n \cdots c_1)\)-gate used to annihilate the \((d_i, 1)\) entry with \(N/2 + 1 \leq d_i < N\) to the \((\tilde{c}_n \cdots \tilde{c}_1)\)-gates for annihilating the \((d_i \oplus \ell, \ell)\) entry as follows, where

\[
\tilde{c}_j = \begin{cases} 
1 & \text{if } j = n \text{ and none of } c_n, \ldots, c_{m+1} \text{ is 1,} \quad \text{(taking care of (b.i))} \\
0 & \text{if } 1 \leq j \leq m \text{ and } c_j = \tilde{\ell}_j = 1, \quad \text{(taking care of (b.ii))} \\
c_j & \text{otherwise.}
\end{cases} \tag{2.3}
\]

Because at least one of \(\tilde{c}_n, \ldots, \tilde{c}_{m+1}\) is 1, for \(1 \leq j \leq 2^m\) the \((j, j)\) entries will not interact with other \((k, j)\) entry with \(1 \leq k \leq N/2\) and \(k \neq j\).

Note also that a \((c_n \cdots c_1)\)-gate with \(c_n, \ldots, c_1 \in \{*, 0, 1, V\}\) corresponding to the unitary matrix

\[
\tilde{V} = I_N + V_n \otimes \cdots \otimes V_1,
\]

\footnote{For instance, the binary form of \(f_2(d_i)\) is the sum of (using \(\oplus\)) the binary sequence \((0 \cdots 01)\) and the binary form of \(d_i\); the binary form of \(f_3(d_i)\) is the sum of the binary sequence \((0 \cdots 010)\) and the binary form of \(d_i\); \ldots, and the binary form of \(f_{N/2}(d_i)\) is the sum of the binary sequence \((01 \cdots 1)\) and the binary form of \(d_i\).}
where

\[ V_i = \begin{cases} |0\rangle\langle 0| & \text{if } c_i = 0, \\ |1\rangle\langle 1| & \text{if } c_i = 1, \\ V - I_2 & \text{if } c_i = V, \\ I_2 & \text{if } c_i = *. \end{cases} \]

For the \((c_n \cdots c_1)\)-gates used in the first columns, we have \(c_n, \ldots, c_1 \in \{*, 1, V\}\) with \(c_1 \neq 1\). So, changing the 1-control in the \(c_i\) position whenever \(\tilde{\ell}_i = 1\) in our rule is equivalent to applying a unitary similarity transform to change \(V\) to \(P_\ell^t V P_\ell\), where \(P_\ell\) is the permutation matrix changing the basis \(\{|j_n \cdots j_1\} : j_r \in \{0, 1\}\}\) to \(\{|j_n \cdots j_1 \oplus \tilde{\ell}_n \cdots \tilde{\ell}_1\} : j_r \in \{0, 1\}\}\), where \(\tilde{\ell}_n \cdots \tilde{\ell}_1\) is the binary number corresponding to \(\ell\).

So, the modified gates can be used to annihilate \((d_j \oplus \ell, \ell)\) entries for \(j = 1, \ldots, N/2 - 1\). After that, only the \((\ell, \ell)\) and \((N/2 + \ell, \ell)\) entries are nonzero in column \(\ell\). We annihilate the \((N/2 + \ell, \ell)\) entry using the \((V \hat{c}_{n-1} \cdots \hat{c}_1)\)-gate to ensure that the annihilation in these steps will not affect the zero entries in the previous steps, where \((\hat{c}_{n-1} \cdots \hat{c}_1)\) is obtained from the binary sequence correspondence \((\tilde{\ell}_{n-1} \cdots \tilde{\ell}_1)\) of \(\ell\) by changing all 0 terms to \(*.\)

Note also that except for the last step one will always get the same set of \((c_n \cdots c_1)\)-gates for the the elimination of the lower half of the entries in Columns \(2k - 1\) and \(2k\) because the modification in (2.3) will have the same effects in these columns. This follows from the fact that the \((c_n \cdots c_1)\)-gates for Column 1 satisfy \(c_n, \ldots, c_1 \in \{*, 1, V\}\) with \(c_1 \neq 1\).

The annihilation steps of Column \(\ell\) can be summarized in the following.

\[\text{For example, for Column 2 we change } (c_n \cdots c_1) \text{ to } G_2(c_n \cdots c_1) \text{ by changing only } c_1 \text{ and } c_n \text{ because 2 corresponds to 0\cdots01, and } (\hat{c}_{n-1} \cdots \hat{c}_1) = (V \cdots \cdots 1)\; ; \text{ for Column 3, we change } (c_n \cdots c_1) \text{ to } G_3(c_n \cdots c_1) \text{ by changing only } c_2 \text{ and } c_n \text{ because 3 corresponds to 0\cdots010, and } (\hat{c}_{n-1} \cdots \hat{c}_1) = (V \cdots \cdots *1)\; ; \text{ for Column 4, we change } (c_n \cdots c_1) \text{ to } G_4(c_n \cdots c_1) \text{ by changing only } c_1, c_2 \text{ and } c_n \text{ because 4 corresponds to 0\cdots011, and } (\hat{c}_{n-1} \cdots \hat{c}_1) = (V \cdots \cdots 11).\]
Suppose in the \((n-1)\)-qubit case, the off-diagonal entries in Column \(\ell\) are eliminated in the order of
\[
(a_1, \ell), \ldots, (a_{N/2-\ell}, \ell) \text{ by } \mathbf{D}_1 - \text{gate}, \ldots \mathbf{D}_{N/2-\ell} - \text{gate}.
\]
For the \(n\)-qubit case, eliminate the entries in the upper half of the column in the order of
\[
(a_1, \ell), \ldots, (a_{N/2-1}, \ell) \text{ by } (*\mathbf{D}_1) - \text{gate}, \ldots, (*\mathbf{D}_{N/2-\ell}) - \text{gate}.
\]
For \(\mathbf{C} = (c_{n-1} \cdots c_1)\) let \(G_\ell(\mathbf{C}) = (\tilde{c}_n \cdots \tilde{c}_1)\) satisfy (2.3), and let \(d_i\) and \(G(\mathbf{C}_i)\) be defined as in Procedure 2.1. Eliminate the entries in the lower half of the column in the order of
\[
(d_1 \oplus \ell, \ell), \ldots, (d_{N/2-1\oplus\ell}, \ell) \text{ by } G_\ell(G(\mathbf{C}_1)) - \text{gate}, \ldots G_\ell(G(\mathbf{C}_{N/2-1})) - \text{gate};
\]
eliminate the \((N/2 + \ell, \ell)\) entry by a \((\tilde{V}\tilde{c}_{n-1} \cdots \tilde{c}_1)\)-gate, where \((\tilde{c}_{n-1} \cdots \tilde{c}_1)\) is obtained from the binary sequence correspondence \((\tilde{\ell}_{n-1} \cdots \tilde{\ell}_1)\) of \(\ell\) by changing all 0 terms to *.

Several remarks concerning Procedures 2.1 and 2.2 are in order.

1. In Column 1, it is easy to determine the order of the entries to be eliminated and the \((c_n \cdots c_1)\)-gates used.

2. For the lower half of Column \(\ell\) with \(2 \leq \ell \leq N/2\), we change the order of entries to be eliminated to \((d_1 \oplus \ell, \ell), \ldots, (d_{N/2} \oplus \ell, \ell)\), and change the \((c_n \cdots c_1)\)-gates to \(G_\ell(c_n \cdots c_1)\)-gates.

3. The \((c_n \cdots c_1)\)-gates used in Column 1 satisfy \(c_n, \ldots, c_1 \in \{*, 1, V\}\) with \(c_1 \neq 1\).

4. The \((c_n \cdots c_1)\)-gates used to eliminate the entries in the lower half of Column \(2k - 1\) and \(2k\) are always the same before the last step, for \(k = 1, \ldots, N/4\).
5. The \((c_n \cdots c_1)\)-gates used in the last steps of Columns 1, \ldots, \(N/2\) satisfy \(c_n = V\), and 
\((c_{n-1} \cdots c_1)\) is obtained from the binary sequences \((0 \cdots 0), \ldots, (1 \cdots 1)\) of length \(n - 1\) 
by replacing 0 with \(*\).

The recurrence scheme is easy to implement. Even the most non-trivial steps of 
adapting the procedures of eliminating the entries in the lower half of the first column to 
other columns are quite straightforward. We illustrate this for the case \(n = 4\).

**Four qubit case, lower left block**

<table>
<thead>
<tr>
<th>Col 1</th>
<th>entries</th>
<th>(10,1)</th>
<th>(12,1)</th>
<th>(11,1)</th>
<th>(14,1)</th>
<th>(16,1)</th>
<th>(15,1)</th>
<th>(13,1)</th>
<th>(9,1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>steps</td>
<td>binary</td>
<td>1001</td>
<td>1011</td>
<td>1010</td>
<td>1101</td>
<td>1111</td>
<td>1110</td>
<td>1100</td>
<td>1000</td>
</tr>
<tr>
<td>8-15</td>
<td>gates</td>
<td>1**V</td>
<td>**1V</td>
<td>1<em>V</em></td>
<td><em>1</em>V</td>
<td>**1V</td>
<td><em>1V</em></td>
<td>1V**</td>
<td>V***</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Col 2</th>
<th>entries</th>
<th>(9,2)</th>
<th>(11,2)</th>
<th>(12,2)</th>
<th>(13,2)</th>
<th>(15,2)</th>
<th>(16,2)</th>
<th>(14,2)</th>
<th>(10,2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>steps</td>
<td>binary</td>
<td>1000</td>
<td>1010</td>
<td>1011</td>
<td>1100</td>
<td>1110</td>
<td>1111</td>
<td>1101</td>
<td>1001</td>
</tr>
<tr>
<td>7-14</td>
<td>gates</td>
<td>1**V</td>
<td>**1V</td>
<td>1<em>V</em></td>
<td><em>1</em>V</td>
<td>**1V</td>
<td><em>1V</em></td>
<td>1V**</td>
<td>V**1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Col 3</th>
<th>entries</th>
<th>(12,3)</th>
<th>(10,3)</th>
<th>(9,3)</th>
<th>(16,3)</th>
<th>(14,3)</th>
<th>(13,3)</th>
<th>(15,3)</th>
<th>(11,3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>steps</td>
<td>binary</td>
<td>1011</td>
<td>1001</td>
<td>1000</td>
<td>1111</td>
<td>1101</td>
<td>1100</td>
<td>1110</td>
<td>1010</td>
</tr>
<tr>
<td>6-13</td>
<td>gates</td>
<td>1**V</td>
<td>1*0V</td>
<td>1<em>V</em></td>
<td><em>1</em>V</td>
<td>1*0V</td>
<td><em>1V</em></td>
<td>1V**</td>
<td>V**1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Col 4</th>
<th>entries</th>
<th>(11,4)</th>
<th>(9,4)</th>
<th>(10,4)</th>
<th>(15,4)</th>
<th>(13,4)</th>
<th>(14,4)</th>
<th>(16,4)</th>
<th>(12,4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>steps</td>
<td>binary</td>
<td>1010</td>
<td>1000</td>
<td>1001</td>
<td>1110</td>
<td>1100</td>
<td>1101</td>
<td>1111</td>
<td>1011</td>
</tr>
<tr>
<td>5-12</td>
<td>gates</td>
<td>1**V</td>
<td>1*0V</td>
<td>1<em>V</em></td>
<td><em>1</em>V</td>
<td>1*0V</td>
<td><em>1V</em></td>
<td>1V**</td>
<td>V**1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Col 5</th>
<th>entries</th>
<th>(14,5)</th>
<th>(16,5)</th>
<th>(15,5)</th>
<th>(10,5)</th>
<th>(12,5)</th>
<th>(11,5)</th>
<th>(9,5)</th>
<th>(13,5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>steps</td>
<td>binary</td>
<td>1101</td>
<td>1111</td>
<td>1110</td>
<td>1001</td>
<td>1011</td>
<td>1010</td>
<td>1000</td>
<td>1100</td>
</tr>
<tr>
<td>4-11</td>
<td>gates</td>
<td>1**V</td>
<td>1*1V</td>
<td>1<em>V</em></td>
<td>10*V</td>
<td>1*1V</td>
<td>10V*</td>
<td>1V**</td>
<td>V1**</td>
</tr>
</tbody>
</table>


TABLE 2.3: Partial scheme table for annihilating the lower left block of a 4–qubit quantum gate

2.4 Total Number of Controls and Comparison to a Previous Study

Let $g_n^k$ denote the number of $k$-controlled qubit gates used in the decomposition scheme for $U \in U_{2^n}$. The following theorem gives the formula for the number $g_n^k$, where $k = 0, 1, \ldots, n - 1$

Theorem 2.4.1. 1. $g_n^0 = n$

2. $g_n^{n-1} = \begin{cases} 1 & \text{if } n = 1 \\ 4 & \text{if } n = 2 \\ 7 + (n-3) & \text{if } n \geq 3 \end{cases}$

3. $g_n^k = g_{n-1}^{k-1} + g_{n-1}^{k-1} + \binom{n-1}{k} \text{ for all } 3 \leq k < n - 1$

4. $g_n^1 = n(n-1)(2^{n-2} + 1) \text{ for all } n \geq 2$

5. $g_n^2 = \frac{1}{3}(4^n - 4) - 2^n(n-1) + \frac{n(n-1)(n-2)}{2} \text{ for all } n \geq 3$
Note that \( \sum_{k=0}^{n-1} g_n^k = 2^{n-1}(2^n - 1) = N(N - 1)/2 \). By convention \( g_0^1 = 1 \). In general, if \( n > 1 \),

\[ g_n^k = A_n^k + B_n^k + C_n^k + D_n^k, \]

where \( A_n^k \) is the number of \( g_n^k \) gates used to annihilate entries in the upper left block of the matrix, \( B_n^k \) is the number of \( g_n^k \) gates used to annihilate entries of the lower half of columns \( 1, \ldots, 2^{n-1} \) excluding the entries of the form \((N/2 + \ell, \ell)\). The number \( C_n^k \) is the number of \( g_n^k \) gates used to annihilate entries \((N/2 + \ell, \ell)\), where \( \ell \in \{1, \ldots, 2^{n-1}\} \). Finally \( D_n^k \) is the number of \( g_n^k \) gates used to annihilate the lower right block entries of the matrix. For example, we saw in section 2 that

\[ g_2^0 = 2 = 1 + 0 + 1 + 0 \quad \text{and} \quad g_2^1 = 4 = 0 + 2 + 1 + 1 \]

and

\[ g_3^0 = 3 = 2 + 0 + 1 + 0, \quad g_3^1 = 18 = 4 + 10 + 2 + 2, \quad \text{and} \quad g_3^2 = 0 + 2 + 1 + 4 \]

Remarks 2.4.2. Immediately, we can see the following recursive properties.

1. \( A_n^k = g_{n-1}^k \) for \( k \in \{0, \ldots, n - 2\} \) and \( A_n^{n-1} = 0 \) as illustrated in the first half of Procedure 2.2.

2. \( D_n^k = g_{n-1}^{k-1} \) for \( k \in \{1, \ldots, n - 1\} \) and \( D_n^0 = 0 \) since the \( k \)-controlled gates in Step 3 can be obtained by appending a 1-control in the leftmost qubit of a \((k1)\)-controlled gate that appears in the n1 scheme.

3. \( C_n^k = \binom{n-1}{k} \) for \( k \in \{0, \ldots, n - 1\} \), because \( C_n^m \) is the number of column indices \( \ell \), with \( 1 \leq \ell \leq 2^{n-1} \), such that the binary sequence of \( \ell \) of length \( n \) has exactly \( k \) digits equal to 1.

4. Observe that the gate \( G_j = \mathcal{G}(C_j) \), \( 1 \leq j \leq N/2 - 1 \), in table 1 has exactly one 1-control. All other gates accounted for by \( B_n^k \) are obtained from the \( G_j \)'s via the transformation \( G_\ell \), for \( 2 \leq \ell \leq N/2 \). But notice that \( G_\ell(G_j) \) either has the same number of controls as \( G_j \) or has one
more control than $G_j$. Hence $B_n^k = 0$ for $k > 2$ and $B_n^1 + B_n^2 = 2^{n-1}(2^{n-1} - 1)$.

Let us observe the recursive scheme for the first column (see Procedure 1.1). The following lemma can be proven inductively from this scheme.

**Lemma 2.4.3.** If

$$\ell = 2^{s_1-1} + \sum_{m=1}^{j} (2^{s_m-1} - 1),$$

where $1 \leq s_1 < s_2 < \cdots < s_j \leq n - 1$ and $1 \leq j \leq n - 1$

then

$$b_\ell = 1 + \sum_{m=1}^{j} 2^{s_m-1}, \quad \text{and} \quad C_\ell = (\cdots * c_{s_2} * \cdots * c_{s_1} * \cdots *),$$

(2.4)

where $(c_{s_2}, c_{s_1}) = (\ast, V)$ when $j = 1$, otherwise $(c_{s_2}, c_{s_1}) = (1, V)$.

**Lemma 2.4.4.** Let $G_1, \ldots, G_{N/2-1}$ be as in remark 2.4.2. Suppose $G_\ell$ is a $(c_{\ell_n} \ldots c_{\ell_1})$-gate. Then the following holds

$$\#\{\ell | c_{\ell_k} = 1\} = \begin{cases} n - 1 & \text{when } k = n, \\ 2^{n-k-1}(k - 1) & \text{otherwise}. \end{cases}$$

**Proof.** We want to know how many of the $G_\ell$’s have a 1-control in the $k^{th}$ bit. By Lemma 2.4.3, we know that the $G_\ell$’s satisfying this annihilate entries $b_\ell$ of the form given in equation (2.4), where $s_2 = k$ and $s_j = n$. If $k = n$, then $j = 2$ and thus we have $(n - 1)$ choices for $s_1$. If $k < n$, we have $k - 1$ choices for $s_1$ and we are free to choose which ones in $\{2^{k+1}, \ldots, 2^{n-1}\}$ to include in the sum defining $b_\ell$. The conclusion then follows.

Next, let us look at the gates used to annihilate entries of column $\ell \in \{1, \ldots, \frac{N}{2}\}$ that contribute to $B_n^1$.

**Theorem 2.4.5.** Let $2^{m-1} < \ell \leq 2^m$ with $1 \leq m \leq n - 1$ and $G_1, \ldots, G_{\frac{N}{2}-1}$ be as in Lemma 2.4.4. Then

$$\#\{i | G_\ell(G_i) \text{ has exactly one control }\} = \begin{cases} n - 1 & \text{if } m = n - 1, \\ (n - 1) + \sum_{k=m+1}^{n-1} 2^{n-k-1}(k - 1) & \text{otherwise}. \end{cases}$$

38
Proof. If $2^{m-1} < \ell \leq 2^m$, then $\ell = \sum_{j=1}^{m} \tilde{\ell}_j 2^{j-1} + 1$. Recall that $G_\ell(G_i)$ has exactly one control if $G_i = (c_{i_1}, \ldots, c_{i_n})$ has its one 1-control in $\{c_{i(m+1)}, \ldots, c_{i_n}\}$. Thus

$$\#\{i | G_\ell(G_i) \text{ has exactly one control} \} = \bigcup_{k=m+1}^{n} \#\{i | c_{ik} = 1\}$$

The conclusion follows from Lemma 2.4.4. \hfill \Box

Proof of Theorem 2.4.1

1. A control-free gate can only be utilized in Column 1. This is because when we transform the matrix to the form $[1] \oplus U'$, the succeeding gates must make sure that the first row does not interact with other rows. As mentioned in Lemma 2.4.3 and illustrated in Table 1, these gates with no control are the gates that annihilate the entries of the form $(1 + 2^s m, 1)$ for $m \in \{1, \ldots, n\}$. Indeed $g_1^0 = n$.

2. We have shown that $g_1^0 = 1$, $g_1^1 = 4$ and $g_2^3 = 7$. From Remark 2.4.2 we deduce that $g_{n-1}^n = \binom{n-1}{n-1} + g_{n-1}^{n-2}$ for all $n \geq 4$ and hence

$$g_{n-1}^n = 1 + g_{n-1}^{n-2} = (n-3) + g_3^2 = (n-3) + 7.$$

3. Now, assume $n - 1 > k \geq 3$. From Remark 2.4.2, we get $g_k^n = g_k^n + \binom{n-1}{k} + 0 + g_{k-1}^{k-1}$.

4. When $n = 2$, we know that that $g_2^1 = 4 = 2(2-1)(2^0+1)$.

Now, assume $n > 2$. From Remark 2.4.2, $g_1^n = g_1^{n-1} + B_1^n + (n-1) + g_0^{n-1}$. Let us look at the summation defining $B_1^n$. From Remark 2.4.2.4, Column 1 contributes $\frac{n}{2} - 1 = 2^{n-1} - 1$ gates to $B_1^n$. From Lemma 2.4.5, we deduce that

$$B_1^n = (2^{n-1} - 1) + 2^{n-2}(n-1) + \sum_{m=1}^{n-2} 2^{m-1} \left[ (n-1) + \sum_{k=m+1}^{n-1} 2^{n-k-1}(k-1) \right]$$

$$= (2^{n-1} - 1) n + [2^{n-3} n(n-3) - 2^{n-2} + n] = 2^{n-3} (n+2)(n-1).$$

39
Thus \( g_1^n - g_{n-1}^1 = 2(n - 1) + 2^{n-3}(n + 3)(n + 2)(n - 1). \) Using a telescoping sum, we get

\[
g_1^n = g_1^1 + \sum_{m=3}^{n} \left[ 2(m - 1) + 2^{m-3}(m + 3)(m + 2)(n - 1) \right] = (2^{n-2} + 1)(n)(n - 1).
\]

5. If \( n = 3, g_3^2 = 7 = \frac{1}{3}(4^3 - 4) - 2^3(3 - 1) + \frac{3 \cdot 2 \cdot 1}{2}. \) Now, assume \( n > 3. \) From Remark 2.4.2 and equation (2.5),

\[
g_3^n = g_3^{n-1} + g_1^{n-1} + \binom{n - 1}{2} + 2^{n-1}(2^{n-1} - 1) - 2^{n-3}(n + 2)(n - 1).
\]

Then

\[
g_3^n - g_3^{n-1} = (2^{n-3} + 1)(n - 1)(n - 2) + \frac{(n - 2)(n - 1)}{2} + 2^{n-1}(2^{n-1} - 1) - 2^{n-3}(n + 2)(n - 1)
\]

\[= 2^{n-1}(2^{n-1} - n) + \frac{3}{2}(n - 2)(n - 1).\]

And hence

\[
g_3^n = g_3^2 + \sum_{m=4}^{n} \left[ 2^{m-1}(2^{m-1} - m) + \frac{3}{2}(m - 2)(m - 1) \right] = \frac{1}{3}(4^n - 4) - 2^n(n - 1) + \frac{n(n - 1)(n - 2)}{2}.
\]

\[\square\]

In [90], the Gray code basis was utilized to achieve the same goal of this chapter. Let us denote the total number of gates with \( k \) controls in the decomposition scheme presented in [90] by \( g_n^k. \) The recursion formula presented in the said study is

\[g_n^k = g_{n-1}^k + g_{n-1}^{k-1} + \max(2^{n-2}, 2^k) + (2^{2n-k-2} - 2^{n-2}) \quad (\text{for } k \geq 1)\]

with the conditions that \( g_n^0 = 2^{n-1} \) and \( g_n^n = 0 \) for all \( n. \) Let us compare values for small \( n. \)
FIG. 2.2: \( n \) versus \( \log_{10}(T_2(n) - T_1(n)) \) graph

<table>
<thead>
<tr>
<th>( n )</th>
<th>( g_n^0 / g_n^0 )</th>
<th>( g_n^1 / g_n^1 )</th>
<th>( g_n^2 / g_n^2 )</th>
<th>( g_n^3 / g_n^3 )</th>
<th>( g_n^4 / g_n^4 )</th>
<th>( T_1(n) / T_2(n) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 / 1</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0 / 0</td>
</tr>
<tr>
<td>2</td>
<td>2 / 2</td>
<td>4 / 4</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>4 / 4</td>
</tr>
<tr>
<td>3</td>
<td>3 / 4</td>
<td>18 / 14</td>
<td>7 / 10</td>
<td>–</td>
<td>–</td>
<td>32 / 34</td>
</tr>
<tr>
<td>4</td>
<td>4 / 8</td>
<td>60 / 50</td>
<td>48 / 40</td>
<td>8 / 22</td>
<td>–</td>
<td>180 / 196</td>
</tr>
<tr>
<td>5</td>
<td>5 / 16</td>
<td>180 / 186</td>
<td>242 / 154</td>
<td>60 / 94</td>
<td>9 / 46</td>
<td>880 / 960</td>
</tr>
</tbody>
</table>

TABLE 2.4: Comparison of total cost of decomposing \( n \)-qubit quantum gates into a product of controlled gates using the scheme presented in this chapter \( (T_1(n)) \) and that of [90] \( (T_2(n)) \).

Here, \( T_1(n) \) (respectively, and \( T_2(n) \)) is the total number of controls in the decomposition of \( U \in U_{2^n} \) using the scheme presented in this chapter (respectively, the scheme in [90]). Starting from \( n = 3 \), we get a small advantage in our decomposition and because both methods are recursive, the discrepancy becomes large as \( n \) gets larger. For example, \( T_2(10) - T_1(10) = 30,720 \). In Figure 2.2, we plot the difference between \( T_2 \) and \( T_1 \) for \( n \) from 1 to 50. We use the log scale in the \( y \)-axis.
2.5 Concluding Remarks and Future Research

In this chapter, we present a recurrence scheme for generating controlled single qubit unitary gates $U_1, \ldots, U_r$ with $r \leq N(N - 1)/2$ such that $U_r \cdots U_1 U = I_N$. Consequently, $U = U_1^\dagger \cdots U_r^\dagger$.

We have the following.

<table>
<thead>
<tr>
<th>Recurrence scheme</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Step 1</strong> Partition $U \in U_n$ into a $2 \times 2$ block matrix with each block is $N/2 \times N/2$, where $N = 2^n$.</td>
</tr>
<tr>
<td><strong>Step 2</strong> Use the scheme of the $(n - 1)$-qubit case to help reduce $U$ to the form $I_{N/2} \oplus \tilde{U}$ with $\tilde{U} \in U_{N/2}$.</td>
</tr>
<tr>
<td><strong>Step 2.1</strong> For Column 1, use Procedure 2.1 in Section 3.</td>
</tr>
<tr>
<td><strong>Step 2.2</strong> For Column $\ell$ with $2 \leq \ell \leq N/2$, use Procedure 2.2 in Section 3.</td>
</tr>
<tr>
<td><strong>Step 3</strong> Apply the scheme of the $(n - 1)$-qubit case to transform $\tilde{U}$ to $I_{N/2}$.</td>
</tr>
</tbody>
</table>

It is worth noting that one can actually describe the entire recursive scheme in terms of the steps used to eliminate the off-diagonal entries of the first column as follows.

- We first generate the $(c_n \cdots c_1)$-gates for eliminating the off-diagonal entries:

  For $n = 1$ use $V$ to eliminate the $(2,1)$ entry; for $n > 1$ modify the $(c_{n-1} \cdots c_1)$-gates to $(c_{n-1} \cdots c_1)$-gates to eliminate the off-diagonal entries in upper half of Column 1 in the $n$-qubit case, and $G(c_{n-1} \cdots c_1)$-gates to eliminate the entries in the lower half.

- Once, we have the $(c_n \cdots c_1)$-gates for Column 1, we can modify them to eliminate the off-diagonal entries for the leading $2^m \times 2^m$ blocks for $m = 1, \ldots, n$, using Steps 2.1 and 2.2 described in Section 3.

We give recursive formulas for the number of controlled single qubit gates needed in the decomposition. The total number of controls used in our scheme is less than that in [90].

For future research, it might be interesting to design other recurrence schemes, which are easy to implement and use even less controls. Moreover, there might be other optimality criteria depending on the physical implementation of qubits. One may take this into consideration and
assign a cost $w_k$ for implementing $k$-controlled single qubit gates, and then study the optimal decomposition by minimizing the cost instead of number of controls.

A Matlab program `decomposition.m` implementing our decomposition scheme can be found in Appendix A.2. Another Matlab script `gatecount.m` counts the total number of controls in our scheme and that of [90].
CHAPTER 3

Optimal Bounds on Functions of Quantum States under Quantum Channels

3.1 Introduction

In quantum sciences research, one often compares a pair of quantum states $\rho_1, \rho_2$ by considering some scalar functions $D(\rho_1, \rho_2)$. For instance, in quantum information and quantum control, one would like to measure the ‘distance’ between a state $\rho_1$ and another state $\rho_2$ which go through a quantum channel or a quantum operation $\Phi$. The following measures are often used [8, 40]:

\[
(\text{tr}|\rho_1 - \rho_2|^2)^{1/2}, \quad \frac{1}{2} \text{tr}|\rho_1 - \rho_2|, \quad \sqrt{2}\sqrt{1 - \text{tr}|\sqrt{\rho_1} \sqrt{\rho_2}|},
\]

which are known as the Hilbert-Schmidt (HS) distance, the trace distance and the Bures distance, respectively. Here $|\rho|$ is the positive semidefinite square root of $\rho^* \rho$. In particular, the Bures

*The material in this chapter is contained in the paper [64], which is a joint work of C.K. Li, K.Z. Wang and the author.
The purpose of this paper is to study the following.

**Problem 3.1.1.** Let $D$ be a scalar function on a pair of quantum states. Suppose $\rho_1, \rho_2$ are two quantum states and $\mathcal{S}$ is a set of quantum channels. Determine the optimal bounds for $D(\rho_1, \Phi(\rho_2))$ for $\Phi \in \mathcal{S}$, and also the states $\sigma = \Phi(\rho_2)$ attaining the optimal bounds.

These optimal bounds provide insight on the geometry of certain sets of quantum states [73, 75] and play an important role in quantum state discrimination [17, 39, 49]. Physically, if quantum state $\rho_2$ goes through some quantum channel $\Phi$, one would like to know $D(\rho_1, \Phi(\rho_2))$ for another fixed quantum state $\rho_1$. If $\Phi$ is under our control, a solution to this problem can help us select $\Phi$ to attain the maximum or minimum value for $D(\rho_1, \Phi(\rho_2))$. On the other hand, if we only know that $\Phi$ lies in a certain class of quantum channels, then the solution will tell us the range of values where $D(\rho_1, \Phi(\rho_2))$ lies.

Recall that quantum channels are trace preserving completely positive map $\Phi : \mathbb{C}^{n \times n} \rightarrow \mathbb{C}^{n \times n}$ with the operator-sum representation

$$
\Phi(X) = \sum_{j=1}^{r} A_j X A_j^* \quad \text{for all } X \in \mathbb{C}^{n \times n},
$$

where $A_1, \ldots, A_r \in \mathbb{C}^{n \times n}$ satisfy $\sum_{j=1}^{r} A_j^* A_j = I_n$. The map $\Phi$ is a unitary channel if $r = 1$ and $F_1$ is unitary; it is a mixed unitary channel if every $A_j$ is a multiple of a unitary matrix; it is unital if $\Phi(I_n) = I_n$.

In the next two sections, we will obtain results for two general classes of functions $D(\cdot, \cdot)$. The first type of functions will cover the Hilbert-Schmidt (HS) distance and the trace distance. The second type will cover the fidelity, the Bures distance, and also the relative entropy defined...
by

\[ H(\rho_1||\rho_2) = \text{tr}(\rho_1(\log \rho_1 - \log \rho_2)). \]

For each class of functions, we will give the complete solution of Problem 3.1.1 when \( S \) is the set of unitary quantum channels, the set of mixed unitary channels and the set of unital quantum channels. These will be done in the next two sections. We also consider the set of all quantum channels and obtain a complete answer for the first class of functions, and partial results for the second class of functions. Some concluding remarks and future research directions will be mentioned in Section 3.4.

Recall that \( \mathcal{D}_n \) is the set of \( n \times n \) density matrices. By the following result ([67], Theorem 3.6), the solutions of Problem 3.1.1 are the same for the set of mixed unitary channels and the set of unital channels.

**Lemma 3.1.2.** Let \( \rho, \sigma \in \mathcal{D}_n \). The following are equivalent.

1. There exists a mixed unitary quantum channel \( \Phi \) such that \( \Phi(\rho) = \sigma \).
2. There exists a unital quantum channel \( \Phi \) such that \( \Phi(\rho) = \sigma \).
3. \( \sigma \prec \rho \).
4. There exist \( U_1, \ldots, U_n \in \mathcal{U}_n \) such that \( \sigma = \frac{1}{n}(U_1^*\rho U_1 + \cdots + U_n^*\rho U_n) \).

### 3.2 Schur Convex Functions

In Chapter 1.5, we defined the Schatten \( p \)-norm \( \| \cdot \|_p \) for \( p \geq 1 \). The Hilbert Schmidt distance is \( \| \cdot \|_2 \) and, up to a multiple, the trace distance is \( \| \cdot \|_1 \).

In [75, Theorem 4], the authors observed that

\[
\max_{U \text{ is unitary}} \| \rho_1 - U\rho_2U^* \|_1 = \| \Lambda^\downarrow(\rho_1) - \Lambda^\uparrow(\rho_2) \|_1,
\]

46
and
\[
\min_{U \text{ is unitary}} \|\rho_1 - U\rho_2 U^*\|_1 = \|\Lambda^\downarrow(\rho_1) - \Lambda^\downarrow(\rho_2)\|_1,
\]
where \(\Lambda^\downarrow(X)\) (respectively, \(\Lambda^\uparrow(X)\)) denotes the diagonal matrix having the eigenvalues of \(X\) as diagonal entries arranged in descending order (respectively, ascending order).

Actually, the same result holds if one replaces \(\|\cdot\|_1\) by any unitary similarity invariant norms \(\|\cdot\|\). To describe the full generalization of the result, we need the notion of majorization and Schur convex functions discussed in 1.5.

Denote by \((\lambda_1(X), \ldots, \lambda_n(X)) = \text{eig}^\downarrow(X)\) — the vector of eigenvalues of \(X \in \mathcal{H}_n\). Note that if \(\rho \in \mathcal{D}_n\), then \(\text{eig}^\downarrow(\rho)\) is in the set
\[
\Omega_n = \{(x_1, \ldots, x_n) : x_1 \geq \cdots \geq x_n \geq 0, \ x_1 + \cdots + x_n = 1\}. \tag{3.1}
\]

We have the following.

**Theorem 3.2.1.** Suppose the function \(D : \mathcal{D}_n \times \mathcal{D}_n \to \mathbb{R}\) is defined by \(D(\sigma_1, \sigma_2) = d(\text{eig}^\downarrow(\sigma_1 - \sigma_2))\) for a Schur convex function \(d : \mathbb{R}^n \to \mathbb{R}\). Then
\[
\max_{U \in \mathcal{U}_n} D(\rho_1, U\rho_2 U^*) = D(\Lambda^\downarrow(\rho_1), \Lambda^\downarrow(\rho_2)),
\]
and
\[
\min_{U \in \mathcal{U}_n} D(\rho_1, U\rho_2 U^*) = D(\Lambda^\downarrow(\rho_1), \Lambda^\downarrow(\rho_2)).
\]
The maximum is attained at \(U\rho_2 U^*\) if there exists a \(V \in \mathcal{U}_n\) such that \(V\rho_1 V^* = \Lambda^\downarrow(\rho_1)\) and \(VU\rho_2 U^*V^* = \Lambda^\downarrow(\rho_2)\). The minimum is attained at \(U\rho_2 U^*\) if there exists a \(V \in \mathcal{U}_n\) such that \(V\rho_1 V^* = \Lambda^\downarrow(\rho_1)\) and \(VU\rho_2 U^*V^* = \Lambda^\downarrow(\rho_2)\). The converses of the two preceding statements are also true if \(d\) is strictly Schur convex.

Theorem 3.2.1 provides a complete solution to Problem 3.1.1 for the set \(\mathcal{S}\) of unitary channels if \(D(\sigma_1, \sigma_2) = d(\text{eig}^\downarrow(\sigma_1 - \sigma_2))\) for a Schur convex function \(d(\cdot)\). In particular, it provides information about the state \(\sigma = \Phi(\rho_2)\) that attains the maximum and minimum values.
For example, take $\rho_1 = \text{diag}(.55, .45, 0)$ and $\rho_2 = \text{diag}(.35, .33, .32)$, and let $|| \cdot ||$ be any u.s.i. norm. Since all u.s.i. norms are Schur-convex, then $||\rho_1 - \rho_2||$ and $||\rho_1 - \text{diag}(.32, .33, .35)||$ will yield the minimum and maximum values in the set $\{||\rho_1 - U\rho_2 U^*|| : U \text{ is unitary}\}$. Furthermore, if we choose a norm $|| \cdot ||$ that corresponds to a strictly Schur-convex function such as the Schatten $p-$norm for $p \in (1, \infty)$, then the lower bound and upper bound can only occur at the matrices $\rho_2$ and $\text{diag}(.32, .33, .35)$, respectively. On the other hand, for the Schatten 1-norm, i.e., the trace norm, the minimum may occur at other matrices such as $U\rho_2 U^* = \text{diag}(.33, .32, .35)$.

Another situation where the optimal is attained by multiple states may arise when $\rho_1$ has repeated eigenvalues. For example, if $\rho = \frac{1}{n}I_n$, then for any $\Phi \in S$, $\Phi(\rho_2)$ attains the maximum/minimum.

Next, we turn to Problem 3.1.1 for the set $S$ of mixed unitary channels and unital channels. By Lemma 3.1.2, and the results in [71], we have the following solution of Problem 3.1.1 if $D(\sigma_1, \sigma_2) = d(\text{eig}^\downarrow(\sigma_1 - \sigma_2))$ for a Schur convex function $d(\cdot)$ and $S$ is the set of mixed unitary channels or the set of unital channels. Furthermore, as shown in Lemma 3.1.2, we can always construct the mixed unitary channel of the form

$$\sigma \mapsto \frac{1}{n}(U_1\rho U_1^* + \cdots + U_n\rho U_n^*)$$

for some $U_1, \ldots, U_n \in U_n$.

**Theorem 3.2.2.** Suppose the function $D : D_n \times D_n \rightarrow \mathbb{R}$ is defined by $D(\sigma_1, \sigma_2) = d(\text{eig}^\downarrow(\sigma_1 - \sigma_2))$ for a Schur convex function $d : \mathbb{R}^n \rightarrow \mathbb{R}$. Let $S$ be the set of mixed unitary channels or the set of unital channels acting on $\mathbb{C}^{n \times n}$. Then

$$\max_{\Phi \in S} D(\rho_1, \Phi(\rho_2)) = D(\Lambda^\downarrow(\rho_1), \Lambda^\downarrow(\rho_2)) \quad \text{and} \quad \min_{\Phi \in S} D(\rho_1, \Phi(\rho_2)) = D\left(\Lambda^\downarrow(\rho_1), \sum_{j=1}^n d_j E_{jj}\right),$$

where $(d_1, \ldots, d_n)$ is determined by the following algorithm:

**Step 0.** Set $(\Delta_1, \ldots, \Delta_n) = \text{eig}^\downarrow(\rho_1) - \text{eig}^\downarrow(\rho_2)$.

**Step 1.** If $\Delta_1 \geq \cdots \geq \Delta_n$, then set $(d_1, \ldots, d_n) = \text{eig}^\downarrow(\rho_1) - (\Delta_1, \ldots, \Delta_n)$ and stop.
Else, go to Step 2.

Step 2. Let $1 \leq j < k \leq \ell \leq n$ be such that

$$\Delta_1 \geq \cdots \geq \Delta_{j-1} > \Delta_j = \cdots = \Delta_{k-1} < \Delta_k = \cdots = \Delta_{\ell} \neq \Delta_{\ell+1}.$$ 

Replace each $\Delta_j, \ldots, \Delta_\ell$ by $(\Delta_j + \cdots + \Delta_\ell)/(\ell - j + 1)$, and go to Step 1.

The maximum is attained at $\Phi \in \mathcal{S}$ if there exists a unitary $V$ satisfying $V\rho_1V^* = \Lambda^\downarrow(\rho_1)$ and $V\Phi(\rho_2)V^* = \Lambda^\uparrow(\rho_2)$. The minimum is attained at $\Phi \in \mathcal{S}$ if there exists a unitary $V$ satisfying $V\rho_1V^* = \Lambda^\downarrow(\rho_1)$ and $V\Phi(\rho_2)V^* = \sum_{j=1}^n d_j E_{jj}$. The converses of the above two statements also hold if $d$ is strictly Schur-convex.

Here is an example illustrating the construction in the theorem.

**Example 3.2.3.** Let $\rho_1 = \frac{1}{10}\text{diag}(4,3,3,0)$ and $\rho_2 = \frac{1}{10}\text{diag}(5,2,2,1)$.

Apply Step 0. Set $(\Delta_1, \ldots, \Delta_4) = \frac{1}{10}\text{diag}(4,3,3,0) - \frac{1}{10}\text{diag}(5,2,2,1) = \frac{1}{10}\text{diag}(-1,1,1,-1)$.

Apply Step 2. Change $(\Delta_1, \ldots, \Delta_4)$ to $\frac{1}{10}\text{diag}(1/3,1/3,1/3,-1)$.

Apply Step 1. Set $(d_1, \ldots, d_4) = \frac{1}{10}\text{diag}(4,3,3,0) - \frac{1}{10}\text{diag}(1/3,1/3,1/3,-1) = \frac{1}{30}\text{diag}(11,8,8,3)$.

Finally, we consider the set $\mathcal{S}$ of all quantum channels. It is known that for any two quantum states, there is a quantum channel sending the first one to the second one. We have the following.

**Theorem 3.2.4.** Suppose the function $D : \mathcal{D}_n \times \mathcal{D}_n \to \mathbb{R}$ is defined by $D(\sigma_1, \sigma_2) = d(\text{eig}^\downarrow(\sigma_1 - \sigma_2))$ for a Schur convex function $d : \mathbb{R}^n \to \mathbb{R}$. Let $\mathcal{S}$ be the set of all quantum channels acting on $\mathbb{C}^{n \times n}$. Then

$$\max_{\Phi \in \mathcal{S}} D(\rho_1, \Phi(\rho_2)) = D(\Lambda^\downarrow(\rho_1), E_{nn}) \quad \text{and} \quad \min_{\Phi \in \mathcal{S}} D(\rho_1, \Phi(\rho_2)) = D(\rho_1, \rho_1).$$

The minimum is attained at $\Phi \in \mathcal{S}$ if $\Phi(\rho_2) = \rho_1$. The maximum is attained at $\Phi \in \mathcal{S}$ if there exists a unitary $V$ satisfying $V\rho_1V^* = \Lambda^\downarrow(\rho_1)$ and $V\Phi(\rho_2)V^* = E_{nn}$. If, in addition, $d$ is strictly Schur-convex, then the converses of the two preceding statements are also true.
Proof. The conclusion on the minimum is clear. For the maximum, note that for any \( \sigma \in \mathcal{D}_n \),

\[
\sum_{j=1}^k \lambda_j (\rho_1 - \sigma) \leq \sum_{j=1}^k \lambda_j (\rho_1) - \sum_{j=1}^k \lambda_j (-\sigma) \leq \sum_{j=1}^k \lambda_j (\rho_1) - \sum_{j=1}^k \lambda_j (-E_{nn}) = \sum_{j=1}^k \lambda_j (\Lambda^\downarrow (\rho_1) - E_{nn})
\]

for \( j = 1, \ldots, n - 1 \), and \( \sum_{j=1}^n \lambda_j (\rho_1 - \sigma) = 0 \). Because \( d(\cdot) \) is Schur convex, the result follows.

\( \square \)

3.3 Fidelity, relative entropy, and other functions

In this section, we consider Problem 3.1.1 for other functions including the fidelity

\[
F(\rho_1, \rho_2) = \text{tr} \left( \sqrt{\sqrt{\rho_2} \rho_1 \sqrt{\rho_2}} \right) = \| \sqrt{\rho_1} \sqrt{\rho_2} \|_1 = \text{tr} |\rho_1^{1/2} \rho_2^{1/2}|,
\]

and the relative entropy

\[
H(\rho|\sigma) = \text{tr}(\rho (\log \rho - \log \sigma)) = \text{tr}(\rho \log \rho) - \text{tr}(\rho \log \sigma).
\]

In [94], it was shown that if \( \mathcal{S} \) is the set of unitary channels, then

\[
\max_{\Phi \in \mathcal{S}} F(\rho_1, \Phi(\rho_2)) = F(\Lambda^\downarrow (\rho_1), \Lambda^\downarrow (\rho_2)) = \sum_{j=1}^n \sqrt{\lambda_j (\rho_1)} \sqrt{\lambda_j (\rho_2)},
\]

and

\[
\min_{\Phi \in \mathcal{S}} F(\rho_1, \Phi(\rho_2)) = F(\Lambda^\uparrow (\rho_1), \Lambda^\uparrow (\rho_2)) = \sum_{j=1}^n \sqrt{\lambda_j (\rho_1)} \sqrt{\lambda_{n-j+1} (\rho_2)}.
\]

If \( \mathcal{S} \) is the set of unital channels, it was also shown that the above minimum is also valid, but determining the maximum is an open problem.

In the following, we consider different functions \( f \) and \( g \) on quantum states and study upper
bounds and lower bounds for a function $D : \mathcal{D}_n \times \mathcal{D}_n \to \mathbb{R}$ of the form

$$D(\rho_1, \rho_2) = \text{tr} f(\rho_1) g(\Phi(\rho_2)) \quad \text{and} \quad D(\rho_1, \rho_2) = |\text{tr} f(\rho_1) g(\Phi(\rho_2))|$$  \hspace{1cm} (3.2)$$

with $\Phi \in \mathcal{S}$ for different sets $\mathcal{S}$ of quantum channels. The results will cover a number of important functions in quantum information research, and the techniques based on the theory of majorization can be further extended to other functions.

To present our results, we need some more definitions and results in majorization (see [76]) to present our general theorem.

A scalar function $f : [0, 1] \to \mathbb{R}$ can be extended to $f : \mathcal{D}_n \to \mathcal{H}_n$ such that $f(\sigma) = U^\dagger \text{diag}(f(\mu_1), \ldots, f(\mu_n)) U$ if $\sigma = U^\dagger \text{diag}(\mu_1, \ldots, \mu_n) U$, where $\mu_1 \geq \cdots \geq \mu_n \geq 0$ and $U$ is unitary.

For two vectors $x, y \in \mathbb{R}^n$, $x$ is weakly majorized by $y$, denoted by $x \prec_w y$ if the sum of the $k$ largest entries of $x$ is not larger than that of $y$ for $k = 1, \ldots, n$. Furthermore, for $x, y \in \mathbb{R}$ have nonnegative entries, $x$ is log majorized by $y$, denoted by $x \prec_{\log} y$ if the product of the entries of $x$ is the same as that of $y$, and the product of the $k$ largest entries of $x$ is not larger than that of $y$ for $k = 1, \ldots, n - 1$. It is known that $x \prec_{\log} y$ then $x \prec_w y$.

We have the following.

**Theorem 3.3.1.** Let $f, g : [0, 1] \to \mathbb{R}$, $\rho_1, \rho_2 \in \mathcal{D}_n$.

(a) If $f(\rho_1)$ and $g(\rho_2)$ have eigenvalues $a_1 \geq \cdots \geq a_n$ and $b_1 \geq \cdots \geq b_n$, then

$$\min_{U \in \mathcal{U}_n} \text{tr}(f(\rho_1) g(U \rho_2 U^\dagger)) = \sum_{j=1}^n a_j b_{n-j+1}, \quad \max_{U \in \mathcal{U}_n} \text{tr}(f(\rho_1) g(U \rho_2 U^\dagger)) = \sum_{j=1}^n a_j b_j$$

The minimum is attained at a unitary $U$ if and only if there exists a unitary $V$ such that $V^* f(\rho_1) V = \text{diag}(a_1, \ldots, a_n)$ and $V^* g(U^* \rho_2 U) V = g(V^* U^* \rho_2 U V) = \text{diag}(b_n, \ldots, b_1)$.

The maximum is attained at a unitary $U$ if and only if there exists a unitary $V$ such that $V^* f(\rho_1) V = \text{diag}(a_1, \ldots, a_n)$ and $V^* g(U^* \rho_2 U) V = g(V^* U^* \rho_2 U V) = \text{diag}(b_1, \ldots, b_n)$.
(b) If \( f(\rho_1) \) and \( g(\rho_2) \) have singular values \( \alpha_1 \geq \cdots \geq \alpha_n \) and \( \beta_1 \geq \cdots \geq \beta_n \), then

\[
\min_{U \in \mathcal{U}_n} \text{tr}|f(\rho_1)g(U^*\rho_2 U)| = \sum_{j=1}^{n} \alpha_j \beta_{n-j+1}, \quad \max_{U \in \mathcal{U}_n} \text{tr}|f(\rho_1)g(U^*\rho_2 U)| = \sum_{j=1}^{n} \alpha_j \beta_j.
\]

The minimum is attained at a unitary \( U \) if and only if there exists a unitary \( V \) such that \( |V^*f(\rho_1)V| = \text{diag}(\alpha_1, \ldots, \alpha_n) \) and \( |V^*g(U^*\rho_2 U)V| = |g(V^*U^*\rho_2 U V)| = \text{diag}(\beta_n, \ldots, \beta_1) \).

The maximum is attained at a unitary \( U \) if and only if there exists a unitary \( V \) such that \( |V^*f(\rho_1)V| = \text{diag}(\alpha_1, \ldots, \alpha_n) \) and \( |V^*g(U^*\rho_2 U)V| = |g(V^*U^*\rho_2 U V)| = \text{diag}(\beta_1, \ldots, \beta_n) \).

Proof. Let \( f(\rho_1), g(\rho_2) \) have eigenvalues \( a_1 \geq \cdots \geq a_n \) and \( b_1 \geq \cdots \geq b_n \), respectively. Suppose \( V \in \mathcal{U}_n \) such that \( V^*f(\rho_1)V = \text{diag}(a_1, \ldots, a_n) \). Then

\[
\text{tr}(f(\rho_1)g(U^*\rho_2 U)) = \text{tr}(\text{diag}(a_1, \ldots, a_n)Vg(U^*\rho_2 U)V^*) = \text{tr}(\text{diag}(a_1, \ldots, a_n)g(U^*\rho_2 UV^*)).
\]

By [76, II.9 Theorem H.1.g-h], we have

\[
\sum_{j=1}^{n} a_j b_{n-j+1} \leq \sum_{j=1}^{n} a_j d_j \leq \sum_{j=1}^{n} a_j b_j.
\]  \hspace{1cm} (3.3)

Evidently, the bounds are attained if the unitary matrices \( U \) have the said properties. Assertion (a) follows.

Next, suppose \( f(\rho_1), g(\rho_2) \) have singular values \( \alpha_1 \geq \cdots \geq \alpha_n \geq 0 \) and \( \beta_1 \geq \cdots \geq \beta_n \geq 0 \), respectively. Suppose \( f(\rho_1)g(U^*\rho_2 U) \) has singular values \( s_1, \ldots, s_n \). By [76, II.9 Theorem H.1.g-h],

\[
(\alpha_1 \beta_n, \ldots, \alpha_n \beta_1) \prec_{\log} (s_1, \ldots, s_n) \prec_{\log} (\alpha_1 \beta_1, \ldots, \alpha_n \beta_n),
\]

and \( \text{tr}|f(\rho_1)g(U^*\rho_2 U)| = \sum_{j=1}^{n} s_j \) satisfies

\[
\sum_{j=1}^{n} \alpha_j \beta_{n-j+1} \leq \sum_{j=1}^{n} s_j \leq \sum_{j=1}^{n} \alpha_j \beta_j.
\]

Suppose \( V \in \mathcal{U}_n \) such that \( V^*f(\rho_1)V = \text{diag}(\xi_1 \alpha_1, \ldots, \xi_n \alpha_n) \) with \( \xi_1, \ldots, \xi_n \in \{-1, 1\} \). One
easily construct $U \in \mathcal{U}_n$ so that $g(U^*\rho_2 U)$ attaining the lower and upper bounds. Evidently, only those unitary matrices having the said properties will yield the optimal bounds. Assertion (b) follows.

If $\mathcal{S}$ is the set of all unitary channels, then the lower bounds and upper bounds in Theorem 3.3.1 are attainable by $\text{tr}f(\sigma_1)g(\Psi(\sigma_2))$ for some $\Psi \in \mathcal{S}$. There are no restrictions to the real valued functions $f$ and $g$ in Theorem 3.3.1. So, it can be applied to a wide variety of situations. For example, if $f(x) = g(x) = \sqrt{x}$, we obtain the result for the fidelity function $F(\sigma_1, \sigma_2) = \text{tr}|f(\sigma_1)g(\sigma_2)|$ and conclude that for any $U \in \mathcal{U}_n$,

$$
\sum_{j=1}^{n} [\lambda_j(\rho_1)\lambda_{n-j+1}(\rho_2)]^{1/2} \leq F(\rho_1, U^*\rho_2 U) \leq \sum_{j=1}^{n} [\lambda_j(\rho_1)\lambda_j(\rho_2)]^{1/2}.
$$

If $f(x) = x$ and $g(x) = \log(x)$, then for any $U \in \mathcal{U}_n$,

$$
\sum_{j=1}^{n} \lambda_j(\rho_1) \log \lambda_{n-j+1}(\rho_2) \leq \text{tr}(\rho_1 \log(U^*\rho_2 U)) \leq \sum_{j=1}^{n} \lambda_j(\rho_1) \log \lambda_j(\rho_2).
$$

Here we use the convention that $0 \log 0 = 0$ and $a \log 0 = -\infty$ if $a \in (0, 1]$. Applying this result to $H(\sigma_1||\sigma_2) = \text{tr}\sigma_1(\log \sigma_1 - \log \sigma_2)$, we have

$$
\sum_{j=1}^{n} \lambda_j(\rho_1) \log(\lambda_j(\rho_1)/\lambda_j(\rho_2)) \leq H(\rho_1||U^*\rho_2 U) \leq \sum_{j=1}^{n} \lambda_j(\rho_1) \log(\lambda_j(\rho_1)/\lambda_{n-j+1}(\rho_2))
$$

for any $U \in \mathcal{U}_n$.

Next, we consider the set $\mathcal{S}$ of mixed unitary channels and the set of unital channels. Given $\rho_1, \rho_2 \in \mathcal{D}_n$, from Lemma 3.1.2, the following statements are true.

(i) For any $\Phi \in \mathcal{S}$, we have $\text{eig}^\downarrow(\Phi(\rho_2)) \prec \text{eig}^\downarrow(\rho_2)$.

(ii) If $f(\rho_1)$ has eigenvalues $a_1 \geq \cdots \geq a_n \geq 0$, then for any $(x_1, \ldots, x_n) \prec \text{eig}^\downarrow(\rho_2)$, there is $\Phi \in \mathcal{S}$ such that

$$
\text{tr}(f(\rho_1)g(\Phi(\rho_2))) = \sum_{j=1}^{n} a_j g(x_j), \quad \text{and} \quad \text{tr}|f(\rho_1)g(\Phi(\rho_2))| = \sum_{j=1}^{n} |a_j g(x_j)|.
$$
Hence, we have the following.

**Theorem 3.3.2.** Let \( f, g : [0, 1] \to \mathbb{R}, \rho_1, \rho_2 \in \mathcal{D}_n, \) and \( \Phi \) be a unital channel. Suppose \( f(\rho_1) \) have eigenvalues \( a_1 \geq \cdots \geq a_n, \) singular values \( \alpha_1 \geq \cdots \geq \alpha_n, \) and \( \rho_2 \) has eigenvalues \( b_1 \geq \cdots \geq b_n. \)

(a) The best lower upper and upper bounds of \( \sum_{j=1}^n \text{tr}(f(\rho_1)g(\Phi(\rho_2))) \) equal

\[
\inf \left\{ \sum_{j=1}^n a_j \lambda_{n-j+1}(g(\sigma)) : \sigma \in \mathcal{D}_n, \ eig^+(\sigma) \prec (b_1, \ldots, b_n) \right\}, \quad \text{and}
\]

\[
\sup \left\{ \sum_{j=1}^n a_j \lambda_j(g(\sigma)) : \sigma \in \mathcal{D}_n, \ eig^+(\sigma) \prec (b_1, \ldots, b_n) \right\}, \quad \text{respectively.}
\]

Suppose the function \( g(x) \) is increasing concave. Then the infimum value \( \sum_{j=1}^n a_j g(b_{n-j+1}) \) is attainable, and a unital channel \( \Phi \) will attain the infimum value if and only if there is a unitary \( V \) satisfying \( V^\dagger f(\rho_1)V = \text{diag}(a_1, \ldots, a_n) \) and \( V^\dagger g(\Phi(\rho_2))V = \text{diag}(g(b_n), \ldots, g(b_1)). \)

In particular, the infimum can be attained at a unitary channel.

(b) The best lower upper and upper bounds of \( \sum_{j=1}^n \text{tr}|f(\rho_1)g(\Phi(\rho_2))| \) equal

\[
\inf \left\{ \sum_{j=1}^n \alpha_j \lambda_{n-j+1}(|g(\sigma)|) : \sigma \in \mathcal{D}_n, \ eig^+(\sigma) \prec (b_1, \ldots, b_n) \right\}, \quad \text{and}
\]

\[
\sup \left\{ \sum_{j=1}^n \alpha_j \lambda_j(|g(\sigma)|) : \sigma \in \mathcal{D}_n, \ eig^+(\sigma) \prec (b_1, \ldots, b_n) \right\}, \quad \text{respectively.}
\]

If the functions \( f(x) \) and \( g(x) \) have non-negative values on \([0, 1] \), then the lower and upper bounds are the same as those in (a). If in addition that \( g \) is increasing concave, then the minimum exists and occurs at the same \( \Phi(\rho_2) \) matrix as in (a) so that the minimum equals \( \sum_{j=1}^n a_j g(b_{n-j+1}). \)

**Proof.** (a) We may assume that \( V^* f(\rho_1)V = \text{diag}(a_1, \ldots, a_n). \) For any \( \Phi \in \mathcal{S}, \) we have

\[
\sum_{j=1}^n a_j d_j = \text{tr}(f(\rho_1)g(\Phi(\rho_2))),
\]
where \((d_1, \ldots, d_n)\) are the diagonal entries of \(Vg(\Phi(\rho_2))V^*\). Hence, \((d_1, \ldots, d_n)\) is majorized by \(\lambda(g(\Phi(\rho_2)))\), where \(\lambda(\Phi(\rho_2))\) are majorized by \((b_1, \ldots, b_n)\). Similar to the proof of Theorem 3.3.1,

\[
\sum_{j=1}^{n} a_j \lambda_{n-j+1}(g(\Phi(\rho_2))) \leq \sum_{j=1}^{n} a_j \lambda_j(g(\Phi(\rho_2))) \leq \sum_{j=1}^{n} a_j \lambda_j(\Phi(\rho_2))
\]

Hence the forms of the best lower upper and upper bounds of \(\sum_{j=1}^{n} \text{tr}(f(\rho_1)g(\Phi(\rho_2)))\) holds. If \(g\) is increasing concave, we can apply (vi) of Table 2 in [76, I.3.B.2] to the negative of the function \(\psi : (x_1, \ldots, x_n) \in \Omega_n \mapsto \sum_{j=1}^{n} a_j g(x_{n-j+1})\) to show that \(\psi\) is Schur-concave. Thus the minimum occurs at \((x_1, \ldots, x_n) = (b_1, \ldots, b_n)\).

(b) Note that the singular values of \(g(\Phi(\rho_2))\) are \(\gamma_1 \geq \cdots \geq \gamma_n\), which are rearrangement of \(|g(x_1)|, \ldots |g(x_n)|\), where \(x_1, \ldots, x_n\) are the eigenvalues of \(\Phi(\rho_2)\) satisfies \((x_1, \ldots, x_n) \prec (b_1, \ldots, b_n)\). Now, the eigenvalues of \(|f(\rho_1)g(\Phi(\rho_2))|\) are the singular values of \(f(\rho_1)g(\Phi(\rho_2))\), which is log majorized by \((\alpha_1 \gamma_1, \ldots, \alpha_n \gamma_n)\) and log majorizes \((\alpha_1 \gamma_n, \ldots, \alpha_n \gamma_1)\). Thus,

\[
\sum_{j=1}^{n} \alpha_j \gamma_{n-j+1} \leq \text{tr}|f(\rho_1)g(\Phi(\rho_2))| \leq \sum_{j=1}^{n} \alpha_j \gamma_j.
\]

If \(f(x)\) has nonnegative values, then the eigenvalues of \(f(\rho_1)\) are the its singular values, and the same holds for \(g(\Phi(\rho_2))\). Thus, the results in (a) applies. \(\Box\)

We can specialize the result to the function \(f(x) = x\) and \(g(x) = \log(x)\) to conclude that

\[
\sum_{j=1}^{n} \lambda_j(\rho_1) \log \lambda_{n-j+1}(\rho_2) \leq \text{tr}\rho_1 \log \Phi(\rho_2)
\]

for any unital channel \(\Phi\), and hence

\[
H(\rho_1||\Phi(\rho_2)) = \text{tr}\rho_1(\log \rho_1 - \log \Phi(\rho_2)) \leq \sum_{j=1}^{n} \lambda_j(\rho_1) \log(\lambda_j(\rho_1)/\lambda_{n-j+1}(\rho_2)).
\]

For the Fidelity function

\[
F(\rho_1, \Phi(\rho_2)) = |\text{tr}\rho_1^{1/2}\Phi(\rho_2)^{1/2}|
\]
we can deduce the following result in [73]

$$\min_{\Phi \in S} F(\rho, \Phi(\sigma)) = F(\Lambda^\dagger(\rho), \Lambda^\tau(\sigma)) = \sum_{i=1}^{n} \sqrt{\lambda_i(\rho)} \sqrt{\lambda_{n-i+1}(\sigma)}.$$ 

It was noted in [73] that the maximum value is not easy to determine. As shown in Theorem 3.3.2, the upper bound of $F(\rho_1, \Phi(\rho_2)) = \text{tr}|\rho_1^{1/2}\Phi(\rho_2)^{1/2}|$ is the same as the upper bound of $\text{tr}(\rho_1^{1/2}\Phi(\rho_2))$, and one needs to determine

$$\sup\{\sum_{j=1}^{n} \lambda_j(\rho_1)^{1/2} x_j^{1/2} : x_1 \geq \cdots \geq x_n \geq 0, (x_1, \ldots, x_n) \prec \text{eig}^i(\rho_2)\}.$$ 

By the continuity of the function $f(x) = g(x) = \sqrt{x}$ and the compactness of the set $R = \{(x_1, \ldots, x_n) : x_1 \geq \cdots \geq x_n \geq 0, (x_1, \ldots, x_n) \prec \text{eig}^i(\rho_2)\}$, we see that supremum is attainable.

On the other hand, the determination of the maximum depends heavily on $\text{eig}^i(\rho_1)$ and $\text{eig}^i(\rho_2)$. For instance, if $\text{eig}^i(\rho_2) = (1/n, \ldots, 1/n)$, then $R$ is a singleton and $F(\rho_1, \Phi(\rho_2)) = \text{tr}|\rho_1^{1/2}|/\sqrt{n}$.

If $\rho_2 = \text{diag}(1, 0, \ldots, 0)$, then $R$ contains all quantum states, and $F(\rho_1, \Phi(\rho_2)) = 1$ if $\Phi(\rho_2) = \rho_1$.

On the other hand, if $\rho_1 = I_n/n$, then $I_n/n \in R$ for any $\rho_2$ so that $F(\rho_1, \Phi(\rho_2)) = 1$ for some unital channel $\Phi$.

In the following, we describe how to determine the unital channel $\Phi$ that gives rise to $\max F(\rho_1, \Phi(\rho_2))$ for given $\rho_1, \rho_2 \in D_n$. The result actually covers a larger class of functions.

**Theorem 3.3.3.** Let $D : D_n \times D_n \to \mathbb{R}$ be defined as follows.

(a) $D(\sigma_1, \sigma_2) = \text{tr}(f(\sigma_1)g(\sigma_2))$ or $D(\sigma_1, \sigma_2) = \text{tr}|f(\sigma_1)g(\sigma_2)|$, where $f(x) = x^p$ and $g(x) = x^q$

with $p, q > 0$ such that $p + q = 1$, or

(b) $D(\sigma_1, \sigma_2) = \text{tr}(f(\sigma_1)g(\sigma_2))$ with $f(x) = x$ and $g(x) = \log x$.

Suppose $S$ is the set of mixed unitary channels or the set of unital channels acting on $\mathbb{C}^{n \times n}$. If $\rho_1, \rho_2 \in D_n$ have eigenvalues $a_1 \geq \cdots \geq a_n \geq 0$ and $b_1 \geq \cdots \geq b_n \geq 0$, respectively, then

$$\max_{\Phi \in S} D(\rho_1, \Phi(\rho_2)) = \sum_{j=1}^{n} f(a_j)g(d_j),$$
where \((d_1, \ldots, d_n)\) is determined by the algorithm below.

If \(\Phi \in \mathcal{S}\) such that there exists a unitary \(V\) satisfying \(V^\dagger \rho_1 V = \Lambda^\dagger (\rho_1)\) and \(V^\dagger \Phi(\rho_2) V = (\sum_{j=1}^n d_j E_{jj})\), then the upper bound is attained.

**Algorithm 3.3.4. Algorithm for determining \(d_1 \geq \cdots \geq d_n\)**

**Step 0.** If \(a_r > 0\) and \(a_{r+1} = \cdots = a_n = 0\), let \(a = (a_1, \ldots, a_r)\) and \(b = (b_1, \ldots, b_r)\), and set \((d_{r+1}, \ldots, d_n) = (b_{r+1}, \ldots, b_n)\). (if \(a_n > 0\), then \(r = n\) and \((d_{r+1}, \ldots, d_n)\) is vacuous.)

**Step 1.** Let \(k \in \{1, \ldots, r\}\) be the largest integer such that

\[
\frac{1}{a_1 + \cdots + a_k} (a_1, \ldots, a_k) < \frac{1}{b_1 + \cdots + b_k} (b_1, \ldots, b_k).
\] (3.4)

**Step 2.** Set \((d_1, \ldots, d_k) = \frac{b_1 + \cdots + b_k}{a_1 + \cdots + a_k} (a_1, \ldots, a_k)\). **Stop if** \(k = r\). **Otherwise,** change \(r\) to

\[r - k, a = (a_{k+1}, \ldots, a_r), b = (b_{k+1}, \ldots, b_r);\] **repeat Steps 1 and 2.**

Note that in Step 0 of the algorithm above, we can alternatively choose \((d_{r+1}, \ldots, d_n) = (b_{r+1}, \ldots, b_n)\) for any doubly stochastic matrix \(S\). Also, in Step 1, \(a_1 + \cdots + a_k \neq 0\). This implies that \(b_1 + \cdots + b_k \neq 0\) because otherwise, the maximality of the choice for \(k\) in the previous iteration will be contradicted.

By Theorem 3.3.3, we see that \(H(\rho_1||\Phi(\rho_2)) \geq \text{tr}(\lambda_j(\rho_1) \log(\lambda_j(\rho_j)/d_j))\), where we use the usual convention that \(0 \log 0 = 0\) and \(a \log 0 = -\infty\) if \(a > 0\). The proof of Theorem 3.3.3 is quite involved, and will be presented in Section 3.4. We illustrate the results in Theorem 3.3.3 and Theorem 3.3.2 in the following example.

**Example 3.3.5.** Let \(\rho_1 = \frac{1}{10} \text{diag}(4, 3, 3, 0)\) and \(\rho_2 = \frac{1}{10} \text{diag}(5, 2, 2, 1)\).

Apply Step 0. Set \(d_4 = 0.1\), \(a = (.4, .3, 3)\) and \(b = (.5, .2, .2)\).

Apply Step 1. Because \((0.4, 0.3)/0.7 < (0.5, 0.2)/0.7, (0.4, 0.3, 0.3) < (0.5, 0.2, 0.2)/0.9,\)

we set \((d_1, d_2, d_3) = (0.36, 0.27, 0.27),\) and stop.

Hence, \((d_1, d_2, d_3, d_4) = (0.36, 0.27, 0.27, 0.1)\). For the set \(S\) of unital channels,

\[
\min_{\Phi \in S} F(\rho_1, \Phi(\rho_2)) = (\sqrt{4}, \sqrt{3}, \sqrt{3}, 0)(1, \sqrt{2}, \sqrt{2}, \sqrt{3})^T/10 = (2 + 2\sqrt{6})/10,
\]

57
\[
\max_{\Phi \in S} F(\rho_1, \Phi(\rho_2)) = (\sqrt{4}, \sqrt{3}, \sqrt{3}, 0)(\sqrt{3.6}, \sqrt{2.7}, 1)^T/10 = 3/\sqrt{10}
\]

and

\[
\min_{\Phi \in S} S(\rho_1||\Phi(\rho_2)) = (4, 3, 3)(\log(10/9), \log(10/9), \log(10/9))^T/10,
\]

\[
\max_{\Phi \in S} S(\rho_1||\Phi(\rho_2)) = (4, 3, 3)(\log 4, \log(3/2), \log(3/2))^T/10.
\]

The Matlab script `maxFidvN.m` in Appendix A.3 can be used to carry out the steps in algorithm 3.3.4 to find the maximum value of \(F(\rho_1, \Phi(\rho_2))\) and the minimum value of \(S(\rho_1||\Phi(\rho_2))\) over all mixed unitary or over all unital channels.

Next we consider the set \(S\) of all quantum channels. It is known that for any \(\sigma_1, \sigma_2 \in D_n\), there is a quantum channel \(\Phi\) such that \(\Phi(\sigma_1) = \sigma_2\). Recall that \(\Omega_n = \{(x_1, \ldots, x_n) : x_1 \geq \cdots \geq x_n \geq 0, x_1 + \cdots + x_n = 1\}\). Similar to Theorem 3.3.2, we have the following.

**Theorem 3.3.6.** Let \(f, g : [0, 1] \to \mathbb{R}, \rho_1, \rho_2 \in D_n, \) and \(\Phi\) be a quantum channel. Suppose \(f(\rho_1)\) have eigenvalues \(a_1 \geq \cdots \geq a_n\) and singular values \(\alpha_1 \geq \cdots \geq \alpha_n\).

(a) The best lower upper and upper bounds of \(\sum_{j=1}^n \text{tr}(f(\rho_1)g(\Phi(\rho_2)))\) equal

\[
\inf \left\{ \sum_{j=1}^n a_j \lambda_{n-j+1}(g(\sigma)) : \text{eig}^t(\sigma) = (x_1, \ldots, x_n) \in \Omega_n \right\} \quad \text{and}
\]

\[
\sup \left\{ \sum_{j=1}^n a_j \lambda_j(g(\sigma)) : \text{eig}^t(\sigma) = (x_1, \ldots, x_n) \in \Omega_n \right\},
\]

respectively.

Suppose \(g(x)\) is increasing concave, then the infimum value equal to \(g(0)\sum_{j=1}^{n-1} a_j + a_n g(1)\) is attainable, and \(\Phi \in S\) attains the infimum if and only if there is a unitary \(V\) such that \(V^* f(\rho_1)V = \text{diag}(a_1, \ldots, a_n)\) and \(V^* g(\Phi(\rho_2))V = \text{diag}(g(0), \ldots, g(0), g(1))\).

(b) The best lower upper and upper bounds of \(\sum_{j=1}^n \text{tr}|f(\rho_1)g(\Phi(\rho_2))|\) equal

\[
\inf \left\{ \sum_{j=1}^n \alpha_j \lambda_{n-j+1}(|g(\sigma)|) : \text{eig}^t(\sigma) = (x_1, \ldots, x_n) \in \Omega_n \right\} \quad \text{and}
\]

58
\[
\sup \left\{ \sum_{j=1}^{n} \alpha_j \lambda_j(|g(\sigma)|) : \text{eig}^k(\sigma) = (x_1, \ldots, x_n) \in \Omega_n \right\}, \quad \text{respectively.}
\]

If the functions \( f(x) \) and \( g(x) \) have non-negative values on \([0, 1]\), then the lower and upper bounds are the same as that in (a). If in addition that \( g \) is increasing concave, then the infimum value equals \( \text{tr}(f(\rho_1)g(\Phi(\rho_2))) = g(0) \sum_{j=1}^{n-1} a_j + a_n g(1) \) is attainable, and will occur at \( \Phi(\rho_2) \) satisfying the same conditions as in (a).

In [73], it was proved that if \( S \) is the set of all quantum channels, then

\[
\max_{\Phi \in S} F(\rho_1, \Phi(\rho_2)) = F(\rho_1, \rho_1) = 1 \quad \text{and} \quad \min_{\Phi \in S} F(\rho_1, \Phi(\rho_2)) = \lambda_{\min}(\rho_1)^{\frac{1}{2}}.
\]

By Theorem 3.3.6 and Lemma 3.4.1 in the next section, we have the following.

Corollary 3.3.7. Suppose \( S \) is the set of all quantum channels, and \( \rho_1, \rho_2 \in \mathcal{D}_n \) have eigenvalues \( a_1 \geq \cdots \geq a_n \geq 0 \) and \( b_1 \geq \cdots \geq b_n \geq 0 \), respectively. The following statements hold.

(a) If \( D(\sigma_1, \sigma_2) = \text{tr}(f(\sigma_1)g(\sigma_2)) \) or \( D(\sigma_1, \sigma_2) = \text{tr}|f(\sigma_1)g(\sigma_2)| \) with \( f(x) = x^p, g(x) = x^q \) such that \( p, q > 0 \) and \( p + q = 1 \), then

\[
\max_{\Phi \in S} D(\rho_1, \Phi(\rho_2)) = 1 \quad \text{and} \quad \min_{\Phi \in S} D(\rho_1, \Phi(\rho_2)) = f(a_n).
\]

(b) For the relative entropy function,

\[
\max_{\Phi \in S} H(\rho_1||\Phi(\rho_2)) = \infty \quad \text{and} \quad \min_{\Phi \in S} H(\rho_1||\Phi(\rho_2)) = 0.
\]

Proof. Similar to the proof of Theorem 3.3.1, we can focus on

\[
\sum_{j=1}^{n} f(a_j)g(z_j) \quad \text{and} \quad \sum_{j=1}^{n} a_j \log a_j - \sum_{j=1}^{n} a_j \log z_j
\]

over the set \( \Omega_n = \{(x_1, \ldots, x_n) : x_1 \geq \cdots \geq x_n \geq 0, \ x_1 + \cdots + x_n = 1\} \).
(a) The lower bound follows readily from Theorem 3.3.2. For the upper bound, by Lemma
3.4.1(b), we have
\[
\sum_{j=1}^{n} f(a_j)g(z_j) \leq \sum_{j=1}^{n} f(a_j)g(a_j) = \sum_{j=1}^{n} a_j^p a_j^q = 1
\]
for all \((z_1, \ldots, z_n) \in \Omega_n\).

(b) Choose \((z_1, \ldots, z_n) = (0, \ldots, 0, 1)\). Since \(a_1 > 0\), we have
\[
\sum_{j=1}^{n} a_j \log a_j - \sum_{j=1}^{n} a_j \log z_j = \infty.
\]

From Lemma 3.4.1(b), \(\sum_{j=1}^{n} a_j \log z_j \leq \sum_{j=1}^{n} a_j \log a_j\) for all \((z_1, \ldots, z_n) \in \Omega_n\). Hence
\[
\min_{(z_1, \ldots, z_n) \in \Omega_n} \left( \sum_{j=1}^{n} a_j \log a_j - \sum_{j=1}^{n} a_j \log z_j \right) = 0.
\]
The result follows.

\[\square\]

3.4 Proof of Theorem 3.3.3

To prove Theorem 3.3.3, we need some auxiliary results.

Lemma 3.4.1. Suppose \(f, g\) are defined as in Theorem 3.3.3. Given \(p_1, \ldots, p_\eta, t_1 \in [0, 1]\), let
\[
F_{p_1, \ldots, p_\eta, t_1}(x_1, \ldots, x_{\eta-1}) = f(p_1)g(x_1) + \cdots + f(p_\eta)g(t_1 - x_1 - \cdots - x_{\eta-1})
\]
for \(0 \leq x_j \leq t_1\) and \(x_1 + \cdots + x_{\eta-1} \leq t_1\). Then the following statements are true:

(a) \(F_{p_1, p_2, t_1}(x_1)\) is concave for \(x_1 \in [0, t_1]\);

(b) For any \((x_1, \ldots, x_{\eta-1}) \neq \alpha(p_1, \ldots, p_{\eta-1})\) such that \(\alpha = \frac{t_1}{p_1 + \cdots + p_\eta}\), the following holds
\[
F_{p_1, \ldots, p_\eta, t_1}(x_1, \ldots, x_{\eta-1}) < F_{p_1, \ldots, p_\eta, t_1}(\alpha p_1, \ldots, \alpha p_{\eta-1})
\]
Proof. For \( \eta = 2 \), we have \( F'_{p_1,p_2,t_1}(\frac{p_{11}}{p_1+p_2}) = 0 \) and \( F''_{p_1,p_2,t_1}(x_1) < 0 \) for all \( x_1 \in (0,t_1) \). Hence, (a) holds and in the case \( \eta = 2 \), (b) is true. Assume that \( \eta = k > 2 \). \( F_{p_1,...,p_k,t_1} \) is continuous in \( \Gamma_k \equiv \{ (x_1,...,x_{k-1}) : 0 \leq x_i \leq t_1, x_1 + \cdots + x_{k-1} \leq t_1 \} \). Since \( \Gamma_k \) is compact, there exists \( (\hat{x}_1,...,\hat{x}_{k-1}) \in \Gamma_k \) such that
\[
F_{p_1,...,p_k,t_1}(\hat{x}_1,...,\hat{x}_{k-1}) = \max_{\Gamma_k} F_{p_1,...,p_k,t_1}(\Gamma_k).
\]
From the case \( \eta = 2 \), we get \( \hat{x}_j = \hat{x}_j + \hat{x}_i p_j \) for all \( i,j \) and \( i \neq j \). This implies that \( \hat{x}_j = \alpha p_j \) for all \( j \). Since \( \hat{x}_1 + \cdots + \hat{x}_k = t_1 \), we obtain \( \alpha = \frac{t_1}{p_1+\cdots+p_k} \) and then (b) holds. \( \square \)

**Theorem 3.4.2.** Let \( f,g \) be defined as in Theorem 3.3.3 and suppose \( a = (a_1,...,a_n), b = (b_1,...,b_n), x = (x_1,...,x_n) \) are nonnegative decreasing sequences and that \( x < b \) satisfies
\[
\sum_{j=1}^{n} f(a_j)g(x_j) \equiv f(a)g(x) \geq f(a)g(y) \quad \text{for all } y < b.
\]
Then the following statements hold.

(a) There exist \( n_0 = 0 < 1 < n_1 < n_2 < \cdots < n_k = n \) such that for \( 0 \leq i < k \),
\[
\sum_{j=n_{i-1}+1}^{n_{i+1}} x_{j} = \sum_{j=n_{i-1}+1}^{n_{i+1}} b_{j} \quad \text{and} \quad (x_{n_{i-1}+1},...,x_{n_{i+1}}) = \alpha_i(a_{n_{i-1}+1},...,a_{n_{i+1}}),
\]
where \( b_{n_{i}+1} + \cdots + b_{n_{i+1}} = \alpha_i(a_{n_{i}+1} + \cdots + a_{n_{i+1}}) \).

(b) The values \( n_1,...,n_k \) in (a) can be determined as follows:
\[
n_1 = \max\{r : \alpha(a_1,...,a_r) < (b_1,...,b_r)\}, \quad \text{and} \quad n_j = \max\{r : \alpha(a_{n_{j-1}+1},...,a_r) < (b_{n_{j-1}+1},...,b_r)\} \quad \text{for } 1 < j \leq k.
\]

Proof. (a) Let \( n_1 = \max\{k : (x_1,...,x_k) = \alpha(a_1,...,a_k) \text{ for some } \alpha\} \). If \( n_1 = n \), then the proof is done.
Suppose that \( n_1 < n \). Then \((x_1, \ldots, x_{n_1}) = \alpha_0(a_1, \ldots, a_{n_1})\) and \( x_{n_1+1} \neq \alpha_0a_{n_1+1} \). We claim that \( \sum_{j=1}^{n_1} x_j = \sum_{j=1}^{n_1} b_j \). Suppose that \( \sum_{j=1}^{n_1} x_j < \sum_{j=1}^{n_1} b_j \). Let \( \beta = \frac{x_{n_1} + x_{n_1+1}}{a_{n_1} + a_{n_1+1}} \). If \( x_{n_1} = \beta a_{n_1} \), then \( x_{n_1+1} = \beta a_{n_1+1} \). Since \( x_{n_1+1} \neq \alpha_0a_{n_1+1} \) and \( x_{n_1} = \alpha_0 a_{n_1}, \beta \neq \alpha_0 \). Thus, \( \beta < \alpha_0 \) or \( \beta > \alpha_0 \).

**Case 1.** \( \beta < \alpha_0 \). Let \( \hat{x} = (x_1, \ldots, x_{n_1-1}, \beta a_{n_1}, \beta a_{n_1+1}, x_{n_1+2}, \ldots, x_n) \). We have \( \beta a_{n_1} < \alpha_0 a_{n_1} = x_{n_1} \) and \( \beta a_{n_1+1} = x_{n_1} + x_{n_1+1} - \beta a_{n_1} > x_{n_1+1} \). Hence \( \hat{x} \) is decreasing and \( \hat{x} < b \). On the other hand,

\[
\begin{align*}
    f(a)g(\hat{x}) - f(a)g(x) &= f(a_{n_1})g(\beta a_{n_1}) + f(a_{n_1+1})g(\beta a_{n_1+1}) - (f(a_{n_1})g(x_{n_1}) + f(a_{n_1+1})g(x_{n_1+1})) \\
    &= F_{a_{n_1}, a_{n_1+1}, x_{n_1} + x_{n_1+1}}(a_{n_1} x_{n_1} + x_{n_1+1}) - F_{a_{n_1}, a_{n_1+1}, x_{n_1} + x_{n_1+1}}(x_{n_1}) \\
    &> 0 \quad \text{(by Lemma 3.4.1(b)).}
\end{align*}
\]

This is a contradiction.

**Case 2.** \( \beta > \alpha_0 \). There exist \( m_1 \leq n_1 < m_2 \) such that

\[
x_{m_1-1} > x_{m_1} = \cdots = x_{n_1} \geq x_{n_1+1} = \cdots = x_{m_2} > x_{m_2+1}.
\]

We will show that \( \sum_{j=1}^{r} x_j < \sum_{j=1}^{r} b_j \) for \( m_1 \leq r < m_2 \).

**Assertion 1.** \( \sum_{j=1}^{r} x_j < \sum_{j=1}^{r} b_j \) for \( n_1 + 1 \leq r < m_2 \).

If not, then \( \sum_{j=1}^{r_0} x_j = \sum_{j=1}^{r_0} b_j \) for some \( n_1 + 1 \leq r_0 < m_2 \). Because \( \sum_{j=1}^{r_0+1} x_j \geq \sum_{j=1}^{r_0+1} b_j \), we see that \( x_{r_0+1} \geq b_{r_0+1} \). Since \( \sum_{j=1}^{n_1} x_j < \sum_{j=1}^{n_1} b_j \), we may assume \( \sum_{j=1}^{r} x_j < \sum_{j=1}^{r} b_j \) for \( n_1 \leq r < r_0 \). We get \( x_{r_0} > b_{r_0} \geq b_{r_0+1} \). But \( x_{r_0} = x_{r_0+1} \leq b_{r_0+1} \). This is a contradiction. Thus,

\[
\sum_{j=1}^{r} x_j < \sum_{j=1}^{r} b_j \quad \text{for } n_1 + 1 \leq r < m_2.
\]

**Assertion 2.** \( \sum_{j=1}^{r} x_j < \sum_{j=1}^{r} b_j \) for \( m_1 \leq r < n_1 \).

If not, then \( \sum_{j=1}^{x_1} x_j = \sum_{j=1}^{x_1} b_j \) for some \( m_1 \leq r_1 < n_1 \). Then \( x_{r_1} \geq b_{r_1} \). Since \( x_{r_1} = \cdots =
\( x_{n_1} \) and \( b_{r_1} \geq \cdots \geq b_{n_1} \), we have \( \sum_{j=1}^{r_1} x_j \geq \sum_{j=1}^{r_1} b_j \) for \( r_1 \leq r \leq n_1 \). This is impossible since \( \sum_{j=1}^{n_1} x_j < \sum_{j=1}^{n_1} b_j \). Hence \( \sum_{j=1}^{r} x_j < \sum_{j=1}^{r} b_j \) for \( m_1 \leq r < n_1 \).

By the above argument, \( \sum_{j=1}^{r} x_j < \sum_{j=1}^{r} b_j \) for \( m_1 \leq r < m_2 \). Now, let

\[ \hat{x} = (x_1, \ldots, x_{m_1-1}, x_{m_1} + \delta, x_{m_1+1}, \ldots, x_{m_2-1}, x_{m_2} - \delta, x_{m_2+1}, \ldots, x_n) \]

For sufficiently small \( \delta > 0 \), \( \hat{x} \) is decreasing and \( \hat{x} < b \). In fact,

\[ \alpha_0 < \beta = \frac{x_{n_1} + x_{n_1+1}}{a_{n_1} + a_{n_1+1}} = \frac{x_{m_1} + x_{m_2}}{a_{m_1} + a_{m_1+1}} = \frac{x_{m_1} + x_{m_2}}{a_{m_1} + a_{m_2}} \]

The third equality holds because \( \alpha_0 a_{m_1} = x_{m_1} = x_{n_1} = \alpha_0 a_{n_1} \). Hence

\[ \frac{x_{m_1} + x_{m_2}}{a_{m_1} + a_{m_2}} > \alpha_0 a_{m_1} = x_{m_1} \]

Then for sufficiently small \( \delta > 0 \),

\[
\begin{align*}
f(a)g(\hat{x}) - f(a)g(x) &= f(a_{m_1})g(x_{m_1} + \delta) + f(a_{m_2})g(x_{m_2} - \delta) - (f(a_{m_1})g(x_{m_1}) + f(a_{m_2})g(x_{m_2})) \\
&= F_{a_{m_1}, a_{m_2}, x_{m_1} + x_{m_2}}(x_{m_1} + \delta) - F_{a_{m_1}, a_{m_2}, x_{m_1} + x_{m_2}}(x_{m_1}) \\
&> 0 \text{ (by Lemma 3.4.1).}
\end{align*}
\]

This is a contradiction and then \( \sum_{j=1}^{n_1} x_j = \sum_{j=1}^{n_1} b_j \). Let

\[ n_2 = \max\{k : (x_{n_1+1}, \ldots, x_k) = \alpha(a_{n_1+1}, \ldots, a_k) \text{ for some } \alpha\} \]

From the above proof, we also have \( \sum_{j=n_1+1}^{n_2} x_j = \sum_{j=n_1+1}^{n_2} b_j \). By induction, we get the desired conclusion.
(b) Suppose \( n_1 < \eta \equiv \max\{r : \alpha(a_1, \ldots, a_r) < (b_1, \ldots, b_r)\}\). We have

\[
\sum_{j=1}^{n_1} x_j = \sum_{j=1}^{n_1} \alpha_0 a_j = \sum_{j=1}^{n_1} b_j < \sum_{j=1}^{\eta} b_j = \sum_{j=1}^{\eta} \alpha' a_j
\]

for some \( \alpha' \). Let \( 1 < r < k \) with \( n_{r-1} < \eta \leq n_r \). Then

\[
\sum_{j=1}^{\eta} \alpha' a_j = \sum_{j=1}^{n_{r-1}} b_j \geq \sum_{j=1}^{\eta} x_j = \sum_{j=1}^{n_{r-1}} b_j + \sum_{j=n_{r-1}+1}^{\eta} x_j.
\]

There is \( 0 < \alpha'' \leq \alpha' \) such that \( \sum_{j=1}^{\eta} \alpha'' a_j = \sum_{j=1}^{\eta} x_j \). Then \( \sum_{j=1}^{p} \alpha'' a_j \leq \sum_{j=1}^{p} b_j \) for \( 1 \leq p \leq \eta \).

We have \( \sum_{j=1}^{n_{r-1}} \alpha'' a_j \leq \sum_{j=1}^{n_{r-1}} b_j = \sum_{j=1}^{n_{r-1}} x_j \). So

\[
\sum_{j=n_{r-1}+1}^{\eta} \alpha'' a_j \geq \sum_{j=n_{r-1}+1}^{\eta} x_j = \sum_{j=n_{r-1}+1}^{\eta} \alpha_{n_{r-1}} a_j.
\]

Thus \( \alpha'' \geq \alpha_{n_{r-1}} \), and hence \( \alpha'' a_{\eta} \geq \alpha_{n_{r-1}} a_{\eta} = x_{\eta} \). Let \( \hat{x} = (\alpha'' a_1, \ldots, \alpha'' a_\eta, x_{\eta+1}, \ldots, x_n) \).

Then \( \hat{x} \) is decreasing and \( \hat{x} < b \). By (a), \( n_1 = \max\{k : (x_1, \ldots, x_k) = \alpha(a_1, \ldots, a_k) \text{ for some } \alpha\} \) and \( n_1 < \eta \). Hence, \( (x_1, \ldots, x_\eta) \neq \alpha'' (a_1, \ldots, a_\eta) \). We also have \( \alpha'' = \frac{x_1 + \cdots + x_\eta}{a_1 + \cdots + a_\eta} \). By Lemma 3.4.1(b),

\[
f(a)g(\hat{x}) - f(a)g(x) = \sum_{j=1}^{\eta} f(a_j)g(\alpha'' a_j) - \sum_{j=1}^{\eta} f(a_j)g(x_j) > 0.
\]

This is a contradiction. Hence \( n_1 = \eta \).

By induction, we only need to show the case \( n_2 \). From the \( n_1 \) case, we have \( \sum_{j=1}^{n_1} x_j = \sum_{j=1}^{n_1} b_j \). Thus, \( (x_{n_1+1}, \ldots, x_n) < (b_{n_1+1}, \ldots, b_n) \), and

\[
\sum_{j=n_{1}+1}^{n} f(a_j)g(x_j) \leq \max_{(y_{n_1+1}, \ldots, y_n) < (b_{n_1+1}, \ldots, b_n)} \sum_{j=n_{1}+1}^{n} f(a_j)g(y_j).
\]

On the other hand, if \( (y_{n_1+1}, \ldots, y_n) < (b_{n_1+1}, \ldots, b_n) \), then \( (x_1, \ldots, x_{n_1}, y_{n_1+1}, \ldots, y_n) < b \).
Then

\[
\sum_{j=1}^{n} f(a_j)g(x_j) = \max_{y<b} f(a)g(y) \\
\geq \sum_{j=1}^{n} f(a_j)g(x_j) + \max_{(y_{n+1}, \ldots, y_n) \prec (b_{n+1}, \ldots, b_n)} \sum_{j=n+1}^{n} f(a_j)g(y_j).
\]

This implies that

\[
\sum_{j=n+1}^{n} f(a_j)g(x_j) = \max_{(y_{n+1}, \ldots, y_n) \prec (b_{n+1}, \ldots, b_n)} \sum_{j=n+1}^{n} f(a_j)g(y_j).
\]

From the proof of the case \(n_1\), the result follows. \(\square\)

**Proof of Theorem 3.3.3** From Theorem 6.3.2, we need only to determine the maximum of

\[
\sum_{j=1}^{n} f(a_j)g(x_j)
\]

for \(x_1 \geq \cdots \geq x_n \geq 0\) and \((x_1, \ldots, x_n) \prec (b_1, \ldots, b_n)\). Suppose that \(a_r > 0\) and \(a_{r+1} = \cdots = a_n = 0\). Let \(\alpha \equiv \sum_{j=1}^{n} f(a_j)g(d_j)\) attain the maximum for \(d_1 \geq \cdots \geq d_n \geq 0\) and \((d_1, \ldots, d_n) \prec (b_1, \ldots, b_n)\). Then \(\alpha = \sum_{j=1}^{r} f(a_j)g(d_j)\) and \((d_1, \ldots, d_r) \prec_w (b_1, \ldots, b_r)\). Since \(f\) is nonnegative and \(g\) is increasing,

\[
\max\{\sum_{j=1}^{r} f(a_j)g(x_j) : x_1 \geq \cdots \geq x_r \geq 0, (x_1, \ldots, x_r) \prec_w (b_1, \ldots, b_r)\}
\]

\[
\leq \max\{\sum_{j=1}^{r} f(a_j)g(x_j) : x_1 \geq \cdots \geq x_r \geq 0, (x_1, \ldots, x_r) \prec (b_1, \ldots, b_r)\} \equiv \beta. \tag{3.5}
\]

Hence \(\alpha \leq \beta\). Given \(x_1 \geq \cdots \geq x_r \geq 0\) and \((x_1, \ldots, x_r) \prec (b_1, \ldots, b_r)\), choose

\[
(y_1, \ldots, y_n) = (x_1, \ldots, x_r, b_{r+1}, \ldots, b_n).
\]

Then \(y_1 \geq \cdots \geq y_n \geq 0\), \((y_1, \ldots, y_n) \prec (b_1, \ldots, b_n)\), and \(\sum_{j=1}^{r} f(a_j)g(x_j) = \sum_{j=1}^{n} f(a_j)g(x_j)\). We obtain \(\alpha = \beta\). By Theorem 3.4.2, we see that the algorithm will produce the state of the form \(\Phi(\rho_2)\) attaining the maximum. \(\square\)
3.5 Concluding remarks and further research

Let \((\sigma_1, \sigma_2) \mapsto D(\sigma_1, \sigma_2)\) be a scalar function on quantum states \(\rho_1, \rho_2\), such as the trace distance, the fidelity function, and the relative entropy. For two given quantum states \(\rho_1, \rho_2\), we determine optimal bounds for \(D(\rho_1, \Phi(\rho_2))\) for \(\Phi \in \mathcal{S}\) for different classes of functions \(D(\cdot, \cdot)\), where \(\mathcal{S}\) is the set of unitary quantum channels, the set of mixed unitary channels, the set of unital quantum channels, and the set of all quantum channels. Specifically, we obtain results for functions of the following form

(a) \(D(\sigma_1, \sigma_2) = d(eig^\downarrow(\sigma_1 - \sigma_2))\), where \(d(X)\) is a Schur convex function on the eigenvalues of \(X \in \mathcal{H}_n\),

(b) \(D(\sigma_1, \sigma_2) = \text{tr}(f(\sigma_1)g(\sigma_2))\), and \(D(\sigma_1, \sigma_2) = \text{tr}|f(\sigma_1)g(\sigma_2)|\), where \(f, g : [0, 1] \to \mathbb{R}\).

For the class of function in (a), optimal bounds for \(D(\rho_1, \Phi(\rho_2))\) are given for \(\Phi \in \mathcal{S}\) for the four classes of quantum channels mentioned above. Actually, the results and techniques in Section 3.2 can be extended to functions of the form

\[
D(\sigma_1, \sigma_2) = d(eig^\downarrow(\alpha \sigma_1 - \beta \sigma_2))
\]

for given \(\alpha, \beta \in \mathbb{R}\), and a Schur convex function \(d\).

For the class of functions in (b), the optimal lower and upper bounds for \(D(\rho_1, \Phi(\rho_2))\) are given for \(\Phi \in \mathcal{S}\), where \(\mathcal{S}\) is the set of unitary channels. For the set of mixed unitary channels, the set of unital channels, and the set of all quantum channels, we determine the best lower bound if \(g\) is an increasing concave function; we also find the best upper bounds for special functions including the fidelity and relative entropy functions. The results and techniques in Section 3.3 can be extended to cover functions \(D : \mathcal{D}_n \times \mathcal{D}_n \to \mathbb{R}\) of the form \(D(\sigma_1, \sigma_2) = \psi(f(\sigma_1)g(\sigma_2))\), where \(\psi(X)\) is a Schur concave function on the singular values (eigenvalues or diagonal entries) of the matrix \(X\).
There are many related problems deserving further study. For instance, one may consider Problem 3.1.1 for a wider class of functions $D$ and different classes of $S$. More generally, one may study the optimal bounds for the set

$$\{D(\rho_1, \Phi(\sigma)) : \Phi \in S, \sigma \in T\}$$

for a set $S$ of quantum channels, and a set $T$ of quantum states. If $T = \{\sigma_1, \ldots, \sigma_k\}$ is a finite set, then one can apply our results to $D(\rho_1, \Phi(\sigma_j))$ for each $j$ to get the optimal bounds for each $j$, and compare them.
CHAPTER 4

Bipartite Qubit-Qudit States with Maximally Mixed Reduced State*

4.1 Introduction

In this chapter, we look at the compact convex set

\[ S_2\left(\frac{1}{n}I_n\right) = \{ \rho \in \mathcal{D}_{2n} \mid \text{tr}_1(\rho) = \frac{1}{n}I_n \} \]  

(4.1)

Recall that when viewed as a quantum state, \(\frac{1}{n}I_n\) represents a maximally mixed system. Hence, we are looking at possible states of bipartite systems \(X = (A, B)\) such that the reduced state of \(B\) is maximally mixed. This indicates entanglement of \(A\) and \(B\) since a measurement on subsystem \(A\) will cause a loss of information on the subsystem \(B\).

Using the Choi matrix representation of a channel (up to a scalar), we see that the set (4.1) also has a one-one correspondence with the set of unital completely positive maps from \(\mathcal{H}_2\) to \(\mathcal{H}_n\) and similarly, to the set of quantum channels from \(\mathcal{H}_n\) to \(\mathcal{H}_2\). In fact, this correspondence is

*This chapter contains work done by the author with C.K. Li, and two undergraduate students E. Berry and D. Katsaros during the 2014 EXTREEMS-QED summer research program.
used to define the entropy of a quantum channel [82].

We are interested in the spectral properties of the $S_2\left(\frac{1}{n}I_n\right)$. In particular, we look at the set

$$\mathcal{E}_n = \{(a_1, \ldots, a_{2n}) \in \Omega_{2n} \mid e^{\uparrow}(A) = (a_1, \ldots, a_{2n}) \text{ for some } A \in S_2\left(\frac{1}{n}I_n\right)\}.$$  (4.2)

which is a compact convex set described by a special set of inequalities [66]. As $m$ and $n$ increase, the number of inequalities grow fast and many of these inequalities may be redundant. We wish to determine the minimal set of inequalities that describe this set.

In Section 4.2, we describe the general set of necessary inequalities that define $\mathcal{E}_n$ and deduce some general properties of $\mathcal{E}_n$. In Section 4.3, we describe $\mathcal{E}_n$ for $n = 2, 3, 4, 5, 6$. In the case of $n = 5, 6$, we describe a geometric approach to prove that a given set of inequalities is necessary and sufficient to describe $\mathcal{E}_n$. We also give some necessary conditions for $n = 7$. We will end this section by the following proposition that can easily be proven using results from [66] and the fact that if

$$\rho = \begin{bmatrix} \rho_{11} & \rho_{12} \\ \rho_{21} & \rho_{22} \end{bmatrix}$$

where $\rho_{ij} \in \mathbb{C}^{n \times n}$, then $\text{tr}(\rho) = \rho_{11} + \rho_{22}$.

**Proposition 4.1.1.** The following are equivalent.

a. $a_1, \ldots, a_{2n} \in \mathcal{E}_n$.

b. There exists $D = \text{diag}(d_1, \ldots, d_n)$, with $0 \leq d_j \leq \frac{1}{n}$ for all $j$ and a matrix $X \in \mathbb{R}^{n \times n}$ such that the matrix

$$\begin{bmatrix} D & X \\ X^T & \frac{1}{n}I_n - D \end{bmatrix}$$

(4.3)

has eigenvalues $a_1, \ldots, a_{2n}$ [42, Theorem 3].

c. There exists $\frac{1}{n} \geq d_1 \geq \cdots \geq d_n \geq 0$, and $A, B \in \mathcal{H}_{2n}$ such that $e^{\uparrow}(A) = (d_1, \ldots, d_n, 0, \ldots, 0)$, $e^{\uparrow}(B) = (\frac{1}{n} - d_1, \ldots, \frac{1}{n} - d_n, 0, \ldots, 0)$, and $e^{\uparrow}(A + B) = (a_1, \ldots, a_{2n})$. [13].

69
d. There exists \( \frac{1}{n} \geq d_1 \geq \cdots \geq d_n \geq 0 \) such that the vector of eigenvalues

\[
\alpha = \begin{pmatrix} d_1 & \cdots & d_n & 0 & \cdots & 0 \end{pmatrix},
\]

\[
\beta = \begin{pmatrix} 1/n - d_n & \cdots & 1/n - d_1 & 0 & \cdots & 0 \end{pmatrix},
\]

\[
\nu = \begin{pmatrix} a_1 & a_2 & \cdots & a_{2n} \end{pmatrix}
\]

satisfy

\[
\sum_{p \in P} \alpha_p + \sum_{q \in Q} \beta_q \geq \sum_{r \in R} \nu_r
\]

for all \((P, Q, R) \in LR_k(2n)\) and for all \(k = 1, \ldots, n\) [66], [42].

The set \( LR_k(2n) \) is described in detail in [42]. In the next section, we will give a known characterization for elements of \( LR_k(2n) \).

\[\] 4.2 Some Necessary Eigenvalue Inequalities

For an index set \( J = \{j_1, \ldots, j_k\} \subseteq \{1, \ldots, N\} \) such that \( j_1 < i_2 < \cdots < j_k \), define

\[
s(J) = (j_1 - 1, j_2 - 2, \ldots, j_k - k).
\] (4.4)

The following theorem describes triples \((P, Q, R)\) of \( k \)-subsets of \( \{1, \ldots, 2n\} \) that is contained in the set \( LR_k(2n) \).

**Theorem 4.2.1** (Horn’s Conjecture and the Saturation Conjecture [53, 56]). Let \( \alpha = (\alpha_j), \beta = (\beta_j), \nu = (\nu_j) \in \mathbb{R}^N \) arranged in nonincreasing order. There exists \( A, B, A + B \) with eigenvalues sets \( \alpha, \beta, \nu \), respectively, if and only if

1. \( \sum_{j=1}^{N} (\alpha_j + \beta_j - \nu_j) = 0 \)
2. \[ \sum_{p \in P} \alpha_p + \sum_{q \in Q} \beta_q \geq \sum_{r \in R} \nu_r \]
for any \(1 \leq k \leq n\) and any \(k\)-subsets \(P, Q, R\) of \(\{1, \ldots, N\}\) such that \(s(P), s(Q), s(R)\) are eigenvalues of \(\tilde{A}, \tilde{B}, \tilde{A} + \tilde{B}\) for some \(k \times k\) matrices \(\tilde{A}, \tilde{B}\).

To apply this theorem to our problem, we take \(\alpha, \beta, \nu\) as described in Proposition 4.1.1 d. If \(P, Q, R \subseteq \{1, \ldots, 2n\}\) such that \(|P| = |Q| = |R|\) and there exists hermitian matrices \(\tilde{A}, \tilde{B}\) satisfying \(\text{eig}(\tilde{A}) = s(P), \text{eig}(\tilde{B}) = s(Q), \text{eig}(\tilde{A} + \tilde{B}) = s(R)\), then a necessary condition for \(\nu = (a_1, \ldots, a_{2n}) \in \mathcal{E}_n\) is given by

\[ \sum_{p \in P, i \leq n} d_p + \sum_{q \in Q, q \leq n} \frac{1}{n} - d_{n-q+1} \geq \sum_{r \in R} a_r \]  \hspace{1cm} (4.5)

for some \(\frac{1}{n} \geq d_1 \geq \ldots \geq d_n \geq 0\). In particular, we can take

\[ P = \{j_1, \ldots, j_k, n+1, \ldots, n+k\} \quad \text{and} \quad Q = \{n-j_k+1, \ldots, n-j_1+1, n+1, \ldots, n+k\} \]  \hspace{1cm} (4.6)

for any \(1 \leq k \leq n\) and \(1 \leq j_1 \leq \cdots \leq j_k \leq n\). In this case, we get

\[ s(P) = \{j_1 - 1, \ldots, j_k, n-k, \ldots, n-k\}, \]  \hspace{1cm} (4.7)

\[ s(Q) = \{n-j_k, \ldots, n-j_1 - k + 1, n-k, \ldots, n-k\} \]  \hspace{1cm} (4.8)

and hence

\[ \sum_{s=1}^{2k} a_{rs} \leq k/n \]  \hspace{1cm} (4.9)

for some compatible \(R = (r_1, \ldots, r_{2k})\). Note that applying Theorem 4.2.1 to \(\nu = (-a_{2n}, \ldots, -a_1)\), \(\alpha = (0, \ldots, 0, -d_n, \ldots, -d_1)\) and \(\beta = (0, \ldots, 0, d_1 - \frac{1}{n}, \ldots, d_n - \frac{1}{n})\), we see that if (4.9) is a necessary condition for \((a_1, \ldots, a_{2n}) \in \mathcal{E}_n\), then so is

\[ \sum_{s=1}^{2k} a_{2n-r_s+1} \geq k/n. \]  \hspace{1cm} (4.10)

As an example, consider \(P = (1, n, n+1, n+2) = Q\) and \(R = (1, 2n - 2, 2n - 1, 2n)\). Then
s(P) = (0, n − 2, n − 2, n − 2) = Q and s(R) = (0, 2n − 4, 2n − 4, 2n − 4). Clearly, we can choose diagonal Hermitian \( \tilde{A}, \tilde{B}, \tilde{A} + \tilde{B} \) such that \( \text{eig}(A) = s(P) \), \( \text{eig}(B) = s(Q) \) and \( \text{eig}(A + B) = s(R) \). Using equations (4.9) and (4.10), we see that if \((a_1, \ldots, a_{2n}) \in \mathcal{E}_n\), then

\[
a_1 + a_{2n-2} + a_{2n-1} + a_{2n} \leq \frac{2}{n} \leq a_1 + a_2 + a_3 + a_{2n} \quad (4.11)
\]

In particular,

\[
a_j \leq \frac{2}{n} \quad \text{for all } j = 1, \ldots, 2n. \quad (4.12)
\]

As a result, \( \text{rank}(\rho) \geq \frac{n}{2} \) for any \( \rho \in S_2(\frac{1}{n}I_n) \). In fact, we can say more about elements of \( S_2(\frac{1}{n}I_n) \) having minimal rank.

**Proposition 4.2.2.** Suppose \( \rho \in S_2(\frac{1}{n}I_n) \). Then

a. \( \text{rank}(\rho) \geq \left\lceil \frac{n}{2} \right\rceil \)

b. Suppose \( n = 2m \) for some positive integer \( m \) and \( \rho \in S_2(\frac{1}{n}I_n) \). We have \( \text{rank}(\rho) = m \) if and only if \( \rho \) is of the form

\[
(W \otimes I_2) \begin{bmatrix}
\frac{1}{n}I_m & 0 & 0 & \frac{1}{n}I_m \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
\frac{1}{n}I_m & 0 & 0 & \frac{1}{n}I_m
\end{bmatrix} (W^* \otimes I_2) \quad (4.13)
\]

for some \( W \in \mathcal{U}_n \).

c. Suppose \( n \geq 5 \) is odd and \( \text{rank}(\rho) = \frac{n+1}{2} \). Then \( \text{eig}^1(\rho) = (\frac{2}{n}, \ldots, \frac{2}{n}, a_{\frac{n+1}{2}}, a_{\frac{n+1}{2}}, 0, \ldots, 0) \) for some \( \frac{1}{n} \leq a_{\frac{n+1}{2}} = \frac{3}{n} - a_{\frac{n+1}{2}} \leq \frac{3}{2n} \).

**Proof:** First, we prove that the following inequality holds for any \((a_i) \in \mathcal{E}_n\).

\[
a_1 + \ldots + a_j + a_{2n-3j+1} + \ldots + a_{2n} \leq \frac{2j}{n} \quad \text{for any } j \leq \left\lfloor \frac{n}{2} \right\rfloor \quad (4.14)
\]
Define \( R = (1, \ldots, j, 2n - 3j + 1, \ldots, 2n) \) and \( P = (1, \ldots, j, n - j + 1, \ldots, n, n + 1, \ldots, 2n) = Q \). Then \( \text{diag}(s(P)) + \text{diag}(s(Q)) = \text{diag}(s(R)) \). The above inequality then follows. As a consequence

\[ a_{j+1} + \ldots + a_{2n-j} \geq \frac{n-2j}{n} \quad \text{for any } j \leq \left\lceil \frac{n}{2} \right\rceil \] (4.15)

In particular, if we choose \( j = \left\lceil \frac{n}{2} \right\rceil - 1 \), we have

\[ a_{\left\lceil \frac{n}{2} \right\rceil} + \ldots + a_{2n-\left\lceil \frac{n}{2} \right\rceil+1} \geq \left(1 - \frac{2}{n} \left\lceil \frac{n}{2} \right\rceil \right) + \frac{2}{n} > 0 \] (4.16)

If \( n = 2m \) for some positive integer \( m \), and \( a_j = 0 \) for all \( j > m \), then, together with inequality (4.12), this implies \( a_1 = \ldots = a_m = \frac{2}{n} \). Suppose \( \rho \in S_2 \left( \frac{1}{n} I_n \right) \)

To prove (c), assume \( n = 2m+1 \) for some integer \( m \) and let \( P = (1, m+1, n, n+1, n+2, n+3) = Q \) and \( R = (m, m+1, 2n-3, 2n-2, 2n-1, 2n) \). Take \( \tilde{A} = (0, m-1, n-3, n-3, n-3, n-3) \) and \( \tilde{B} = (m-1, 0, n-3, n-3, n-3, n-3) \). Then \( \text{eig}(A) = s(P), \text{eig}(B) = s(Q) \) and \( \text{eig}(A+B) = s(R) \) and hence

\[ a_m + a_{m+1} + a_{2n-3} + a_{2n-2} + a_{2n-1} + a_{2n} \leq \frac{3}{n} \quad \text{(if } n = 2m+1) \] (4.17)

If \( m > 1 \) and \( a_j = 0 \) for all \( j > m+1 \), this implies \( a_1 + \ldots + a_{m-1} \geq \frac{2(m-1)}{n} \). So, by equation (4.12), \( a_1 = \ldots = a_{m-1} = \frac{2}{n} \). Also by equation (4.16), we have \( a_{m+1} \geq \frac{1}{n} \). \( \square \)

**Corollary 4.2.3.** For any \( \rho \in S_2 \left( \frac{1}{n} I_n \right) \) we get the following lower bound for the entropy of \( \rho \)

\[ H(\rho) = -\text{tr}(\rho \log(\rho)) \geq \begin{cases} 
\log n - \log 2 & \text{if } n \text{ is even} \\
\log n - \frac{n-1}{n} \log 2 & \text{if } n \text{ is odd}
\end{cases} \] (4.18)

**Proof:** Let \( \rho \in S_2 \left( \frac{1}{n} I_n \right) \) have eigenvalues \( a_1, \ldots, a_{2n} \). Define the density matrix \( \sigma \in S_2 \left( \frac{1}{n} I_n \right) \) by

\[ \sigma = \begin{cases} 
\text{diag}(\frac{2}{n}, \ldots, \frac{2}{n}, 0, \ldots, 0) & \text{if } n \text{ is even} \\
\text{diag}(\frac{2}{n}, \ldots, \frac{1}{n}, 0, \ldots, 0) & \text{if } n \text{ is odd}
\end{cases} \] (4.19)

73
It follows from (4.12) that \( \rho < \sigma \). Since \( H(\cdot) \) is a Schur-concave function, then \( H(\rho) \geq H(\sigma) \), which gives the desired conclusion. \( \square \)

Next we will look at elements of \( E_n \) that are of the form \((\frac{1}{k}, \ldots, \frac{1}{k}, 0, \ldots, 0)\).

**Theorem 4.2.4.** Let \((a_j) \in \Omega_{2n}\) satisfy \(a_1 = \ldots = a_k = \frac{1}{k}\) for some \(k\). Then \((a_j) \in E_n\) if and only if

\[
k \in \{n, 2n\} \cup \{\frac{sn}{s+1} | 1 \leq s \leq n-1 \text{ and } (s+1)|n\} \cup \{2n - \frac{sn}{s+1} | 1 \leq s \leq n-1 \text{ and } (s+1)|n\}. \tag{4.20}
\]

**Proof:** Let \(s \in \{1, \ldots, n-1\}\) and \(\frac{(s-1)n}{s} < r < n\). We will show that

\[
\left( \sum_{t=r-s+1}^{s+1} a_t \right) + a_{(s+1)n-sr-s} + \left( \sum_{t=2n-s}^{2n} a_t \right) \leq \frac{s+1}{n} \tag{4.21}
\]

\[
\left( \sum_{t=1}^{s+1} a_t \right) + a_{s(r+1)-(s-1)n} + \left( \sum_{t=2n-r+1}^{2n-r+s} a_t \right) \geq \frac{s+1}{n} \tag{4.22}
\]

Note that the since \(1 \leq s \leq n-1\) and \(\frac{(s-1)n}{s} < r < n\) it follows that \(r < (s+1)n-sr-s < 2n-s\). Thus, we can let \(R = (r-s+1, r-s+2, \ldots, r-1, r, (s+1)n-sr-s, 2n-s, 2n-s+1, \ldots, 2n-1, 2n)\).

Let \(P, Q\) be of the form described in (4.6) with \(k = s+1\) and \(j_t = (s-t+1)r - (s-t)n\) for \(t = 1, \ldots, s+1\). Note that \(j_{s+1} = n\) and for \(t = 1, \ldots, s\) we have \(0 < j_t < j_{t+1}\) because

\[
0 < (s-t+1)(r - \frac{s-1}{s}n) \leq (s-t+1)(r - \frac{s-t}{s+n}n) = j_t = j_{t+1} - (n-r) < j_{t+1}.
\]

Now, \(s(R) = (r-s, \ldots, r-s, s(n-r-2)+n-1, 2n-2s-2, \ldots, 2n-2s-2)\). Define \(A = \text{diag}(s(I))\) and \(B = (W \oplus I_{s+2})\text{diag}(s(J))(W^T \oplus I_{s+2})\), where \(W\) is the \(s \times s\) permutation matrix that switches \(s-t+1\) and \(t\) for all \(t = 1, \ldots, s\). That is,

\[
A = \text{diag}(j_1 - 1, \ldots, j_s - s, j_{s+1} - (s+1), n-s-2, \ldots, n-s-2)
\]

\[
B = \text{diag}(n-j_2 - (s-1), \ldots, n-j_{s+1}, n-j_1 - s, n-s-2, \ldots, n-s-2)
\]
and hence \(\text{eig}(A + B) = s(R)\). By equations (4.9) and (4.10), we get the desired inequalities in (4.21).

To prove the necessity part of the theorem, assume \((a_j) \in \mathcal{E}_n\) satisfies \(a_1 = \cdots = a_k = \frac{1}{k}\). We consider the following two cases.

Case 1: Suppose \(k < n\). Define \(s = \max\{t \mid 1 \leq t \leq n - 1 \text{ such that } \frac{(s-1)n}{s} < k\}\). That is, \(\frac{(s-1)n}{s} < k \leq \frac{sn}{s+1}\). Applying the left side of (4.20) to \(r = k\), we get \(a_{k-s+1} + \cdots + a_k = \frac{s}{k} \leq \frac{s+1}{n}\). Thus \(k \geq \frac{sn}{s+1}\) and hence \(k = \frac{sn}{s+1}\) and consequently, \(s + 1\) must divide \(n\).

Case 2: Suppose \(k = 2n - r\) for some \(0 < r < n\). Define \(s = \max\{t \mid 1 \leq t \leq n - 1 \text{ such that } \frac{(s-1)n}{s} < r\}\). That is, \(\frac{(s-1)n}{s} < r \leq \frac{sn}{s+1}\). Note that \(a_{2n-r+1} = \cdots = a_{2n-r+s} = 0\) and so the right side of (4.20) gives \(\frac{s+2}{2n-r} = \frac{s+2}{k} \geq \frac{s+1}{n}\), which implies \(r \geq \frac{sn}{s+1}\). Hence \(r = \frac{sn}{s+1}\).

Next, we prove the converse. For \(k \in \{n, 2n\}\), consider \(\rho = E_{11} \otimes \frac{1}{n} I_n\) and \(\rho = \frac{1}{2n} I_{2n}\). If \(k = \frac{s}{s+1} n\) for some \(1 \leq s \leq n - 1\) such that \((s+1)|n\), define \(\rho\) as in (4.3) such that

\[
D = \bigoplus_{j=1}^{s+1} \frac{s+1-2}{sn} I_n \quad \text{and} \quad X = \begin{bmatrix} 0 & Y \\ 0 & 0 \end{bmatrix}, \quad \text{where} \quad Y = \bigoplus_{j=1}^{s} \frac{\sqrt{(s+1-j)}}{sn} I_n.
\]

It is easy to verify that \(\text{eig}^4(\rho) = (\frac{1}{k}, \ldots, \frac{1}{k}, 0, \ldots, 0) = (\frac{\frac{sn}{s+1}}{s}, \ldots, \frac{\frac{sn}{s+1}}{s}, 0, \ldots, 0)\). Lastly, if \(k = 2n - \frac{s}{s+1} n\) for some \(1 \leq s \leq n - 1\) such that \((s+1)|n\), define \(\rho\) as in (4.3) such that

\[
D = \bigoplus_{j=1}^{s+1} \frac{j}{(s+2)n} I_n \quad \text{and} \quad X = \begin{bmatrix} 0 & Y \\ 0 & 0 \end{bmatrix}, \quad \text{where} \quad Y = \bigoplus_{j=1}^{s} \frac{\sqrt{(s+1-j)}}{(s+2)n} I_n.
\]

It is easy to verify that \(\text{eig}^4(\rho) = (\frac{1}{k}, \ldots, \frac{1}{k}, 0, \ldots, 0) = (\frac{(s+1)n}{s+2}, \ldots, \frac{(s+1)n}{s+2}, 0, \ldots, 0)\).

\[
\square
\]

**Theorem 4.2.5.** Let \((a_j) \in \mathcal{E}_n\).

a. If \(n = 2k + 1\), then \(a_{3k+1} + a_{3k+2} \leq \frac{1}{n} \leq a_{k+1} + a_{k+2}\).

b. If \(n = 2k\), then \(a_{3k-2} + a_{3k-1} + a_{3k} + a_{3k+1} \leq \frac{1}{k} \leq a_{k} + a_{k+1} + a_{k+2} + a_{k+3}\)

**Proof:** For a, let \(P = (k+1, n+1) = Q, R = (n+k, n+k+1)\) and \(\tilde{A} = \text{diag}(k, n-1)\) and \(\tilde{B} = ...
diag(n-1,k). For b, let \( P = (k,k+1,n+1,n+2) = Q \) and \( R = (n+k-2,n+k-1,n+k,n+k+1) \) and \( \tilde{A} = \text{diag}(k-1,k-1,n-2,n-2) \) and \( \tilde{B} = \text{diag}(n-2,n-2,k-1,k-1) \). Applying the same arguments as the preceding theorems, we get the desired conclusion.

\[ \square \]

### 4.3 Low Dimension Solutions

In this section, we will give the necessary and sufficient conditions for \((a_j) \in E_n\) for \(n = 2, \ldots, 6\). We also give necessary conditions for \((a_j) \in E_7\).

**Theorem 4.3.1.** \( E_2 = \Omega_4 = Co \{ (1,0,0,0), (1\frac{1}{2},1\frac{1}{2},0,0), (1\frac{1}{3},1\frac{1}{3},1\frac{1}{3},0), (1\frac{1}{4},1\frac{1}{4},1\frac{1}{1},1\frac{1}{4}) \} \).

**Proof:** Indeed, for any \(a_1,a_2,a_3,a_4 \geq 0\) with \( \sum_{j=1}^{4} a_j = 1\), the matrix

\[
\begin{bmatrix}
  a_1 + a_2 & 0 & 0 & a_1 - a_2 \\
  0 & a_3 + a_4 & a_3 - a_4 & 0 \\
  0 & a_3 - a_4 & a_3 + a_4 & 0 \\
  a_1 - a_2 & 0 & 0 & a_1 + a_2
\end{bmatrix}
\in S_2(I_2/2)
\]

has eigenvalues \(a_1,a_2,a_3,a_4\). \( \square \)

**Theorem 4.3.2.** Suppose \((a_j) \in \Omega_6\). Then \((a_j) \in E_3\) if and only if

\[
a_4 + a_5 \leq \frac{1}{3} \leq a_2 + a_3.
\]

**Proof:** If \((a_j) \in E_3\), then (4.23) follows from Theorem (4.2.5) a. To prove the converse, assume \((a_i) \in \Omega_6\) satisfies (4.23). Since \((a_i) \in \Omega_6\), the following are true

(a) \(a_1 + a_4 \geq \frac{1}{3}\) \hspace{1cm} (c) \(a_3 + a_6 \leq \frac{1}{3}\)

(b) \(a_1 + a_4 + a_5 \geq \frac{1}{3}\) \hspace{1cm} (d) \(0 \leq a_3 \leq \frac{1}{3}\).
and from (4.23),

(e) $0 \leq a_4 \leq \frac{1}{3}$  \hspace{1cm} (f) $a_1 + a_4 + a_5 \leq \frac{2}{3}$

Define $\rho$ to be of the form (4.3) with

$$D = \text{diag} \left( \frac{1}{3} - a_3, a_1 + a_4 + a_5 - \frac{1}{3}, a_4 \right)$$

and

$$X = \begin{bmatrix}
0 & \sqrt{(a_2 + a_3 - \frac{1}{3})(\frac{1}{3} - a_3 - a_6)} & 0 \\
0 & 0 & \sqrt{(a_1 + a_4 - \frac{1}{3})(\frac{1}{3} - a_4 - a_5)} \\
0 & 0 & 0
\end{bmatrix}.$$ 

Inequalities (b), (e), (d), and (f) guarantee that $0 \leq D \leq \frac{1}{3}I_3$ and inequalities (a), (c), together with (4.23) guarantee that $X$ is well-defined and hence $eig^i(A) = (a_1, \ldots, a_6)$. \hfill \blackslug

Note that $E_3$ is the convex hull of the following extreme elements

$$\left( \frac{1}{2}, \frac{1}{2}, 0, 0, 0, 0 \right), \left( \frac{2}{3}, \frac{1}{3}, 0, 0, 0, 0 \right), \left( \frac{2}{5}, \frac{3}{5}, 0, 0, 0, 0 \right), \left( \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, 0, 0 \right),$$

$$\left( \frac{1}{2}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, 0, 0 \right), \left( \frac{1}{2}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, 0, 0 \right), \left( \frac{1}{3}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, 0, 0 \right),$$

$$\left( \frac{1}{2}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, 0, 0 \right), \left( \frac{1}{3}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, 0, 0 \right), \left( \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, 0 \right).$$

**Theorem 4.3.3.** Suppose $(a_j) \in \Omega_8$. Then $(a_j) \in E_4$ if and only if

$$a_4 + a_5 + a_6 + a_7 \leq \frac{1}{2} \leq a_2 + a_3 + a_4 + a_5 \quad (4.24)$$

**Proof:** It follows from Theorem 4.2.5 that if $(a_j) \in E_4$, then the inequalities in (4.24) holds.

To prove the converse, assume $(a_i) \in \Omega_8$ satisfies (4.24). The following inequalities hold since $(a_j) \in \Omega_8$.

(a) $a_4 + a_8 \leq \frac{1}{4} \leq a_1 + a_5$  \hspace{1cm} (b) $a_2 + a_4 + a_6 + a_8 \leq \frac{1}{2}$

The following inequalities can be obtained from (4.24)
(c) \(a_6 + a_7 \leq \frac{1}{4} \leq a_2 + a_3\) \hspace{1cm} (d) \(a_5 + a_7 \leq \frac{1}{4} \leq a_2 + a_4\) \hspace{1cm} (e) \(a_1 + a_6 + a_7 + a_8 \leq \frac{1}{2}\)

We will construct a matrix \(\rho\) of the form (4.3), where

\[
D = \text{diag}(x_1, x_2, x_3, x_4), \quad X = \begin{bmatrix} 0 & Y \\ 0 & 0 \end{bmatrix}, \quad Y = \text{diag}(y_1, y_2, y_3, y_4)
\]

for some \(0 \leq x_i \leq \frac{1}{4}\) and \(y_i \in \mathbb{R}^+\). We will choose the \(x_j\)'s such that

\[
x_1 = 1/4 - a_{j_8} \quad \quad \quad \quad \quad \quad \quad x_4 = a_{j_7} \\
x_1 + 1/4 - x_2 = a_{j_1} + a_{j_2} \quad \quad x_1(1/4 - x_2) - y_1^2 = a_{j_1}a_{j_2} \\
x_2 + 1/4 - x_3 = a_{j_3} + a_{j_4} \quad \quad x_2(1/4 - x_3) - y_2^2 = a_{j_2}a_{j_4} \\
x_3 + 1/4 - x_4 = a_{j_5} + a_{j_6} \quad \quad x_3(1/4 - x_4) - y_3^2 = a_{j_5}a_{j_6}
\]

for some choice of indices \(j_1, \ldots, j_8 \in \{1, \ldots, 8\}\). More explicitly,

\[
x_1 = 1/4 - a_{j_8}, \quad x_2 = 1/2 - (a_{j_8} + a_{j_1} + a_{j_2}), \quad x_3 = a_{j_5} + a_{j_6} + a_{j_7} - 1/4, \quad x_4 = a_{j_7}
\]

and

\[
y_1 = \sqrt{(a_{j_1} + a_{j_6} - 1/4)(1/4 - a_{j_6} - a_{j_2})} \\
y_2 = \sqrt{(a_{j_1} + a_{j_2} + a_{j_3} + a_{j_6} - 1/2)(1/2 - a_{j_1} - a_{j_2} - a_{j_4} - a_{j_6})} \\
y_3 = \sqrt{(a_{j_5} + a_{j_7} - 1/4)(1/4 - a_{j_6} - a_{j_7})}
\]

We can assume without loss of generality that \(j_1 < j_2, j_3 < j_4, j_5 < j_6\), i.e. \(a_{j_1} \geq a_{j_2}\) and so on. To ensure that \(0 < x_j \leq 1/4\), the following must be true:

\[
a_{j_7}, a_{j_8} \leq 1/4 \quad \quad 1/4 \leq a_{j_8} + a_{j_1} + a_{j_2} \leq 1/2 \quad \quad 1/4 \leq a_{j_5} + a_{j_6} + a_{j_7} \leq 1/2
\]
And to ensure that \( y_j \) exists for \( j = 1, 2, 3 \), the following inequalities must be true:

\[
\begin{align*}
    a_{j_6} + a_{j_7} &\leq \frac{1}{4} \leq a_{j_5} + a_{j_7} \quad (4.25) \\
    a_{j_8} + a_{j_2} &\leq \frac{1}{4} \leq a_{j_8} + a_{j_1} \quad (4.26) \\
    a_{j_8} + a_{j_1} + a_{j_2} + a_{j_4} &\leq \frac{1}{2} \leq a_{j_8} + a_{j_1} + a_{j_2} + a_{j_3} \quad (4.27)
\end{align*}
\]

Note that inequalities (4.25)-(4.27) imply the previous three inequalities. We will consider three cases:

**Case 1:** Suppose \( a_2 + a_3 + a_4 + a_8 \geq \frac{1}{2} \). Choose 

\[
j_1 = 2, j_2 = 8, j_3 = 3, j_4 = 6, j_5 = 1, j_6 = 7, j_7 = 5, j_8 = 4
\]

Inequality (4.25) is guaranteed by (a) and (d), while (4.26) follows from (c) and the assumption in this case. Lastly, (4.26) is implied by (b) and the assumption in this case.

**Case 2:** Suppose \( a_2 + a_3 + a_4 + a_8 < \frac{1}{2} \). Then

\[
(f) \quad a_3 + a_8 < \frac{1}{4} < a_1 + a_6 \quad (g) \quad a_1 + a_5 + a_6 + a_7 > \frac{1}{2}
\]

**Case 2.1:** Suppose \( a_4 + a_5 \leq \frac{1}{4} \). Choose 

\[
j_1 = 1, j_2 = 7, j_3 = 3, j_4 = 8, j_5 = 2, j_6 = 5, j_7 = 4, j_8 = 6
\]

Inequality (4.25) follows from (d) and the additional assumption, while (4.26) follows from (f) and (c) and (4.27) follows from (g) and (e).

**Case 2.2:** Suppose \( a_2 + a_3 + a_4 + a_8 < \frac{1}{2} \) and \( a_4 + a_5 > \frac{1}{4} \). Choose 

\[
j_1 = 4, j_2 = 7, j_3 = 1, j_4 = 6, j_5 = 2, j_6 = 8, j_7 = 3, j_8 = 5
\]

Inequality (4.25) follows from (c) and the assumption that \( a_2 + a_3 + a_4 + a_8 < \frac{1}{2} \), while (4.26)
follows from the assumption in this case and (d). Lastly, (4.27) is guaranteed by (g) and (4.24).

In all cases \( \rho \in S_2(\frac{1}{n}I_n) \) and \( \text{eig}^i(\rho) = (a_j) \).

By the Krein-Milman Theorem, we can apply the above argument to \( Q \) and suppose \( \tilde{Q} \) is another compact convex polytope described by a finite subset of these inequalities, say \( Ax \leq \tilde{b} \) for some \( k \times m \) matrix \( A \). Clearly, \( Q \subseteq \tilde{Q} \). Now consider the set of extreme points of \( \tilde{Q} \), that is,

\[
\tilde{Q}_{\text{ext}} = \{ x = (P_V \tilde{A})^{-1}P_V \tilde{b} \mid \text{for some projection } P_V \text{ such that } (P_V \tilde{A})^{-1} \text{ exists and } Ax \leq \tilde{b} \}
\]

By the Krein-Milman Theorem, \( \text{Co}(\tilde{Q}_{\text{ext}}) = \tilde{Q} \supseteq Q \). If \( Ax \leq b \) for all \( x \in Q_{\text{ext}} \), then \( \tilde{Q} = Q \).

We can apply the above argument to \( Q = \mathcal{E}_n \). Suppose that a necessary condition for \( \nu \in \mathcal{E}_n \) is given by \( \nu A \leq b \). Then the set \( \mathcal{E}_n \subseteq \{ \nu \in \Omega_{2n} \mid \nu A \leq b \} = \text{Co}(v_1, \ldots, v_s) \). If \( v_1, \ldots, v_s \in \mathcal{E}_n \), then \( \mathcal{E}_n = \{ \nu \in \Omega_{2n} \mid \nu A \leq b \} \) since \( \mathcal{E}_n \) is also a convex set. We will use this idea to determine the necessary and sufficient conditions for \( (a_i) \in \mathcal{E}_n \) for \( n = 5, 6 \).
Theorem 4.3.4. Suppose \((a_i) \in \Omega_{10}\). Then \((a_i) \in \mathcal{E}_5\) if and only if

\[
a_7 + a_8 \leq \frac{1}{5} \leq a_4 + a_4
\]

(4.28)

\[
a_1 + a_8 + a_9 + a_{10} \leq \frac{2}{5} \leq a_1 + a_2 + a_3 + a_{10}
\]

(4.29)

\[
a_5 + a_6 + a_7 + a_{10} \leq \frac{2}{5} \leq a_1 + a_4 + a_5 + a_6
\]

(4.30)

\[
a_4 + a_7 + a_8 + a_9 \leq \frac{2}{5} \leq a_2 + a_3 + a_4 + a_7
\]

(4.31)

Proof: If \((a_i) \in \mathcal{E}_5\), then (4.28) follows from theorem 4.2.5 and (4.29) follows from (4.11). To see that (4.30) and (4.31) hold, let \(P = (2, 4, 6, 7) = Q, R_1 = (5, 6, 7, 10)\) and \(R_2 = (4, 7, 8, 9)\). Define \(\tilde{A}_1 = \text{diag}(1, 2, 3, 3), \tilde{B}_1 = \text{diag}(3, 2, 1, 3), A_2 = \text{diag}(3, 1, 2, 3)\) and \(B_2 = \begin{bmatrix} 3/2 & \sqrt{3}/2 \\ \sqrt{3}/2 & 5/2 \end{bmatrix} \oplus \text{diag}(3, 2)\). Then \(\text{eig}^+ (\tilde{A}_1 + \tilde{B}_1) = s(R_1)\) and Then \(\text{eig}^+ (\tilde{A}_2 + \tilde{B}_2) = s(R_2)\).

Using the Matlab script \texttt{n5EXT.m}, which can be found in Appendix A.4, we are able to list the extreme points of \(\{(a_j) \in \Omega_{10} \mid (a_j) \text{ satisfies } (4.28), (4.29), (4.30), (4.31)\}\). Each of these extreme points are in \(\mathcal{E}_{10}\). In fact, for each extreme point listed above, one can form \(\rho \in S_2(\frac{1}{n}I_n)\) with the prescribed eigenvalues such that \(\rho\) is permutationally similar to a direct sum of \(2 \times 2\) matrices. The Matlab script \texttt{Findnicesol.m} (see Appendix A.4) can be use to find such a simple solution for any of the 50 extreme points listed in the Appendix B. \qed

Theorem 4.3.5. Suppose \((a_j) \in \Omega_{12}\). Then \((a_j) \in \mathcal{E}_6\) if and only if

\[
a_1 + a_{10} + a_{11} + a_{12} \leq \frac{1}{3} \leq a_1 + a_2 + a_3 + a_{12}
\]

(4.32)

\[
a_4 + a_9 + a_{10} + a_{11} \leq \frac{1}{3} \leq a_2 + a_3 + a_4 + a_9
\]

(4.33)

\[
a_7 + a_8 + a_9 + a_{10} \leq \frac{1}{3} \leq a_3 + a_4 + a_5 + a_6
\]

(4.34)

\[
a_1 + a_6 + a_8 + a_{10} + a_{11} + a_{12} \leq \frac{1}{2} \leq a_1 + a_2 + a_3 + a_5 + a_7 + a_{12}
\]

(4.35)

Proof: If \((a_j) \in \mathcal{E}_{12}\), then inequality (4.32) and (4.34) holds from (4.11) and and theorem 4.2.5 b. We get the inequality (4.35) by letting \(R = (1, 6, 8, 10, 11), P = (1, 3, 6, 7, 8, 9), Q = 81\).
(1, 4, 6, 7, 8, 9), \( \tilde{A} = \text{diag}(0, 1, 3, 3, 3, 3) \) and \( \tilde{B} = \text{diag}(0, 3, 2, 3, 3, 3) \). Finally, we get the inequality (4.33) by letting \( R = (4, 9, 10, 11) \), \( P = (2, 5, 7, 8) = Q \), \( \tilde{A} = \text{diag}(4, 1, 3, 4) \) and

\[
\tilde{B} = \begin{bmatrix}
\frac{7+\sqrt{11}}{4} & \frac{\sqrt{6\sqrt{11}-14}}{4} \\
\frac{\sqrt{6\sqrt{11}-14}}{4} & \frac{13-\sqrt{11}}{4}
\end{bmatrix} \oplus \text{diag}(4, 3).
\]

Now, to prove the converse, we find the extreme points of the set

\[ Q = \{ (a_j) \in \Omega_{12} \mid (a_i) \text{ satisfies (4.32) - (4.35)} \} \]

using the Matlab script n6EXT.m. There are 48 extreme points which are listed in the Appendix.

Using the same method for \( n = 5 \), it can be shown that each of these extreme points are in \( \mathcal{E}_{12} \) and thus \( Q = \mathcal{E}_{12} \).

\[ \square \]

**Theorem 4.3.6.** Let \((a_i) \in \Omega_n\). If \((a_i) \in \mathcal{E}_7\), then

\[
a_{10} + a_{11} \leq \frac{1}{7} \leq a_4 + a_5 \tag{4.36}
\]

\[
a_1 + a_{12} + a_{13} + a_{14} \leq \frac{2}{7} \leq a_1 + a_2 + a_3 + a_{14} \tag{4.37}
\]

\[
a_4 + a_{11} + a_{12} + a_{13} \leq \frac{2}{7} \leq a_2 + a_3 + a_4 + a_{11} \tag{4.38}
\]

\[
a_8 + a_9 + a_{10} + a_{13} \leq \frac{2}{7} \leq a_2 + a_5 + a_6 + a_7 \tag{4.39}
\]

\[
a_7 + a_{10} + a_{11} + a_{12} \leq \frac{2}{7} \leq a_3 + a_4 + a_5 + a_8 \tag{4.40}
\]

\[
a_1 + a_6 + a_{11} + a_{12} + a_{13} + a_{14} \leq \frac{3}{7} \leq a_1 + a_2 + a_3 + a_4 + a_9 + a_{14} \tag{4.41}
\]

\[
a_3 + a_4 + a_{11} + a_{12} + a_{13} + a_{14} \leq \frac{3}{7} \leq a_1 + a_2 + a_3 + a_4 + a_{11} + a_{12} \tag{4.42}
\]

\[
a_1 + a_7 + a_{10} + a_{12} + a_{13} + a_{14} \leq \frac{3}{7} \leq a_1 + a_2 + a_3 + a_5 + a_8 + a_{14} \tag{4.43}
\]

\[
a_1 + a_8 + a_9 + a_{12} + a_{13} + a_{14} \leq \frac{3}{7} \leq a_1 + a_2 + a_3 + a_6 + a_7 + a_{14} \tag{4.44}
\]

\[
a_6 + a_7 + a_8 + a_9 + a_{13} + a_{14} \leq \frac{3}{7} \leq a_1 + a_2 + a_6 + a_7 + a_8 + a_9 \tag{4.45}
\]

\[
a_5 + a_8 + a_9 + a_{10} + a_{11} + a_{14} \leq \frac{3}{7} \leq a_1 + a_4 + a_5 + a_6 + a_7 + a_{10} \tag{4.46}
\]
a_6 + a_7 + a_9 + a_{10} + a_{11} + a_{14} \leq \frac{3}{7} \leq a_1 + a_4 + a_5 + a_6 + a_8 + a_9 \quad (4.47)
a_4 + a_7 + a_{10} + a_{11} + a_{12} + a_{13} \leq \frac{3}{7} \leq a_2 + a_3 + a_4 + a_5 + a_8 + a_{11} \quad (4.48)
a_5 + a_6 + a_{10} + a_{11} + a_{12} + a_{13} \leq \frac{3}{7} \leq a_2 + a_3 + a_4 + a_5 + a_9 + a_{10} \quad (4.49)
a_4 + a_8 + a_9 + a_{11} + a_{12} + a_{13} \leq \frac{3}{7} \leq a_2 + a_3 + a_4 + a_6 + a_7 + a_{11} \quad (4.50)
a_5 + a_7 + a_9 + a_{11} + a_{12} + a_{13} \leq \frac{3}{7} \leq a_2 + a_3 + a_4 + a_6 + a_8 + a_{10} \quad (4.51)
a_5 + a_8 + a_9 + a_{10} + a_{12} + a_{13} \leq \frac{3}{7} \leq a_2 + a_3 + a_5 + a_6 + a_7 + a_{10} \quad (4.52)
a_6 + a_7 + a_9 + a_{10} + a_{12} + a_{13} \leq \frac{3}{7} \leq a_2 + a_3 + a_5 + a_6 + a_8 + a_9 \quad (4.53)
a_6 + a_8 + a_9 + a_{10} + a_{11} + a_{13} \leq \frac{3}{7} \leq a_2 + a_4 + a_5 + a_6 + a_7 + a_9 \quad (4.54)
a_7 + a_8 + a_9 + a_{10} + a_{11} + a_{12} \leq \frac{3}{7} \leq a_3 + a_4 + a_5 + a_6 + a_7 + a_8 \quad (4.55)

**Proof:** Let \((a_j) \in \Omega_{14}\). The first inequality follows from Theorem 4.2.5 while the second is given by inequality (4.11).

For a given triple \(P, Q, R\) of \(k\)-subsets of \(\{1, \ldots, 2n\}\), the existence of Hermitian matrices \(A, B, A+B\) satisfying \(eig^t(A) = s(P), eig^t(B) = s(Q)\) and \(eig^t(A+B) = s(R)\) can be formulated in terms of Young tableaux. Specifically, such \(A, B\) exist if and only if there is a Littlewood-Richardson (LR) skew-tableaux of shape \(\tilde{s}(R)/\tilde{s}(P)\) and content \(\tilde{s}(Q)\), where \(\tilde{s}(\cdot)\) is \(s(\cdot)\) rearranged in nonincreasing order [48, 41]. In Figures 4.1 and 4.2, we provide skew-tableaux that will guarantee the necessity of inequalities (4.38)-(4.55).

\[\square\]

### 4.4 Further Remarks

In this chapter we looked at the possible eigenvalues of an element a bipartite quantum state \(\rho\) such that \(tr_1(\rho) = \frac{1}{n} I_n\). This is a special case of the quantum marginal problem which is a very difficult to solve in its general form. In chapter 5.3 we state the quantum marginal problem (see Problem 5.3.1) and use the alternating projection method to find a solution.

We were only able to completely describe \(S_2(\frac{1}{n} I_n)\) for \(n \leq 6\) and partially for \(n \leq 7\). Note
that if one of the necessary inequalities we have listed in Theorems 4.23-4.3.5 are removed, the list will not provide a sufficient condition any longer. The conditions given in Proposition 4.1.1 b and c depend on a certain choice of \( d_1, \ldots, d_n \in [0, 1] \) satisfying certain inequalities. However, the results in this chapter 4.3 leads us to the following conjecture.

**Conjecture 4.4.1.** An element \((a_1, \ldots, a_{2n})\) of \(\Omega_{2n}\) is in \(E_n\) if and only if

\[
\sum_{r=1}^{2k} a_r \leq \frac{k}{n} \sum_{r=1}^{2k} a_{2n-r+1}
\]

for any \(1 \leq k \leq n\) and any \(R \subseteq \{1, \ldots, 2n\}\) such that \(|R| = 2k\) and \((P, Q, R) \in LR_{2k}(2n)\) for some \(P, Q\) of the form given in equation (4.6).

If the above conjecture is true, then \(E_n\) is always described by a finite set of inequalities, making it a convex polytope. An observation that the author made is that for \(n \leq 6\), any necessary inequality written in the form \(a_{r_1} + \cdots + a_{r_{2k}} \leq \frac{k}{n}\), satisfies

1. \(1 \leq r_1 < r_2 < \cdots < r_{2k} \leq 2n\)

2. \(\sum_{s=1}^{2k} r_s = k(3n - k + 1)\)
and \((r_1, \ldots, r_{2k})\) can in fact be obtained from \((n - r + 1, \ldots, n, 2n - r + 1, \ldots, 2n)\) by a sequence of pinchings. By a pinching, we mean adding 1 to an index and subtracting 1 to another index when there is room to do so.

For general \(n\) we described the possible eigenvalues of an element of \(S_2 \left( \frac{1}{n} I_n \right) \) that is of minimal rank. As a consequence, we were able to give a lower bound for the von Neumann entropy of any \(\rho \in S_2 \left( \frac{1}{n} I_n \right) \). Moreover, if \(\rho\) is an element of \(S_2 \left( \frac{1}{n} I_n \right) \), its entropy \(H(\rho)\) is also defined to be the entropy of the channel \(\Phi : \mathcal{H}_n \rightarrow \mathcal{H}_2\) whose Choi matrix is permutationally similar to \(n\rho\). For future research, one may consider the problem of finding the minimum entropy of \(H(\rho)\) for a different set of quantum channels.
FIG. 4.2: LR skew-tableaux of shape $\tilde{s}(R)/\tilde{s}(P)$ and content $\tilde{s}(Q)$ for inequalities (4.44)-(4.55).
CHAPTER 5

Projection Methods

In this chapter, we utilize projection methods to solve feasibility problems of the form

\[
\text{find: } x \in S_1 \cap S_2
\]

that arise in the context of quantum information theory and matrix theory.

5.1 Introduction

We begin by describing the method of alternating projections (MAP) and the Douglas-Rachford method (DR) in full generality. To this end, consider a Euclidean space \( \mathcal{E} \) with an inner product \( \langle \cdot, \cdot \rangle \) and norm \( \| \cdot \| \). We are interested in finding a point \( x \) lying in the intersection of two closed subsets \( S_1 \) and \( S_2 \) of \( \mathcal{E} \). Projection based methods then presuppose that given a point \( x \in \mathcal{E} \), finding a point in the nearest-point set

\[
\text{proj}_{S_1}(x) = \arg\min_{a \in S_1} \{ \| x - a \| \} \equiv \{ \hat{a} \in S_1 \mid \| x - \hat{a} \| = \min_{a \in S_1} \| x - a \| \}
\]

(5.1)
is easy, as is finding a point in \( \text{proj}_{S_2}(x) \). When \( S_1 \) and \( S_2 \) are convex, the nearest-point sets \( \text{proj}_{S_1}(x) \) and \( \text{proj}_{S_2}(x) \) are singletons, of course.

Given a current point \( a_l \in S_1 \), the method of alternating projections then iterates the
following two steps

\[
\begin{align*}
\text{choose } & \ b_l \in \text{proj}_{S_2}(a_l) \\
\text{choose } & \ a_{l+1} \in \text{proj}_{S_1}(b_l)
\end{align*}
\]

When \( S_1 \) and \( S_2 \) are convex and there exists a pair of nearest points of \( S_1 \) and \( S_2 \), the method always generates iterates converging to such a pair. In particular, when the convex sets \( S_1 \) and \( S_2 \) intersect, the method converges to some point in the intersection \( S_1 \cap S_2 \). Moreover, when the relative interiors of \( S_1 \) and \( S_2 \) intersect, convergence is R-linear with the rate governed by the cosines of the angles between the vectors \( a_{l+1} - b_l \) and \( a_l - b_l \). For details, see for example [36, 4, 5, 15]. When \( S_1 \) and \( S_2 \) are not convex, analogous convergence guarantees hold, but only if the method is initialized sufficiently close to the intersection [59, 60, 7, 29].

The Douglas-Rachford algorithm takes a more asymmetric approach. Given a point \( x \in \mathcal{E} \), we define the reflection operator

\[
\text{refl}_{S_1}(x) = \text{proj}_{S_1}(x) + (\text{proj}_{S_1}(x) - x).
\]

The Douglas-Rachford algorithm is then a “reflect-reflect-average” method; that is, given a current iterate \( x_l \in \mathcal{E} \), it generates the next iterate by the formula

\[
x_{l+1} = \frac{x_l + \text{refl}_{S_1}(\text{refl}_{S_2}(x_l))}{2}.
\]

It is known that for convex instances, the “projected iterates” converge [74]. The rate of convergence, however, is not well-understood. On the other hand, the method has proven to be extremely effective empirically for many types of problems; see for example [2, 35, 6].

The salient point here is that for MAP and DR to be effective in practice, the nearest point mappings \( \text{proj}_{S_1} \) and \( \text{proj}_{S_2} \) must be easy to evaluate. For example, when \( S_1 \) is the convex cone \( PSD_N \), we may use the following classical result in matrix theory [34] to find \( \text{proj}_{PSD_N}(X) \).
Theorem 5.1.1. Suppose \( X = U \text{diag}(d_1, \ldots, d_N)U^* \in \mathcal{H}_N \), where \( U \in \mathcal{U}_N \). Define \( c_j = \max\{0, d_j\} \). Then \( \text{proj}_{\text{PSD}_N}(X) = U \text{diag}(c_1, \ldots, c_N)U^* \). That is, for any \( Z \in \text{PSD}_N \) and any unitary similarity invariant norm \( || \cdot || \), \( ||X - U \text{diag}(c_1, \ldots, c_N)U^*|| \leq ||X - Z|| \).

We will also consider an affine space of the form \( \mathcal{S}_2 = \{X \in \mathcal{H}_N \mid \mathcal{L}(X) = B\} \) for some linear map \( \mathcal{L} \) and some Hermitian matrix \( B \). In this case, a classical result in linear algebra gives

\[
\text{proj}_{\mathcal{S}_2}(X) = X + \mathcal{L}^\dagger(B - \mathcal{L}(X)),
\]

(5.2)

where \( \mathcal{L}^\dagger \) denotes the Moore-Penrose inverse generalized inverse of \( \mathcal{L} \).

In Section 5.2, we consider the problem of finding a quantum channel \( \Phi \) such that for given sets \( \{\rho^{(1)}, \ldots, \rho^{(k)}\} \subset \mathcal{D}_n \) and \( \{\sigma^{(1)}, \ldots, \sigma^{(k)}\} \subset \mathcal{D}_m \), the channel \( \Phi \) maps the state \( \rho^{(j)} \) to \( \sigma^{(j)} \) for \( j = 1, \ldots k \). Using the Choi representation of quantum channels, we know that this problem is equivalent to \( P \in \mathcal{S}_1 \cap \mathcal{S}_2 \), where \( \mathcal{S}_1 = \text{PSD}_{mn} \) and

\[
\mathcal{S}_2 = \{(P_{st})^n_{s,t=1} \mid P_{st} \in \mathbb{C}^{m \times m} \text{ with } \text{tr}(P_{st}) = \delta_{st} \text{ and } \sum_{k,l=1}^n \rho_{kl}^{(j)} P_{kl} = \sigma^{(j)} \text{ for } j = 1, \ldots, k \}
\]

where \( \delta_{st} = 1 \) when \( s = t \) and \( \delta_{st} = 0 \) otherwise. Note that \( \mathcal{S}_1 \) is a convex cone and \( \mathcal{S}_2 \) is an affine, and therefore a convex, space. Moreover, \( \mathcal{S}_1 \) and \( \mathcal{S}_2 \) are subsets of the set of \( mn \times mn \) hermitian matrices \( \mathcal{H}_n \), which is an \( (mn)^2 \)-dimensional real linear space. Thus, projections \( \text{proj}_{\mathcal{S}_1}(P) \) will be a unique element of \( \mathcal{S}_1 \) and \( \text{proj}_{\mathcal{S}_2}(P) \), will be a unique element of \( \mathcal{S}_2 \). We will illustrate the effectiveness of the MAP and DR algorithms in solving this problem.

In Section 5.3, we consider the problem of finding a global state \( \rho \) of a multipartite system \( X = (X_1, \ldots, X_k) \) having prescribed reduced state \( \rho_{J_s} = \text{tr}_{J_s}(\rho) \) on subsystem \( X_{J_s} = (X_j)_{j \in J_s} \) for \( s = 1, \ldots, r \) and \( J_s \subset \{1, \ldots, k\} \). We can view a solution \( \rho \) to this problem as an element of \( \mathcal{S}_1 \cap \mathcal{S}_2 \), where \( \mathcal{S}_1 = \text{PSD}_N \), where \( N = n_1 \ldots n_k \), and

\[
\mathcal{S}_2 = \{P \in \mathcal{H}_N \mid \text{tr}_{J_s}(P) = \rho_{J_s} \text{ for all } s = 1, \ldots, r\}
\]
Here $S_2$ is an affine subspace of $H_N$. We will consider the same problem with additional required properties for the solution $\rho$, such as prescribed eigenvalues, low rank or low entropy. In Section 5.4, we present algorithms to find low rank solutions and in Section 5.5, we will suggest a possible projection-based algorithm that can find solutions of low entropy.

In Section 5.6, we consider a problem of interest in matrix theory. Given a matrix $A \in \mathbb{C}^{n \times n}$, determine an easily-verifiable condition for $A$ to be a product of two positive contractions, that is $A = P_1 P_2$ such that $0_n \leq P_1, P_2 \leq I_n$. Equivalently $A = P_1 P_2$ such that $P_1, P_2, I_n - P_1, I_n - P_2 \in PSD_n$. A necessary and sufficient condition of the form $S_1(A) \cap S_2(A) \cap S_3(A) \neq \emptyset$, where $S_i(A)$ are convex sets whose descriptions depend on $A$, will be given and projection methods will be used to demonstrate this condition.

### 5.2 Quantum Channel Construction*

A basic problem in quantum information science is to construct, if it exists, a quantum channel sending a given set of quantum states $\{\rho^{(1)}, \ldots, \rho^{(k)}\} \subseteq D_n$ to another set of quantum states $\{\sigma^{(1)}, \ldots, \sigma^{(k)}\} \subseteq D_m$; see e.g., [52, 43, 67, 68, 18, 79] and the references therein. Using the Choi representation of completely-positive maps discussed in Section 1.4, we know that a map $\Phi : \mathbb{C}^{n \times n} \rightarrow \mathbb{C}^{m \times m}$ is a quantum channel if and only if the $mn \times mn$ matrix

$$
\mathcal{C}(\Phi) := \begin{bmatrix}
P_{11} & \ldots & P_{1n} \\
\vdots & \ddots & \vdots \\
P_{11} & \ldots & P_{nn}
\end{bmatrix} := \begin{bmatrix}
\Phi(E_{11}) & \ldots & \Phi(E_{1n}) \\
\vdots & \ddots & \vdots \\
\Phi(E_{11}) & \ldots & \Phi(E_{nn})
\end{bmatrix}, \quad (5.3)
$$

is positive semidefinite and $\text{tr}(P_{st}) = \delta_{st}$ for $1 \leq s,t \leq n$. Hence, the existence of a quantum channel $\Phi$ satisfying $\Phi(\rho^{(j)}) = \sigma^{(j)}$ is equivalent to the positive semidefinite feasibility problem

*The material in this section is contained in the paper [30], which is a joint work of Y.-L. Cheung and D. Drusvyatskiy, C.-K. Li, H. Wolkowicz and the author.
of finding $P = [P_{st}]_{s,t=1}^n$, where $P_{st} \in \mathbb{C}^{m \times m}$ such that

$\begin{align*}
\sum_{ij} \rho^{(j)}_{st} P_{st} &= \sigma^{(j)}, \quad j = 1, \ldots, k \\
\text{tr}(P_{st}) &= \delta_{st}, \quad 1 \leq s \leq t \leq n \\
P &\in \text{PSD}_{nm}
\end{align*}$

Moreover, the rank of the Choi matrix $P$ has a natural interpretation: it is equal to the minimal number of summands needed in any Kraus representation of $\Phi$. That is, if $\text{rank}(C(\Phi)) = r$, then there exists $F_1, \ldots, F_r \in \mathbb{C}^{m \times n}$ such that

$\Phi(X) = \sum_{j=1}^r F_j X F_j^\dagger$ for all $X \in \mathbb{C}^{n \times n}$.  

Because of the trace preserving constraints, the solution set of (5.4) is bounded. Thus, the problem is never weakly infeasible, i.e., infeasible but contains an asymptotically feasible sequence, e.g., [33]. In particular, one can use standard primal-dual interior point semidefinite programming packages to solve the feasibility problem. However, when the size of the problem $(n, m)$ grows, the efficiency and especially the accuracy of the semidefinite programming approach is limited. To illustrate, even for a reasonable sized problem $m = n = 100$, the number of complex variables involved is $10^8/2$. In this paper, we exploit the special structure of the problem and develop projection based methods to solve high dimensional problems with high accuracy. We present numerical experiments based on the alternating projection (MAP) and the Douglas-Rachford (DR) projection/reflection methods. We see that the DR method significantly outperforms MAP for this problem. Our numerical results show promise of projection based approaches for many other types of feasibility problems arising in quantum information science.

5.2.1 Projection Operators

In the current work, regard the space of Hermitian matrices $\mathcal{H}_{nm}$ as a Euclidean space, that is, an inner product space over $\mathbb{R}$. As usual, we then endow $\mathcal{H}_{nm}$ with the Frobenius norm
\[ \|X\| = \sum_{p,q=1}^{nm} (\text{Re} X_{pq})^2 + (\text{Im} X_{pq})^2, \]

where \( X_{pq} = \text{Re} X_{pq} + i\text{Im} X_{pq} \) is the \((p, q)\) entry of \( X \).

Recall that our basic problem is to find a Hermitian block matrix \( P = [P_{st}]_{s,t=1}^n \), where \( P_{st} \in \mathbb{C}^{m \times m} \), satisfying (5.4). We aim to apply MAP and DR to this formulation. To this end, we first need to introduce some notation to help with the exposition. Define the linear mappings

\[
\mathcal{L}_1(P) := \left( \sum_{st} \rho_{st}^{(j)} P_{st} \right)_{j=1}^k \quad \text{and} \quad \mathcal{L}_2(P) = \left( \text{tr}(P_{st}) \right)_{1 \leq s \leq t \leq n},
\]

and let

\[
\mathcal{L}(P) = (\mathcal{L}_1(P), \mathcal{L}_2(P)).
\] (5.6)

Moreover assemble the vectors

\[
\sigma = (\sigma^{(1)}, \ldots, \sigma^{(k)}) \quad \text{and} \quad \Delta = (\delta_{st})_{1 \leq s \leq j \leq n}.
\] (5.7)

Thus, we aim to find a matrix \( P \) in the intersection of \( PSD_{nm} \) with the affine subspace

\[
\mathcal{S} := \{ P : \mathcal{L}(P) = (\sigma, \Delta) \}.
\] (5.8)

Projecting a Hermitian matrix \( P \) onto \( PSD_{nm} \) is standard due to the Eckart-Young Theorem 5.1.1. Note that projecting a Hermitian matrix onto \( PSD_{nm} \) requires a single eigenvalue decomposition — a procedure for which there are many efficient and well-tested codes (e.g., [25]).

Next, we need to find the projection of \( X \) onto the affine subspace \( \mathcal{S} \), that is how to solve the nearest point problem

\[
\min \left\{ \frac{1}{2} \|P - \hat{P}\|^2 : \mathcal{L}(\hat{P}) = (\sigma, \Delta) \right\}.
\] (5.9)

Classically, the solution is

\[
\text{proj}_{\mathcal{S}}(P) = P + \mathcal{L}^\dagger R,
\] (5.10)

where \( \mathcal{L}^\dagger \) is the Moore-Penrose generalized inverse of \( \mathcal{L} \) and \( R := (\sigma, \Delta) - \mathcal{L}(P) \) is the residual.
Finding the Moore-Penrose generalized inverse of a large linear mapping, like the one we have here, can often be time consuming and error prone. Luckily, the special structure of the affine constraints in our problem allow us to find $L^\dagger$ both very quickly and very accurately, so that in all our experiments the time to compute the projection onto $S$ is negligible compared to the computational effort needed to perform the eigenvalue decompositions. We now describe how to compute $L^\dagger$ in more detail.

For a fixed positive integer $\ell$ and $0 \leq p, q \leq \ell - 1$, define

$$\hat{E}_{p+1, q+1}^\ell = \begin{cases} 
\frac{1}{\sqrt{2}}(|p\rangle\langle q| + |q\rangle\langle p|) & \text{if } p < q, \\
\frac{1}{\sqrt{2}}(|q\rangle\langle p| - |p\rangle\langle q|) & \text{if } p > q, \\
|q\rangle\langle q| & \text{if } p = q.
\end{cases} \quad (5.11)$$

where $|q\rangle$ is the $(q+1)^{th}$ standard basis vector for $\mathbb{R}^n$. Then $E_{\text{real,offdiag}} \cup E_{\text{imag,offdiag}} \cup E_{\text{diag}}$ forms an orthonormal basis of $\mathcal{H}_\ell$, where

- $E_{\text{real,offdiag}} := \{\hat{E}_{p+1, q+1}^\ell : 0 \leq p < q \leq \ell - 1\}$ collects the real zero-diagonal basis matrices,

- $E_{\text{imag,offdiag}} := \{\hat{E}_{p+1, q+1}^\ell : 0 \leq q < p \leq \ell - 1\}$ collects the imaginary zero-diagonal basis matrices, and

- $E_{\text{diag}} := \{\hat{E}_{q+1, q+1}^\ell : 0 \leq q \leq \ell - 1\}$ collects the real diagonal basis matrices.

We define a total ordering $\preceq$ on the tuples $(p, q)$ for $p, q = 1, \ldots, \ell$, so that the matrices are ordered with $E_{\text{real,offdiag}} \preceq E_{\text{imag,offdiag}} \preceq E_{\text{diag}}$ in the element-wise sense. For example, when $\ell = 3$,

$$(1, 2) \preceq (1, 3) \preceq (2, 3) \preceq (2, 1) \preceq (3, 1) \preceq (3, 2) \preceq (1, 1) \preceq (2, 2) \preceq (3, 3)$$

For any $(i, j), (\tilde{i}, \tilde{j}) \in \{1, \ldots, \ell\}^2$, we say that $(i, j) \preceq (\tilde{i}, \tilde{j})$ if one of the following holds.

- Case 1: $i < j$ (so that $\hat{E}_{ij}$ is a real matrix with zero diagonal).
- $i < j$ and $\tilde{i} \geq \tilde{j}$. 93
• Case 2: $i > j$ (so that $\hat{E}_{ij}$ is a imaginary matrix with zero diagonal).

In this case we must have $\tilde{i} \geq \tilde{j}$.

- $j < i$ and $\tilde{j} = i$.
- $j < i$ and $\tilde{j} < \tilde{i}$, but $\tilde{i} > i$.
- $j < i$ and $\tilde{j} < \tilde{i} = i$, but $\tilde{j} > j$.

• Case 3: $i = j$ (so that $\hat{E}_{jj}$ is a real diagonal matrix).

In this case we must have $\tilde{i} = \tilde{j}$.

- $j < \tilde{j}$.

From this, we define an ordered orthonormal basis $B_\ell = \{V_1, \ldots, V_\ell\} = \{\hat{E}_{p+1,q+1}\}_{p,q}$ for $\mathcal{H}_\ell$.

Using this basis, we can define the corresponding symmetric vectorization of Hermitian matrices:

$$sHvec : \mathcal{H}_\ell \to \mathbb{R}^{\ell^2} : H \mapsto v,$$

where $v = [v_j] \in \mathbb{R}^{\ell^2}$ is the unique vector such that $H = \sum_{j=1}^{\ell^2} v_j V_j$, is well-defined. The map $sHvec$ is a linear isometry (i.e., $sHvec$ is a linear map and $||sHvec(H)||^2 = \text{tr}(H^2)$ for all $H \in \mathcal{H}_\ell$), and its adjoint is given by

$$sHMat : \mathbb{R}^{\ell^2} \to \mathcal{H}_\ell : v \mapsto \sum_{j=1}^{\ell^2} v_j V_j,$$

(5.12)

which is also the inverse map of $sHvec$.

For example, when $\rho = [a_{kl} + ib_{kl}] \in \mathbb{D}_3$,

$$sHvec(\rho) = \left[ \sqrt{2}a_{12} \quad \sqrt{2}a_{13} \quad \sqrt{2}a_{23} \quad \sqrt{2}b_{12} \quad \sqrt{2}b_{13} \quad \sqrt{2}b_{23} \quad a_{11} \quad a_{22} \quad a_{33} \right]^T$$

94
We now construct the matrix $M \in \mathbb{R}^{k \times m^2}$ by declaring

$$M^T = \begin{bmatrix} \text{shvec}(\rho(1)) & \text{shvec}(\rho(2)) & \ldots & \text{shvec}(\rho(k)) \end{bmatrix}.$$ \hspace{1cm} (5.13)

We then separate $M$ into three blocks

$$M = \begin{bmatrix} M_{\text{Re}} & M_{\text{Im}} & M_D \end{bmatrix},$$ \hspace{1cm} (5.14)

where $M_D \in \mathbb{R}^{k \times m}$ has rows formed from the diagonals of matrices $\rho(j)$, and $M_{\text{Re}}$ and $M_{\text{Im}}$ have rows formed from the real and imaginary parts of $\rho(j)$, respectively, for $j = 1, \ldots, k$. Define now the matrices

\begin{align*}
M_{\text{Re} \text{Im} D} &:= \begin{bmatrix} M_{\text{Re}} & -M_{\text{Im}} & M_D \end{bmatrix}, \\
N_{\text{Re} \text{Im} D} &:= \begin{bmatrix} \frac{1}{\sqrt{2}} M_{\text{Re}} & \frac{1}{\sqrt{2}} M_{\text{Re}} & -\frac{1}{\sqrt{2}} M_{\text{Im}} & -\frac{1}{\sqrt{2}} M_{\text{Im}} & M_D & 0 \\
-\frac{1}{\sqrt{2}} M_{\text{Im}} & \frac{1}{\sqrt{2}} M_{\text{Im}} & -\frac{1}{\sqrt{2}} M_{\text{Re}} & \frac{1}{\sqrt{2}} M_{\text{Re}} & 0 & M_D \end{bmatrix}. \hspace{1cm} (5.15)
\end{align*}

Let $P$ be an $nm \times nm$ hermitian matrix oartitioned as $P = [P_{st}]_{s,t=1}^n$, where $P_{s,t} \in \mathbb{C}^{m \times m}$ for all $s,t$. For $1 \leq p < q \leq m$, define

$$F_{pq} = \begin{bmatrix} \text{Re}(A) & \text{Re}(B) & \text{Im}(A) & \text{Im}(B) & \text{Re}(C) & \text{Im}(C) \end{bmatrix}^T$$

and for $1 \leq q \leq m$, define

$$G_{qq} = \begin{bmatrix} \text{Re}(A) & \text{Im}(A) & C \end{bmatrix},$$

where $A = \begin{bmatrix} (P_{12})_{pq} & (P_{12})_{pq} & \cdots & (P_{n-1,n})_{pq} \end{bmatrix}$, $B = \begin{bmatrix} (P_{12})_{qp} & (P_{12})_{qp} & \cdots & (P_{n-1,n})_{qp} \end{bmatrix}$ and $C = \begin{bmatrix} (P_{11})_{pq} & (P_{22})_{pq} & \cdots & (P_{n,n})_{pq} \end{bmatrix}$. Then the linear constraints defining $L_1$ can be written as

$$N_{\text{Re} \text{Im} D F_{pq}} = \begin{bmatrix} \text{Re}\sigma_{pq}^{(1)} & \cdots & \text{Re}\sigma_{pq}^{(k)} & \text{Im}\sigma_{pq}^{(1)} & \cdots & \text{Im}\sigma_{pq}^{(k)} \end{bmatrix}^T.$$
for all $1 \leq p < q \leq m$ and
\[ M_{\text{Re Im}} G_{qq} = \begin{bmatrix} \sigma_{qq}^{(1)} & \cdots & \sigma_{qq}^{(k)} \end{bmatrix}^T \] (5.16) for all $1 \leq q \leq m$. Meanwhile, the linear constraints defining $L_2$ is given by
\[ \begin{bmatrix} I_{n^2} & \cdots & I_{n^2} \\ G_{11} \\ \vdots \\ G_{mm} \end{bmatrix} = e_m^T \otimes I_{n^2} = \begin{bmatrix} G_{11} \\ \vdots \\ G_{mm} \end{bmatrix} = \begin{bmatrix} 0_{n^2-n,1} \\ e_n \end{bmatrix}, \]
where $e_n$ denotes the all ones vector in $\mathbb{R}^n$. Thus,

Therefore, $L$ can be represented by the following coefficient matrix:
\[ \tilde{L} := \begin{bmatrix} I_{\frac{m(m-1)}{2}} \otimes N_{\text{Re Im D}} & 0 \\ 0 & I_m \otimes M_{\text{Re Im D}} \\ e_m^T \otimes I_{n^2} \end{bmatrix}, \] (5.17)

Note however that some of the rows of the second block of $\tilde{L}$ are linearly independent. In particular, the last constraint describing $L_1$, that is, the equation obtained from 5.16 when $q = m$, is redundant and can be obtained from the constraints in $L_2$. Thus, we replace $\tilde{L}$ by
\[ L := \begin{bmatrix} I_{\frac{m(m-1)}{2}} \otimes N_{\text{Re Im D}} & 0 \\ 0 & \begin{bmatrix} I_{m-1} \otimes M_{\text{Re Im D}} & 0 \end{bmatrix} \\ e_m^T \otimes I_{n^2} \end{bmatrix}, \] (5.18)

Let the matrix $(M_{\text{Re Im D}})_{null}$ have orthonormal columns that yield a basis for $\text{null}(M_{\text{Re Im D}})$, i.e.,
\[ \text{null}(M_{\text{Re Im D}}) = \text{range}((M_{\text{Re Im D}})_{null}). \]

The generalized inverse of the top-left block is trivial to find from $N_{\text{Re Im D}}$. An explicit expression
for the generalized inverse of the bottom right-block can also be found. Therefore, we get an explicit blocked structure for the Moore-Penrose generalized inverse of the complete matrix representation.

\[
L^\dagger = \begin{bmatrix}
I_{(n-1)} \otimes N_{\text{Re Im D}}^\dagger & 0 \\
0 & \begin{bmatrix}
I_{n-1} \otimes M_{\text{Re Im D}}^\dagger & e_{n-1} \otimes (M_{\text{Re Im D}})_{\text{null}} \\
e_{n-1}^T \otimes -M_{\text{Re Im D}}^\dagger & I_{n^2} - (n-1)(M_{\text{Re Im D}})_{\text{null}}
\end{bmatrix}
\end{bmatrix},
\]

as claimed. Thus \(L^\dagger\) is easy to construct by simply stacking various small matrices together in blocks. Moreover, this means that both expressions \(L^p\) and \(L^\dagger R\) can be vectorized and evaluated efficiently and accurately.

### 5.2.2 Numerical Experiments

In this subsection, we numerically illustrate the effectiveness of the projection/reflection methods for solving quantum channel construction problems. The large/huge problems were solved on an AMD Opteron(tm) Processor 6168, 1900.089 MHz cpu running LINUX. The smaller problems were solved using an Optiplex 9020, Intel(R) Core(TM), i7-4770 CPUs, 3.40GHz, RAM 16GB running Windows 7. The Matlab scripts used in this section can be found in


For simplicity of exposition, in our numerical experiments, we set \(n = m\). Moreover, we will impose the common unital constraint \(\Phi(I_n) = I_n\) condition. We note in passing that the unital constraint implies that the last constraint in each density matrix block of constraints for each \(i\) is redundant. To generate random instances for our tests we proceed as follows. We start with given integers \(m = n, k\) and a value for \(r\). We generate a Choi matrix \(P\) using \(r\) random unitary matrices \(F_i, i = 1, \ldots, r\) and a positive probability distribution \(d\), i.e., we set

\[
P = \sum_{i=1}^{r} d_i F_i F_i^*.
\]

97
Note that, given a density matrix $X$, then the trace preserving completely positive map can now be evaluated using the blocked form of $P$ in (5.3) as

$$\Phi(X) = \sum_{ij} X_{ij} P_{ij}.$$ 

We then generate random density matrices $\rho^{(j)}, j = 1, \ldots, k$ and set $\sigma^{(j)}$ as the image of the corresponding trace preserving completely positive map $\Phi$ on $\rho^{(j)}$, for all $j$. This guarantees that we have a feasible instance of rank $r$ and larger/smaller $r$ values result in larger/smaller rank for the feasible Choi matrix $P$. We set $\rho^{(k+1)}$ to be $I_n$ to enforce the unital constraint.

**Solving the basic problem with DR**

We first look at our basic feasibility problem (5.4). We illustrate the numerical results only using the DR algorithm since we found it to be vastly superior to MAP; see Section 5.2.2, below. We found solutions of huge problems with surprisingly high accuracy and very few iterations. The results are presented in Table 5.1. We give the size of the problem, the number of iterations, the norm of the residual (accuracy) at the end, the maximum value of the cosine values indicating the linear rate of convergence, and the total computational time to perform a projection on the PSD cone. The projection on the PSD cone dominates the time of the algorithm, i.e., the total time is roughly the number of iterations times the projection time. To fathom the size of the problems considered, observe that a problem with $m = n = 10^2$ finds a PSD matrix of order $10^4$ which has approximately $10^8/2$ variables. Moreover, we reiterate that the solutions are found with extremely high accuracy in very few iterations.

Note that the CPU time depends approximately linearly in the size $m = n$.

**Heuristic for finding max-rank feasible solutions using DR and MAP**

We now look at the problem of finding high rank feasible solutions. Recall that this corresponds to finding a trace preserving completely positive map $\Phi$ mapping $\rho^{(j)}$ to $\sigma^{(j)}$, so that
<table>
<thead>
<tr>
<th>m=n.k.r</th>
<th>iters</th>
<th>norm-residual</th>
<th>max-cos</th>
<th>PSD-proj-CPUs</th>
</tr>
</thead>
<tbody>
<tr>
<td>90,50,90</td>
<td>6</td>
<td>5.88e-15</td>
<td>.7014</td>
<td>233.8</td>
</tr>
<tr>
<td>100,60,90</td>
<td>7</td>
<td>7.243e-15</td>
<td>.8255</td>
<td>821.7</td>
</tr>
<tr>
<td>110,65,90</td>
<td>7</td>
<td>7.983e-15</td>
<td>.8222</td>
<td>1484</td>
</tr>
<tr>
<td>120,70,90</td>
<td>8</td>
<td>8.168e-15</td>
<td>.8256</td>
<td>2583</td>
</tr>
<tr>
<td>130,75,90</td>
<td>8</td>
<td>7.19e-15</td>
<td>.8288</td>
<td>3607</td>
</tr>
<tr>
<td>140,80,90</td>
<td>9</td>
<td>8.606e-15</td>
<td>.8475</td>
<td>5832</td>
</tr>
<tr>
<td>150,85,90</td>
<td>11</td>
<td>8.938e-15</td>
<td>.8606</td>
<td>6188</td>
</tr>
<tr>
<td>160,90,90</td>
<td>11</td>
<td>9.295e-15</td>
<td>.8718</td>
<td>1.079e+04</td>
</tr>
<tr>
<td>170,95,90</td>
<td>12</td>
<td>9.412e-15</td>
<td>.8918??</td>
<td>1.139e+04</td>
</tr>
</tbody>
</table>

**TABLE 5.1:** Using DR algorithm; for solving huge problems

Φ necessarily has a long operator sum representation (5.5). We moreover use this section to compare the DR and MAP algorithms. Our numerical tests fix \( m = n, k \) and then change the value of \( r \), i.e., the value used to generate the test problems.

The heuristic for finding a large rank solution starts by finding a (current) feasible solution \( P_c \) using a multiple of the identity as the starting point \( P_0 = mnI_{mn} \) and finding a feasible point \( P_c \) using DR. We then set the current point \( P_c \) to be the barycenter of all the feasible points currently found. The algorithm then continues by changing the starting point to the other side and outside of the PSD cone, i.e., the new starting point is found by traveling in direction \( d = mnI_{mn} - \text{tr}(P_c)P_c \) starting from \( P_c \) so that the new starting point \( P_n := P_c + \alpha d \) is not PSD. For instance, we may set \( \alpha = 2^i \|d\|^2 \) for sufficiently large \( i \). We then apply the DR algorithm with the new starting point until we find a PSD matrix \( P \) or no increase in the rank occurs.

Again, we see that we find very accurate solutions and solutions of maximum rank. We find that DR is much more efficient both in the number of iterations in finding a feasible solution from a given starting point and in the number of steps in our heuristic needed to find a large rank solution. In Tables 5.2 and 5.3 we present the output for several values of \( r \) when using DR and MAP, respectively. We use a randomly generated feasibility instance for each value of \( r \) but we start MATLAB with the \textit{rng(default)} settings so the same random instances are generated. We note that the DR algorithm is successful for finding a maximum rank solution and usually after only the first step of the heuristic. The last three \( r = 12, 10, 8 \) values required 8, 9, 12 steps,
respectively. However, the final $P$ solution was obtained to (a high) 9 decimal accuracy.

The MAP always requires many more iterations and at least two steps for the maximum rank solution. It then fails completely once $r \leq 12$. In fact, it reaches the maximum number of iterations while only finding a feasible solution to 3 decimals accuracy for $r = 12$ and then 2 decimals accuracy for $r = 10, 8$. We see that the cosine value has reached 1 for $r = 12, 10, 8$ and the MAP algorithm was making no progress towards convergence.

For each value of $r$ we include:

1. the number of steps of DR that it took to find the max-rank $P$;
2. the minimum/maximum/mean number of iterations for the steps in finding $P^\dagger$;
3. the maximum of the cosine of the angles between three successive iterates $\dagger$;
4. the value of the maximum rank found. $\S$

**Heuristic for finding low rank and rank constrained solutions**

In quantum information science, one might want to obtain a feasible Choi matrix solution $P = (P_{ij})$ with low rank, e.g., [91, Section 4.1]. If we have a bound on the rank, then we could change the algorithm by adding a rank restriction when one projects the current iterate of $P = (P_{ij})$ onto the PSD cone. That is instead of taking the positive part of $P = (P_{ij})$, we take the nonconvex projection

$$P_r := \sum_{j \leq r, \lambda_j > 0} \lambda_j x_j x_j^\ast,$$

where $P$ has spectral decomposition $\sum_{j=1}^{mn} \lambda_j x_j x_j^\ast$ with $\lambda_1 \geq \cdots \geq \lambda_{mn}$.

$\dagger$Note that if the maximum value is the same as `iterlimit`, then the method failed to attain the desired accuracy `toler` for this particular value of $r$.

$\ddagger$This is a good indicator of the expected number of iterations.

$\S$We used the `rank` function in MATLAB with the default tolerance, i.e., rank($P$) is the number of singular values of $P$ that are larger than $mn * \text{eps}(||P||)$, where eps(||$P$||) is the positive distance from ||$P$|| to the next larger in magnitude floating point number of the same precision. Here we note that we did not fail to find a max-rank solution with the DR algorithm.
\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|c|c|}
\hline
r=30 & rank steps & min-iters & max-iters & mean-iters & max-cos & max rank \\
\hline
   1 &       6   &       6   &       6   & 7.008801e-01 & 900 & \\
\hline
r=28 & 1 & 7 & 7 & 7 & 7.323953e-01 & 900 & \\
\hline
r=26 & 1 & 7 & 7 & 7 & 7.550174e-01 & 900 & \\
\hline
r=24 & 1 & 8 & 8 & 8 & 7.911440e-01 & 900 & \\
\hline
r=22 & 1 & 9 & 9 & 9 & 8.238539e-01 & 900 & \\
\hline
r=20 & 1 & 9 & 9 & 9 & 8.454781e-01 & 900 & \\
\hline
r=18 & 1 & 11 & 11 & 11 & 8.730321e-01 & 900 & \\
\hline
r=16 & 1 & 15 & 15 & 15 & 8.995266e-01 & 900 & \\
\hline
r=14 & 1 & 23 & 23 & 23 & 9.288445e-01 & 900 & \\
\hline
r=12 & 8 & 194 & 3500 & 1.916375e+03 & 9.954262e-01 & 900 & \\
\hline
r=10 & 9 & 506 & 3500 & 2.605778e+03 & 9.968120e-01 & 900 & \\
\hline
r=8  & 12 & 2298 & 3500 & 3.350833e+03 & 9.986002e-01 & 900 & \\
\hline
\end{tabular}
\caption{Using DR algorithm; with $[m \ k \ n \ mn \ toler \ iterlimit] = [30 \ 30 \ 16 \ 900 \ 1e-14 \ 3500]$; max/min/mean iter and number rank steps for finding max-rank of $P$. The 3500 here means 9 decimals accuracy attained for last step.}
\end{table}

Alternatively, we can do the following. Suppose a feasible Choi matrix $C(\Phi) = P_c = ((P_c)_{ij})$ is found with $\text{rank}(P_c) = r$. We can then attempt to find a new Choi matrix of smaller rank restricted to the face $F$ of the PSD cone where the current $P_c$ is in the relative interior of $F$, i.e., the minimal face of the PSD cone containing $P_c$. We do this using facial reduction, e.g., [11, 12]. More specifically, suppose that $P_c = VDV^*$ is a compact spectral decomposition, where $D \in PSD_r$ is diagonal, positive definite and has rank $r$. Then the minimal face $F$ of the PSD cone containing $P_c$ has the form $F = V(PSD_r)V^*$. Recall $Lp = b$ denotes the matrix/vector equation corresponding to the linear constraints in our basic problem with $p = sHvec(P)$. Let $L_{i,:}$ denote the rows of the matrix representation $L$. We let $sHMat = sHvec^{-1}$. Note that $sHMat = sHvec^*$, the adjoint. Then each row of the equation $Lp = b$ is equivalent to

$$\langle L_{i,:}^*, sHvec(P) \rangle = \langle sHMat(L_{i,:}^*), V\tilde{P}V^* \rangle = \langle V^*sHMat(L_{i,:}^*), V, \tilde{P} \rangle, \quad \tilde{P} \in PSD_r.$$ 

Therefore, we can replace the linear constraints with the smaller system $\bar{L}\bar{p} = b$ with equations $\langle \bar{L}_{i,:}, \bar{p} \rangle$, where $\bar{L}_{i,:} = sHvec \left( V^*sHMat(L_{i,:}^*)V \right)$. In addition, since the current feasible point $P_c$ is in the relative interior of the face $V(PSD_r)V^*$, if we start outside the PSD cone $PSD_r$ for our
TABLE 5.3: Using MAP algorithm; with \([m \ n \ k \ m \ n \ \text{toler} \ \text{iterlimit}] = [30 \ 30 \ 16 \ 900 \ 1e-14 \ 3500]\); max/min/mean iter and number rank steps for finding max-rank of \(P\). The 3500 mean-iters means max iterlimit reached; low accuracy attained.

feasibility search, then we get a singular feasible \(\bar{P}\) if one exists and so have reduced the rank of the corresponding initial feasible \(P\). We then repeat this process as long as we get a reduction in the rank.

The MAP approach we are using appears to be especially well suited for finding low rank solutions. In particular, the facial reduction works well because we are able to get extremely high accuracy feasible solutions before applying the compact spectral decomposition. If the initial \(P_0\) that is projected onto the affine subspace is not positive semidefinite, then successive iterates on the affine subspace stay outside the semidefinite cone, i.e., we obtain a final feasible solution \(\bar{P}\) that is not positive definite if one exists. Therefore, the rank of \(\bar{V}\bar{V}^*\) is reduced from the rank of \(P\). The code for this has been surprisingly successful in reducing rank. We provide some typical results for small problems in Table 5.4. We start with a small rank (denoted by \(r\)) feasible solution that is used to generate a feasible problem. Therefore, we know that the minimal rank is \(\leq r\). We then repeatedly solve the problem using facial reduction until a positive definite solution is found which means we cannot continue with the facial reduction. Note that we could restart the algorithm using an upper bound for the rank obtained from the last rank we obtained.

Finally, our tests indicate that the rank constrained problem, which is nonconvex, often can
TABLE 5.4: Using MAP algorithm with facial reduction for decreasing the rank

<table>
<thead>
<tr>
<th>m=n,k</th>
<th>initial rank r</th>
<th>facial red. ranks</th>
<th>final rank</th>
<th>final norm-residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>12,10</td>
<td>11</td>
<td>100,50,44,39</td>
<td>39</td>
<td>1.836e-15</td>
</tr>
<tr>
<td>12,10</td>
<td>10</td>
<td>92,61,43,44</td>
<td>44</td>
<td>1.786e-15</td>
</tr>
<tr>
<td>20,14</td>
<td>20</td>
<td>304,105,71</td>
<td>71</td>
<td>9.648e-15</td>
</tr>
<tr>
<td>22,13</td>
<td>20</td>
<td>374,121,75</td>
<td>75</td>
<td>9.746e-15</td>
</tr>
</tbody>
</table>

be solved efficiently. Moreover, this problem helps in further reducing the rank. To see this, suppose that we know a bound, rbnd, on the rank of a feasible \( P \). Then, as discussed above, we change the projection onto the PSD cone by using only the largest rbnd eigenvalues of \( P \). In our tests, if we use \( r \), the value from generating our instances, then we were always successful in finding a feasible solution of rank \( r \). Our final tests appear in Table 5.5. We generate problems with initial rank \( r \). We then start solving a constrained rank problem with starting constraint rank \( r_s \) and decrease this rank by 1 until we can no longer find a feasible solution; the final rank with a feasible solution is \( r_f \). At each successful reduction, we found a feasible solution to the requested tolerance \( 1e-14 \).

<table>
<thead>
<tr>
<th>m = n,k</th>
<th>initial rank r</th>
<th>starting constr. rank ( r_s )</th>
<th>final constr. rank ( r_f )</th>
</tr>
</thead>
<tbody>
<tr>
<td>12,9</td>
<td>15</td>
<td>20</td>
<td>7</td>
</tr>
<tr>
<td>25,16</td>
<td>35</td>
<td>45</td>
<td>19</td>
</tr>
<tr>
<td>30,21</td>
<td>38</td>
<td>48</td>
<td>27</td>
</tr>
</tbody>
</table>

TABLE 5.5: Using DR algorithm for rank constrained problems with ranks \( r_s \) to \( r_f \)

Table 5.6 illustrates the DR algorithm for finding a low rank solution for the first instance in Table 5.5. We begin with starting rank 20. We see the increase in max-cos and simultaneously the number of iterations needed to find a feasible solution as the rank constraint decreases. We stop in reducing rank once we cannot find a feasible solution with the iteration limit for DR set at 3,500.
TABLE 5.6: Using DR algorithm for rank constrained problem instance one in Table 5.5 with $m = n = 12$, $k = 9$, $r = 15$ and starting constrained rank 20 till final successful constrained rank 7; feasibility failed for constrained rank 6 with iteration limit 3,500.

<table>
<thead>
<tr>
<th>current constrained rank</th>
<th>max-cos</th>
<th>norm(residual)</th>
<th>iterations</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>9.5183e-01</td>
<td>8.6510e-15</td>
<td>6.4700e+02</td>
</tr>
<tr>
<td>19</td>
<td>9.4773e-01</td>
<td>9.1083e-15</td>
<td>6.9600e+02</td>
</tr>
<tr>
<td>18</td>
<td>9.5347e-01</td>
<td>9.8330e-15</td>
<td>7.4700e+02</td>
</tr>
<tr>
<td>17</td>
<td>9.5947e-01</td>
<td>9.6879e-15</td>
<td>8.2300e+02</td>
</tr>
<tr>
<td>16</td>
<td>9.6289e-01</td>
<td>9.9593e-15</td>
<td>8.9700e+02</td>
</tr>
<tr>
<td>15</td>
<td>9.7182e-01</td>
<td>9.4914e-15</td>
<td>9.9700e+02</td>
</tr>
<tr>
<td>14</td>
<td>9.7775e-01</td>
<td>9.3193e-15</td>
<td>1.1670e+03</td>
</tr>
<tr>
<td>13</td>
<td>9.7630e-01</td>
<td>9.8646e-15</td>
<td>1.2830e+03</td>
</tr>
<tr>
<td>12</td>
<td>9.8125e-01</td>
<td>9.6170e-15</td>
<td>1.4250e+03</td>
</tr>
<tr>
<td>11</td>
<td>9.8389e-01</td>
<td>9.8741e-15</td>
<td>1.6660e+03</td>
</tr>
<tr>
<td>10</td>
<td>9.8834e-01</td>
<td>9.8033e-15</td>
<td>1.9860e+03</td>
</tr>
<tr>
<td>9</td>
<td>9.9109e-01</td>
<td>9.9461e-15</td>
<td>2.4430e+03</td>
</tr>
<tr>
<td>8</td>
<td>9.9260e-01</td>
<td>9.1184e-15</td>
<td>2.9920e+03</td>
</tr>
<tr>
<td>7</td>
<td>9.9704e-01</td>
<td>4.5293e-13</td>
<td>3.5000e+03</td>
</tr>
<tr>
<td>6</td>
<td>9.9960e-01</td>
<td>1.5008e-05</td>
<td>3.5000e+03</td>
</tr>
</tbody>
</table>

5.3 Quantum States with Prescribed Reduced States and Prescribed Eigenvalues

In Section 1.2, we considered a multipartite system $X = (X_1, \ldots, X_k)$ whose state is $\rho \in D_{j_1 \ldots j_k}$, and the state of the component $X_s$ is in $D_{j_s}$. For any subset $J = \{j_1, \ldots, j_r\} \subseteq \{1, \ldots, k\}$, we also defined the partial trace map $\text{tr}_{J^c}$ in equation (1.8), so that $\text{tr}_{J^c}(\rho)$ gives the reduced state of the subsystem $X_J = (X_{j_1}, \ldots, X_{j_r})$.

For example, if $k = 2$, we have a bipartite system. There are two partial traces of the form

$$\rho_1 \otimes \rho_2 \mapsto \rho_1 \quad \text{and} \quad \rho_1 \otimes \rho_2 \mapsto \rho_2$$

for any product states $\rho_1 \otimes \rho_2$. Clearly, the two maps correspond to the case when $J^c = \{2\}$ and $J^c = \{1\}$, respectively. We will use the notation $\text{tr}_2$ and $\text{tr}_1$ for the two maps for notation.

\*\*The material in this section is contained in the paper [32], which is a joint work of X.-F. Duan, C.-K. Li and the author.\*\*
simplicity. For a general state \( \rho = (\rho_{ij})_{1 \leq i, j \leq n_1} \in D_{n_1 \cdots n_k} \) such that \( \rho_{ij} \in \mathbb{C}^{n_2 \times n_2} \), we have

\[
\text{tr}_1(\rho) = \sum_{j=1}^{n_1} \rho_{jj} \in \mathbb{C}^{n_2 \times n_2} \quad \text{and} \quad \text{tr}_2(\rho) = (\text{tr}_1 \rho_{ij})_{1 \leq i, j \leq n_1} \in \mathbb{C}^{n_1 \times n_1}.
\]

If \( k = 3 \), we have a tripartite system, and there are six partial traces such that

\[
\begin{align*}
\text{tr}_1(\rho_1 \otimes \rho_2 \otimes \rho_3) &= \rho_2 \otimes \rho_3, \\
\text{tr}_2(\rho_1 \otimes \rho_2 \otimes \rho_3) &= \rho_1 \otimes \rho_3, \\
\text{tr}_3(\rho_1 \otimes \rho_2 \otimes \rho_3) &= \rho_1 \otimes \rho_2, \\
\text{tr}_{12}(\rho_1 \otimes \rho_2 \otimes \rho_3) &= \rho_3, \\
\text{tr}_{23}(\rho_1 \otimes \rho_2 \otimes \rho_3) &= \rho_1, \\
\text{tr}_{13}(\rho_1 \otimes \rho_2 \otimes \rho_3) &= \rho_2.
\end{align*}
\]

In this section, we study the following problem:

**Problem 5.3.1.** Construct a global state \( \rho \in D_{n_1 \cdots n_k} \) with certain prescribed reduced (marginal) states \( \rho_{J_1}, \ldots, \rho_{J_m} \). Equivalently, if \( N = n_1 \cdots n_k \), find \( \rho \in PSD_N \cap S_2 = \{ \rho : \text{tr}_{J_i} = \rho_{J_i}, \ldots, \text{tr}_{J_m} = \rho_{J_m} \} \). Given \( a_1, \ldots, a_N \geq 0 \) such that \( \sum_{j=1}^{N} a_j = 1 \), find \( \rho \in D_{n_1 \cdots n_k} \) with certain prescribed reduced (marginal) states \( \rho_{J_1}, \ldots, \rho_{J_m} \) and such that \( \rho \) has eigenvalues \( a_1, \ldots, a_N \). That is, find \( \rho \in S_1 \cap S_2 \), where \( S_1 = \{ U \text{diag}(a_1, \ldots, a_N) U^* \mid U \in \mathcal{U}_N \} \).

For a bipartite case, if \( \rho_1 \in D_{n_1} \) and \( \rho_2 \in D_{n_2} \), then \( \rho = \rho_1 \otimes \rho_2 \in D_{n_1 n_2} \) is a global state having reduced states \( \rho_1 \) and \( \rho_2 \). However, it is not easy to construct a global state with prescribed eigenvalues. Researchers have used advanced techniques in representation theory (see [23, 55] and their references) to study the eigenvalues of the global state and the reduced states. The results are described in terms of numerous linear inequalities even for a moderate size problem (see [55]). Moreover, even if one knows that a global state with prescribed eigenvalues exists, it is not possible to construct the density matrix based on the proof. It is not easy to use these results to answer basic problems, test conjectures, or find general patterns of global states with prescribed properties. For multipartite system with more than two subsystems, the problem is more challenging. Not much results are available. For example, for a tripartite system, determining whether there is a state \( \rho \in D_{n_1 n_2 n_3} \) with given reduced states \( \rho_{12} \in D_{n_1 n_2} \) and \( \rho_{23} \in D_{n_2 n_3} \) is an open problem.
We employ the alternating projection method in the following algorithm to study Problem 5.3.1.

**Algorithm 5.3.2.** For constructing a state \( \rho \in PSD_N \cap S_2 \) (respectively, \( \rho \in S_1 \cap S_2 \))

**Step 1.** Choose a positive integer \( L \) (say \( L = 1000 \)) as iteration limit and a small positive integer \( \delta \) (say \( \delta = 10^{-15} \)) as an error/tolerance value and set \( k = 0 \).

**Step 2.** Generate a random density matrix \( \rho^{(0)} \). Do the next step for \( k \leq N \).

**Step 3.** For \( k \geq 1 \), let \( \rho^{(2k-1)} = \text{proj}_{S_2}(\rho^{(2k-2)}) \) and

\[
\rho^{(2k)} = \text{proj}_{PSD_N}(\rho^{(2k-1)}) \quad (\text{respectively, } \rho^{(2k)} \in \text{proj}_{S_1}(\rho^{(2k-1)})).
\]

If \( ||\rho^{(2k+1)} - \rho^{(2k)}||_2 < \delta \), then stop and declare \( \rho^{(2k)} \) as a solution.

We know that if \( \sigma \) is a hermitian matrix with spectral decomposition \( \sigma = UDU^* \), then \( \text{proj}_{PSD_N}(\sigma) = UD_+U^* \), where \( D_+ \) is the diagonal matrix obtained from \( D \) by replacing the negative eigenvalues by 0. The set \( S_2 \) is a non-convex linear manifold. We can determine \( \text{proj}_{S_2} \) using the following result due to Hoffman and Wielandt; for example, see [76, Theorem 10.B.10].

**Theorem 5.3.3.** Let \( ||\cdot|| \) be a unitary similarity invariant norm and suppose \( P = UDU^* \in \mathcal{H}_N \), where \( U \in \mathcal{U}_N \) and \( D \) is a diagonal matrix with diagonal entries arranged in descending order. Then, for all \( Z \in S_1 = \{ V\text{diag}(a_1,\ldots,a_N)V^* \mid V \in \mathcal{U}_N \} \)

\[
||P - U\text{diag}(a_1,\ldots,a_N)U^*|| \leq ||P - Z||
\]

(5.21)

Note that if the eigenvalues \( a_1,\ldots,a_N \) are not distinct, the set \( \text{proj}_{S_1} \) is not a singleton. When implementing algorithm 5.3.2, we may choose any element of \( \text{proj}_{S_1} \). If \( S_1 \cap S_2 \neq \emptyset \), Theorem 4.3 of [60] guarantees local convergence of this algorithm. That is, if we choose a suitable starting point \( \rho_0 \), then the algorithm produces a sequence \( \{\rho^{(k)}\} \) that converges to a \( \rho \in S_1 \cap S_2 \) as \( k \rightarrow \infty \).

In the next two subsections, we will discuss the operator \( \text{proj}_{S_2} \) in detail and illustrate some numerical examples. In our study, we always use the Frobenius norm \( ||X||_2 = [\text{tr}(X^*X)]^{1/2} \),

106
which is unitary similarity invariant.

5.3.1 Projection Operators

To use the projection methods, we need to find the least square projection of a hermitian matrix $Z \in \mathcal{H}_{n_1 \cdots n_k}$ to the linear manifold

$$S_2 = \{ X : \text{tr}_{J^c}(X) = \rho_{J_s}, s = 1, \ldots, m \}. \quad (5.22)$$

Note that if $\mathcal{L} : \mathcal{H}_{n_1 n_2} \rightarrow \mathcal{H}_{n_1}$ such that $\mathcal{L}(X) = \text{tr}_2(X)$, then for any $Y \in \mathcal{H}_{n_1}$, we have

$$\mathcal{L}^\dagger(Y) = Y \otimes \frac{1}{n_2} I_{n_2}. \quad (5.23)$$

Therefore, the following proposition holds.

**Proposition 5.3.4.** Let $J \subseteq \{1, \ldots, k\}$. Given $Z \in \mathcal{H}_{n_1} \otimes \cdots \otimes \mathcal{H}_{n_k}$, the least square projection of $Z$ in $S_2 = \{ \rho \in \mathcal{H}_{n_1} \otimes \cdots \otimes \mathcal{H}_{n_k} : \text{tr}_{J^c}(\rho) = \sigma \}$ is given by

$$\text{proj}_{S_2}(Z) = Z - M_J(Z, \sigma), \quad (5.24)$$

where

$$M_J(Z, \sigma) = P_J^T \left( \frac{I_{n_{J^c}}}{n_{J^c}} \otimes (\text{tr}_{J^c}(Z) - \sigma) \right) P_J, \quad (5.25)$$

$n_{J^c} = \prod_{j \in J^c} n_j$ and $P_J$ is the permutation matrix such that

$$P_J(\alpha_1 \otimes \alpha_2 \otimes \cdots \otimes \alpha_k)P_J^T = \bigotimes_{j \in J^c} \alpha_j \otimes \bigotimes_{j \in J} \alpha_j. \quad (5.26)$$

Now, we use the notation introduced in equation (5.25) to give the formula for the general case. The proof is in Appendix C

**Proposition 5.3.5.** Let $J_1, \ldots, J_m \subseteq \{1, \ldots, k\}$ and $S_2$ be defined as in (5.22). Then $S_2 \neq \emptyset$ if
and only if for any subset \( \{J_{j_1}, \ldots, J_{j_r}\} \) of \( \{J_1, \ldots, J_m\} \), the following partial trace is fixed for all \( t = 1, \ldots, r \)

\[
\text{tr} \left( \bigcap_{s=1}^r J_{j_s} \right) (\rho_{j_{j_1}}) := \rho_{\bigcap_{s=1}^r J_{j_s}}.
\]

Furthermore, the least square projection of a given \( Z \in \mathcal{H}_{n_1 \cdots n_k} \) is

\[
\text{proj}_{S_2}(Z) = Z + \sum_{r=1}^m (-1)^r \sum_{\{J_{j_1}, \ldots, J_{j_r}\} \subseteq \{J_1, \ldots, J_m\}} \mathcal{M} \left( \bigcap_{s=1}^r J_{j_s} \right) \left( Z, \rho_{\bigcap_{s=1}^r J_{j_s}} \right)
\]

As an example, if \( m = 2 \), we get the following projection formula.

**Corollary 5.3.6.** The set \( S_2 = \{ \rho \in \mathcal{H}_{n_1n_2} : \text{tr}_1(\rho) = \sigma_2 \in \mathcal{D}_{n_2} \text{ and } \text{tr}_2(\rho) = \sigma_1 \in \mathcal{D}_{n_1} \} \) is nonempty and the least square projection of a given \( Z \in \mathcal{H}_{n_1n_2} \) onto the set \( S_2 \) is given by

\[
\text{proj}_{S_2}(Z) = Z - \left[ \frac{I_{n_1}}{n_1} \otimes (\text{tr}_1(Z) - \sigma_2) \right] - \left[ (\text{tr}_2(Z) - \sigma_1) \otimes \frac{I_{n_2}}{n_2} \right] + (\text{tr}(Z) - 1)I_{n_1n_2}
\]

Suppose we are interested in looking for a tripartite state \( \rho \in \mathcal{D}_{n_1n_2n_3} \) with given partial traces \( \text{tr}_1(\rho) = \rho_{23} \) and \( \text{tr}_3(\rho) = \rho_{12} \). Then we can use Proposition 5.3.5 to obtain the following projection formula.

**Corollary 5.3.7.** The set

\[
S_2 = \{ \rho \in \mathcal{H}_{n_1n_2n_3} : \text{tr}_1(\rho) = \sigma_2 \in \mathcal{D}_{n_2} \text{ and } \text{tr}_3(\rho) = \sigma_1 \in \mathcal{D}_{n_1} \}
\]

is nonempty if and only if \( \text{tr}_{13} \left( \frac{I_{n_1}}{n_1} \otimes \sigma_2 \right) = \gamma = \text{tr}_{13} \left( \sigma_1 \otimes \frac{I_{n_2}}{n_2} \right) \). In this case, the least square projection of a given \( Z \in \mathcal{H}_{n_1n_2n_3} \) onto the set \( S_2 \) is given by

\[
\text{proj}_{S_2}(Z) = Z - \left[ \frac{I_{n_1}}{n_1} \otimes (\text{tr}_1(Z) - \sigma_2) \right] - \left[ (\text{tr}_3(Z) - \sigma_1) \otimes \frac{I_{n_3}}{n_3} \right] \\
+ \left[ \frac{I_{n_1}}{n_1} \otimes (\text{tr}_{13}(Z) - \gamma) \otimes \frac{I_{n_3}}{n_3} \right] + (\text{tr}(Z) - 1)I_{n_1n_2n_3}
\]
5.3.2 Numerical Experiments

In this section, some examples are tested to illustrate that Algorithms 5.3.2 is feasible and effective to solve Problem 5.3.1. All experiments are performed in MATLAB R2015a on a PC with an Intel Core i7 processor at 2.40GHz with machine precision $\varepsilon = 2.22 \times 10^{-16}$. The programs can be downloaded from http://cklixx.people.wm.edu/mathlib/projection/.

Example 5.3.8. We take $n_1 = n_2 = n_3 = 2$ implement Algorithm 5.3.2, to find a tripartite state $\rho \in \mathcal{D}_8$ such that $\text{tr}_1(\rho) = \rho_{23}$ and $\text{tr}_3(\rho) = \rho_{12}$, where

$$
\rho_{23} = \begin{bmatrix}
0.181375 & 0.161 & 0.1678 & 0.1417 \\
0.161 & 0.314875 & 0.2653 & 0.1937 \\
0.1678 & 0.2653 & 0.307275 & 0.1863 \\
0.1417 & 0.1937 & 0.1863 & 0.196475
\end{bmatrix} \in \mathcal{D}_4,
$$

$$
\rho_{12} = \begin{bmatrix}
0.214875 & 0.1653 & 0.1926 & 0.1934 \\
0.1653 & 0.264475 & 0.2166 & 0.1888 \\
0.1926 & 0.2166 & 0.281375 & 0.1962 \\
0.1934 & 0.1888 & 0.1962 & 0.239275
\end{bmatrix} \in \mathcal{D}_4.
$$

The algorithm produces the solution

$$
\rho = \begin{bmatrix}
0.0811 & 0.0809 & 0.0747 & 0.0654 & 0.0850 & 0.0901 & 0.0923 & 0.07 \\
0.0809 & 0.1338 & 0.1189 & 0.0906 & 0.0898 & 0.1076 & 0.1003 & 0.1011 \\
0.0747 & 0.1189 & 0.1637 & 0.0893 & 0.1053 & 0.0658 & 0.0944 & 0.0947 \\
0.0654 & 0.0906 & 0.0893 & 0.1008 & 0.0728 & 0.1113 & 0.1013 & 0.0944 \\
0.085 & 0.0898 & 0.1053 & 0.0728 & 0.1003 & 0.0801 & 0.0931 & 0.0763 \\
0.0901 & 0.1076 & 0.0658 & 0.1113 & 0.0801 & 0.1811 & 0.1464 & 0.1031 \\
0.0923 & 0.1003 & 0.0944 & 0.1013 & 0.0931 & 0.1464 & 0.1436 & 0.097 \\
0.07 & 0.1011 & 0.0947 & 0.0944 & 0.0763 & 0.1031 & 0.097 & 0.0957
\end{bmatrix}
$$
with an error \( \max(0, - \min(eig(P))) + \|\rho - \text{proj}_{S_2}(\rho)\|_2 < 10^{-16} \). This rank 6 solution is found after approximately 400 iterations, where one iteration consists of a projection on \( \text{PSD}_8 \) and a projection on \( S_2 \). The result was obtained in approximately 0.3 seconds. Note that if \( n_1 = n_3 = 2 \) and \( n_2 \) is increased to \( n = 8 \), this program still obtains a solution relatively fast and accurately.

**Example 5.3.9.** We use the same \( \rho_{23}, \rho_{12} \) in the previous example to find \( \rho \in D_8 \) with \( \text{tr}_1(\rho) = \rho_1, \text{tr}_3(\rho) = \rho_2 \) with the additional condition that the eigenvalues of \( \rho \) are

\[
\begin{bmatrix}
0.8034, 0.0889, 0.05204, 0.0284, 0.0188, 0.0051, 0.0032, 0.0001
\end{bmatrix}
\]

The algorithm ran in under 0.2 seconds and approximately 300 iterations to produce the solution

\[
\rho = \begin{bmatrix}
0.1507 & 0.1056 & 0.0999 & 0.0769 & 0.1047 & 0.0966 & 0.1264 & 0.1293 \\
0.1056 & 0.1209 & 0.0977 & 0.0716 & 0.0813 & 0.0792 & 0.1248 & 0.1018 \\
0.0999 & 0.0977 & 0.1144 & 0.0680 & 0.0879 & 0.0685 & 0.1241 & 0.1100 \\
0.0769 & 0.0716 & 0.0680 & 0.1274 & 0.1053 & 0.0559 & 0.0836 & 0.0821 \\
0.1047 & 0.0813 & 0.0879 & 0.1053 & 0.1160 & 0.0818 & 0.0990 & 0.1055 \\
0.0966 & 0.0792 & 0.0685 & 0.0559 & 0.0818 & 0.0832 & 0.0795 & 0.0870 \\
0.1264 & 0.1248 & 0.1241 & 0.0836 & 0.0990 & 0.0795 & 0.1549 & 0.1297 \\
0.1293 & 0.1018 & 0.1100 & 0.0821 & 0.1055 & 0.0870 & 0.1297 & 0.1324
\end{bmatrix}
\]

with an error \( \|\rho - \text{proj}_{S_2}(\rho)\|_2 + \|eig^+(\rho) - d\|_2 < 10^{-16} \).

**Example 5.3.10.** In this example, we illustrate Algorithm 5.3.2 for the case that \( \rho \in D_8 \) and
\[ \text{tr}_3(\rho) = \rho_{12} = \rho_{13} = \text{tr}_2(\rho). \]

Let
\[
\rho_{12} = \rho_{13} = \begin{bmatrix}
0.2471 & 0.1842 & 0.1738 & 0.2546 \\
0.1842 & 0.2277 & 0.1386 & 0.2144 \\
0.1738 & 0.1386 & 0.182 & 0.2303 \\
0.2546 & 0.2144 & 0.2303 & 0.3432
\end{bmatrix}.
\]

This type of problem is an example of a 2–symmetric extension problem. In [19], the existence of a solution to such a problem was characterized using the concept of separability of quantum states. Using Algorithm 5.3.2, we find a solution
\[
\rho = \begin{bmatrix}
0.1302 & 0.1096 & 0.1111 & 0.1071 & 0.0615 & 0.1156 & 0.1151 & 0.1470 \\
0.1096 & 0.1169 & 0.1147 & 0.0731 & 0.0554 & 0.1123 & 0.1139 & 0.1395 \\
0.1111 & 0.1147 & 0.1169 & 0.0746 & 0.0547 & 0.1152 & 0.1123 & 0.1390 \\
0.1071 & 0.0731 & 0.0746 & 0.1108 & 0.0483 & 0.0839 & 0.0832 & 0.1021 \\
0.0615 & 0.0554 & 0.0547 & 0.0483 & 0.0322 & 0.0649 & 0.0650 & 0.0789 \\
0.1156 & 0.1123 & 0.1152 & 0.0839 & 0.0649 & 0.1498 & 0.1427 & 0.1653 \\
0.1151 & 0.1139 & 0.1123 & 0.0832 & 0.0650 & 0.1427 & 0.1408 & 0.1641 \\
0.1470 & 0.1395 & 0.1390 & 0.1021 & 0.0789 & 0.1653 & 0.1641 & 0.2024
\end{bmatrix}
\]

with an error of order $10^{-17}$ after 2353 iterations in 1.9 seconds.

**Example 5.3.11.** We take $n_1 = 2$ and $n_2 = 3$ and we set
\[
\rho_2 = \begin{bmatrix}
0.4922 & 0.2729 & 0.3138 \\
0.2729 & 0.1980 & 0.1846 \\
0.3138 & 0.1846 & 0.3098
\end{bmatrix}, \quad \rho_1 = \begin{bmatrix}
0.52 & 0.3923 \\
0.3923 & 0.48
\end{bmatrix}.
\]

We use algorithm 5.3.2 to find $\rho \in \mathcal{D}_6$ with $\text{tr}_1(\rho) = \rho_2$, $\text{tr}_2(\rho) = \rho_1$ and prescribed eigenvalues $(0.8329, 0.0781, 0.0529, 0.0238, 0.0109, 0.0015)$. We obtain the following solution after 214
iterations and an error $\approx 3.38 \times 10^{-16}$.

$$
\rho = \begin{bmatrix}
0.2826 & 0.1614 & 0.1582 & 0.1990 & 0.0908 & 0.1861 \\
0.1614 & 0.1234 & 0.0945 & 0.1258 & 0.0601 & 0.1234 \\
0.1582 & 0.0945 & 0.1140 & 0.1088 & 0.0470 & 0.1333 \\
0.1990 & 0.1258 & 0.1088 & 0.2096 & 0.1115 & 0.1556 \\
0.0908 & 0.0601 & 0.0470 & 0.1115 & 0.0746 & 0.0901 \\
0.1861 & 0.1234 & 0.1333 & 0.1556 & 0.0901 & 0.1958 \\
\end{bmatrix}.
$$

### 5.4 Low Rank Bipartite States with Prescribed Reduced States and Rank \[\|\]

In this section, we focus on bipartite states with prescribed reduced states $\rho_1 \in D_{n_1}$ and $\rho_2 \in D_{n_2}$. In particular, we will let

$$S(\rho_1, \rho_2) = \{ \rho \in D_{n_1 \times n_2} : \text{tr}_1(\rho) = \rho_2, \text{tr}_2(\rho) = \rho_1 \}. \quad (5.32)$$

The set $S(\rho_1, \rho_2)$ is compact, convex, and non-empty containing $\rho_1 \otimes \rho_2$. Note that

$$S(\rho_1 \oplus 0_s, \rho_2 \oplus 0_t) = \{ [\rho_{ij} \oplus 0_t] \oplus 0_{s(n_2+s)} : [\rho_{ij}] \in S(\rho_1, \rho_2) \}$$

and for any unitaries $U \in U_{n_1}$ and $V \in U_{n_2}$,

$$S(U \rho_1 U^*, V \rho_2 V^*) = \{ (U \otimes V) \rho (U \otimes V)^* : \rho \in S(\rho_1, \rho_2) \} = (U \otimes V) S(\rho_1, \rho_2) (U \otimes V)^*.$$

Note also that if $T : \mathbb{C}^{n_1 n_2 \times n_1 n_2} \to \mathbb{C}^{n_1 n_2 \times n_1 n_2}$ is the linear map satisfying $T(X_1 \otimes X_2) = X_2 \otimes X_1$ for all $X_1 \in \mathbb{C}^{n_1 \times n_1}$ and $X_2 \in \mathbb{C}^{n_2 \times n_2}$, then

$$S(\rho_2, \rho_1) = \{ T(\rho) : \rho \in S(\rho_1, \rho_2) \}$$

Hence, if convenient, we may focus on the case when $n_1 \leq n_2$ and $\rho_1 \in D_{n_1}, \rho_2 \in D_{n_2}$ are positive definite and are in diagonal form.

\[\|\]The material in this section is also part of [32].
In this section, we discuss methods to find $\rho \in S(\rho_1, \rho_2)$ with a prescribed rank, with special attention to low rank solutions. Note that low rank solutions are of great interest as they are often entangled [83, Theorem 8]. In fact, it was shown in [51, Theorem 1] that if $\text{rank}(\rho) < \max\{\text{rank}(\rho_1), \text{rank}(\rho_2)\}$ then $\rho$ must be distillable. It is also known (for example, see [92]) that if $\rho \in S(\rho_1, \rho_2)$, then

$$\max\left\{ \left\lfloor \frac{\text{rank}(\rho_1)}{\text{rank}(\rho_2)} \right\rfloor, \left\lfloor \frac{\text{rank}(\rho_2)}{\text{rank}(\rho_1)} \right\rfloor \right\} \leq \text{rank}(\rho) \leq \text{rank}(\rho_1)\text{rank}(\rho_2)$$  \hspace{1cm} (5.33)

The upper bound is always attained by $\rho = \rho_1 \otimes \rho_2$ but the lower bound is not always attained. For example, in [54, Subsection 3.3.1], it was shown that there exists a rank one $\rho \in S(\rho_1, \rho_2)$ if and only if $\rho_1$ and $\rho_2$ are isospectral, that is, $\rho_1$ and $\rho_2$ have the same set of nonzero eigenvalues, counting multiplicities.

The following algorithm is an implementation of an alternating projection method to find a low rank solution $\rho \in S(\rho_1, \rho_2)$, if it exists.

**Algorithm 5.4.1.** Alternating projection scheme to find $\rho \in S(\rho_1, \rho_2)$ with $\text{rank}(\rho) \leq k$.

**Step 1:** Set $r = 0$ and choose $X_0 \in D_{n_1n_2}$ and a positive integer $N$ (iteration limit) and a small positive integer $\delta$ (tolerance). Do the next step for $r = 1, \ldots, N$.

**Step 2:** Using Corollary 5.3.6, define

$$\rho^{(2r-1)} = \text{proj}_{S_2}(\rho^{(2r-2)})$$

Then if $\rho^{(2r-1)} = U\text{diag}(d_1, \ldots, d_{n_1n_2})U^*$ for some unitary $U$ and $d_1 \geq d_2 \geq \cdots \geq d_{n_1n_2} \geq 0$, define

$$\rho^{(2r)} = U(s_1, \ldots, s_k, 0, \ldots, 0)U^*,$$

where $s_j = \max\{d_j, 0\}$. If $\max\{|\text{tr}_1(\rho^{(2r)}) - \rho_2|, |\text{tr}_2(\rho^{(2r)}) - \rho_1|\} < \delta$, then declare $\rho^{(2r)}$ as a solution.
Note that we defined $\rho^{(2r)}$ in step 2 of algorithm 5.4.1 so that

$$||\rho^{(2r-1)} - \rho^{(2r)}|| \leq ||\rho^{(2r-1)} - Z||$$

for any positive semidefinite rank matrix $Z$ with rank at most $k$ [76, Theorem 10.B.10]. Convergence of this algorithm is not guaranteed but numerical results shown in Section 5.4.2 illustrate that this algorithm is effective in finding a low rank solution.

### 5.4.1 Constructions of a Low Rank Solution

In view of the fact that the above algorithm may not converge and multiple low rank solutions may exist, we derive other methods to find low rank solutions, namely Proposition 5.4.3 and Algorithms 5.4.6 and 5.4.8. Proposition 5.4.3 provides a simple way to construct a separable $\rho \in S(\rho_1, \rho)$ whose rank can be chosen to be anything between $\max\{\text{rank}(\rho_1), \text{rank}(\rho_2)\}$ up to $\text{rank}(\rho_1) + \text{rank}(\rho_2) - 1$. Meanwhile, Algorithms 5.4.6 and 5.4.8 both construct a specific solution $\rho \in S(\rho_1, \rho_2)$ whose rank is guaranteed to be less than or equal to $\max\{\text{rank}(\rho_1), \text{rank}(\rho_2)\}$. Solutions obtained from these algorithms may not give the minimal rank. However, numerical experiments illustrate that these relatively low rank solutions can be utilized as a starting point for algorithm 5.4.1 to obtain a minimal rank solution. Additionally, as we will see in Section 5.4.2, two of the algorithms produce a solution with low von Neumann entropy.

First, we present the following theorem (see for example [54]) to construct a rank one solution $\rho \in S(\rho_1, \rho_2)$ for isospectral hermitian matrices $\rho_1$ and $\rho_2$, that is, $\rho_1$ and $\rho_2$ have the same nonzero eigenvalues and corresponding multiplicities. In fact, it is known that $S(\rho_1, \rho_2)$ contains a rank 1 element if and only if $\rho_1$ and $\rho_2$ are isospectral. This will be the basis for the three algorithms that we will define in this subsection.

**Theorem 5.4.2.** Let $\rho_1 \in D_{n_1}$ and $\rho_2 \in D_{n_2}$ have spectral decomposition $\rho_1 = \gamma_1|x_1\rangle\langle x_1| + \cdots + \gamma_2|x_2\rangle\langle x_2| + \cdots$ and $\rho_2 = \beta_1|x_1\rangle\langle x_1| + \cdots + \beta_2|x_2\rangle\langle x_2| + \cdots$. Then

1. $\rho_1 = \rho_2$ if and only if $\gamma_k = \beta_k$ for $k = 1, 2, \ldots$.
2. $\rho_1$ and $\rho_2$ are isospectral.
3. $\rho_1$ and $\rho_2$ are isospectral if and only if $\gamma_k = \beta_k$ for $k = 1, 2, \ldots$.

Conversely, if $\rho_1$ and $\rho_2$ are isospectral, then $\rho_1$ and $\rho_2$ are isospectral if and only if $\gamma_k = \beta_k$ for $k = 1, 2, \ldots$.
\( \gamma_k |x_k\rangle \langle x_k| \) and \( \rho_2 = \gamma_1 |y_1\rangle \langle y_1| + \cdots + \gamma_k |y_k\rangle \langle y_k| \), and

\[ |w\rangle = \sum_{j=1}^{k} \sqrt{\gamma_j} |x_j\rangle \otimes |y_j\rangle \]

Then \( P = |w\rangle \langle w| \in \mathcal{S}(\rho_1, \rho_2) \).

In the following proposition, we can choose an integer \( k \) with

\[ \max\{\text{rank}(\rho_1), \text{rank}(\rho_2)\} \leq k \leq \text{rank}(\rho_1) + \text{rank}(\rho_2) - 1 \quad (5.34) \]

and construct a \( \rho \in \mathcal{S}(\rho_1, \rho_2) \) with \( \text{rank}(\rho) = k \). We do this by expressing both \( \rho_1 \) and \( \rho_2 \) as an average of \( k \) pure states.

**Proposition 5.4.3.** Suppose \( k \) satisfies \((5.34)\), then there is a rank \( k \) solution \( \rho \in \mathcal{S}(\rho_1, \rho_2) \) of the form \( \rho = \frac{1}{k} \sum_{s=1}^{k} |u_s\rangle \langle u_s| \otimes |v_s\rangle \langle v_s| \) for some \( |u_j\rangle \in \mathbb{C}^{n_1} \) and \( |v_j\rangle \in \mathbb{C}^{n_2} \) for \( j = 1, \ldots, k \).

**Proof:** Without loss of generality, suppose \( n_1 \leq n_2 \). Suppose \( \rho_1 = \text{diag}(a_1, \ldots, a_{n_1}) \) and \( \rho_2 = \text{diag}(b_1, \ldots, b_{n_2}) \) are positive definite.

Let \( k \) be an integer such that \( n_2 \leq k \leq n_1 + n_2 - 1 \) and denote the principal \( k^{th} \) root of unity by \( \omega_k \). For any \( s = 1, \ldots, k \), define \( |u_s\rangle \in \mathbb{C}^m \) and \( |v_s\rangle \in \mathbb{C}^n \) such that

\[ |u_s\rangle = [\omega_k^{(j-1)(s-1)} \sqrt{a_j}]_j \quad \text{and} \quad |v_s\rangle = [\omega_k^{(l-1)(s-1)} \sqrt{b_l}]_l. \quad (5.35) \]

Then \( \rho_1 = \frac{1}{k} \sum_{s=1}^{k} |u_s\rangle \langle u_s| \) and \( \rho_2 = \frac{1}{k} \sum_{s=1}^{k} |v_s\rangle \langle v_s| \) (see for example, section 6.3.3 of [92]). It is clear that \( \rho = \sum_{s=1}^{k} \frac{1}{k} |u_s\rangle \langle u_s| \otimes |v_s\rangle \langle v_s| \in \mathcal{S}(\rho_1, \rho_2) \). Note that \( \rho = \frac{1}{k} PP^* \), where \( P \) is the \( n_1n_2 \times k \) matrix

\[
P = \begin{bmatrix} u_1 \otimes v_1 & \cdots & u_k \otimes v_k \end{bmatrix} = \text{diag} \left( \sqrt{a_1}, \ldots, \sqrt{a_{n_1}} \right) \otimes \text{diag} \left( \sqrt{b_1}, \ldots, \sqrt{b_{n_2}} \right)
\]

\[
= \begin{bmatrix} F \\ FD \\ \vdots \\ FD^{n_1-1} \end{bmatrix}
\]
and

\[
F = \begin{bmatrix}
1 & 1 & \cdots & 1 \\
1 & \omega_k & \cdots & \omega_k^{k-1} \\
& \vdots & \ddots & \vdots \\
1 & \omega_k^{(n_2-1)} & \cdots & \omega_k^{(k-1)(n_2-1)}
\end{bmatrix}, \quad D = \text{diag}(1, \omega_k, \omega_k^2, \ldots, \omega_k^{k-1}).
\]

Observe that \(FD^s\) consists of the \((1+s)^{th}\) up to the \((n_2+s)^{th}\) row of the discrete \(k \times k\) Fourier matrix, which is a unitary matrix. Hence, \(P\) has \(k\) linearly independent rows consisting of rows \(1, \ldots, n_2, 2n_2, 3n_2, \ldots, (k - n_2 + 1)n_2\). Counting all the linearly independent rows of \(P\), we get that \(\text{rank}(P) = \text{rank}(\rho) = k\).

In \cite{69}, it was proven that if there is a \(\rho \in S(\rho_1, \rho_2)\) with \(\text{rank} k\), then there is \(\rho \in S(\rho_1, \rho_2)\) with \(k \leq \text{rank}(\rho) \leq \text{rank}(\rho_1)\text{rank}(\rho_2)\). The following theorem is a consequence of this but we will give a constructive proof using Proposition 5.4.3 with the advantage of producing a separable global state.

**Theorem 5.4.4.** For any integer \(k\) such that \(\max\{\text{rank}(\rho_1), \text{rank}(\rho_2)\} \leq k \leq \text{rank}(\rho_1)\text{rank}(\rho_2)\), there exists \(\rho \in S(\rho_1, \rho_2)\) with \(\text{rank}(\rho) = k\).

**Proof:** Assume without loss of generality that \(n_1 \leq n_2\), \(\text{rank}(\rho_1) = n_1\), \(\text{rank}(\rho_2) = n_2\) and that \(\rho_1 = \text{diag}(a_1, \ldots, a_{n_1})\) and \(\rho_2 = \text{diag}(b_1, \ldots, b_{n_2})\). Thus, for \(n_2 \leq k \leq n_1n_2\), we need to construct a rank \(k\) solution \(\rho \in S(\rho_1, \rho_2)\).

**Case 1:** If \(k = n_1n_2\), then \(\rho = \rho_1 \otimes \rho_2\) has the desired properties.

**Case 2:** If \(n_2 \geq k < n_1n_2\), then by division algorithm, \(k = pn_2 + r\) for some \(1 \leq p < n_2\) and \(0 \leq r < n_1\).

**Case 2.1** If \(r \leq n_1 - p\), then \(\max\{n_2, n_1 - p\} = n_2 \leq n_2 + r \leq n_2 + n_1 - p\). Let

\[
\hat{\rho}_1 = \frac{1}{c}(0, \ldots, a_p, \ldots, a_{n_1}), \quad \text{where} \ c = a_p + \cdots + a_{n_1}
\]
Using Proposition 5.4.3, there is a \( n_2 + r \) density matrix \( \hat{\rho} \in \mathcal{S}(\hat{\rho}_1, \rho_2) \). Take \( \rho = \rho_A + \rho_B \), where

\[
\rho_A = \text{diag}(a_1, \ldots, a_{p-1}, 0, \ldots, 0) \otimes \rho_2 \quad \text{and} \quad \rho_B = c\hat{\rho} \in \mathcal{S}(\rho_1, \rho_2)
\]

Note that from the definition of \( \hat{\rho} \), we get that \( \text{range}(\rho_A) \cap \text{range}(\rho_B) = \{0\} \). Thus, \( \text{rank}(\rho) = \text{rank}(\rho_A) + \text{rank}(\rho_B) = (p-1)n_2 + n_2 + r = pn_2 + r = k \).

**Case 2.2** If \( r > n_1 - p \), then by division algorithm \( r = q(n_1 - p) + s \), where \( 1 \leq q < \frac{n_2}{n_1 - p} \) and \( 0 \leq s < n_1 - p \). Note that in this case, \( n_2 - p \leq n_1 - q \) since

\[
q(n_1 - p) < n_2 \implies (n_1 - p - 1) \leq q(n_1 - p - 1) \leq n_2 - q - 1.
\]

**Case 2.2.1** If \( n_1 - p = n_2 - q \) and \( s = 0 \), define \( c_1 = a_p + \cdots + a_{n_1} \) and \( c_2 = b_q + \cdots + b_{n_2} \) and

\[
\hat{\rho}_1 = \frac{1}{c_1} \text{diag}(0, \ldots, 0, a_p, \ldots, a_{n_1}) \quad \text{and} \quad \hat{\rho}_2 = \frac{1}{c_2} \text{diag}(0, \ldots, 0, b_q, \ldots, b_{n_2}).
\]

Note that \( n_1 + n_2 - q + p + 1 = \text{rank}(\hat{\rho}_1) + \text{rank}(\hat{\rho}_2) - 1 \). Thus, using Proposition 5.4.3, there exists \( \hat{\rho} \in \mathcal{S}(\hat{\rho}_1, \hat{\rho}_2) \). Take \( \rho = \rho_A + \rho_B + \rho_C \), where

\[
\rho_A = (\rho_1 - c_1\hat{\rho}_1) \otimes \rho_2, \quad \rho_B = c_2\hat{\rho}_1 \otimes (\rho_2 - c_2\hat{\rho}_2), \quad \rho_C = c_1c_2\hat{\rho}
\]

Since \( \text{range}(\rho_A) \cap \text{range}(\rho_B) = \text{range}(\rho_A) \cap \text{range}(\rho_C) = \text{range}(\rho_B) \cap \text{range}(\rho_C) = \{0\} \), we get \( \text{rank}(\rho) = \text{rank}(\rho_A) + \text{rank}(\rho_B) + \text{rank}(\rho_C) = (p-1)n_2 + (n_1 - p + 1)(q-1) + (n_1 + n_2 - q - p + 1) \), which is equal to the desired rank \( k = pn_2 + q(n_1 - p) \).

**Case 2.2.2** For the remaining case, let \( c_1 = a_p + \cdots + a_{n_1} \) and \( c_2 = b_{q+1} + \cdots + b_{n_2} \) and

\[
\hat{\rho}_1 = \frac{1}{c_1} \text{diag}(0, \ldots, 0, a_p, \ldots, a_{n_1}) \quad \text{and} \quad \hat{\rho}_2 = \frac{1}{c_2} \text{diag}(0, \ldots, 0, b_{q+1}, \ldots, b_{n_2}).
\]

Then, \( \max\{\text{rank}(\hat{\rho}_1), \text{rank}(\hat{\rho}_2)\} = \max\{n_1 - p + 1, n_2 - q\} \leq n_2 - q + s \leq n_1 + n_2 - p - q \). By Proposition 5.4.3, there is a rank \( n_2 - q + s \) density matrix \( \hat{\rho} \in \mathcal{S}(\hat{\rho}_1, \hat{\rho}_2) \). Take \( \rho = \rho_A + \rho_B + \rho_C \),
where
\[
\rho_A = (\rho_1 - c_1 \hat{\rho}_1) \otimes \rho_2, \quad \rho_B = c_2 \hat{\rho}_1 \otimes (\rho_2 - c_2 \hat{\rho}_2), \quad \rho_C = c_1 c_2 \hat{\rho},
\]
So that \(\text{range}(\rho_A) \cap \text{range}(\rho_B) = \text{range}(\rho_A) \cap \text{range}(\rho_C) = \text{range}(\rho_B) \cap \text{range}(\rho_C) = \{0\}\).
Hence, \(\text{rank}(\rho) = \text{rank}(\rho_A) + \text{rank}(\rho_B) + \text{rank}(\rho_C) = (p-1)n_2 + (n_1-p+1)(q) + (n_1-q+s)\), which is equal to the desired rank \(k = pn_2 + q(n_1 - p) + s\).

\[\square\]

Once again, when \(\min\{\text{rank}(\rho_1), \text{rank}(\rho_2)\} = 1\), we get the trivial case that \(\mathcal{S}(\rho_1, \rho_2) = \{\rho_1 \otimes \rho_2\}\). Now, what remains to be seen is whether or not we can find a solution with rank \(k\) whenever \(\text{rank}(\rho_1), \text{rank}(\rho_2) \geq 2\). In the next algorithm, we present another scheme to find a low rank solution \(\rho \in \mathcal{S}(\rho_1, \rho_2)\) using the following known result in [46].

**Theorem 5.4.5.** Suppose \(a_1 \geq b_1 \geq a_2 \geq b_2 \geq \cdots \geq a_k \geq b_k \geq 0\). Define \(|d\rangle = [d_s] \in \mathbb{R}^k\) such that

\[
d_s = \begin{cases} 
0 & \text{if } a_s = 0 \text{ or } a_j = a_s \text{ for some } j \neq s \\
\prod_{j=1}^{n} \left( \frac{b_j - a_s}{a_j - a_s} \right) & \text{otherwise}
\end{cases}
\]

Then \(\text{diag}(a_1, \ldots, a_k) - |d\rangle\langle d|\) has eigenvalues \(b_1, \ldots, b_k\).

**Algorithm 5.4.6.** Scheme to find \(\rho \in \mathcal{S}(\rho_1, \rho_2)\) with rank(\(\rho\)) \(\leq \max\{\text{rank}(\rho_1), \text{rank}(\rho_2)\}\).

**Step 1:** Set \(r = 1\) and \(A_1 = \rho_1\) and \(B_1 = \rho_2\).

**Step 2:** If \(A_r = 0\), then proceed to step 3. Otherwise do the following subroutines.

**Step 2.1:** Find unitary \(U_r, V_r\) such that

\[
A_r = U_r(S_1 \oplus \cdots \oplus S_p \oplus T_1 \oplus T_q \oplus L_a)U_r^* \quad \text{and} \quad B_r = V_r(\hat{S}_1 \oplus \cdots \oplus \hat{S}_p \oplus \hat{T}_1 \oplus \hat{T}_q \oplus \hat{L}_b)V_r^*
\]

where
1. \( T_j = \text{diag}(c_{j1}, \ldots, c_{jt_j}) \) and \( \tilde{T}_j = \text{diag}(d_{j1}, \ldots, d_{jt_j}) \) satisfy \( d_{j1} \geq c_{j1} \geq \cdots \geq d_{jt_j} \geq c_{jt_j} \),

2. \( S_\ell = \text{diag}(\hat{c}_{\ell 1}, \ldots, \hat{c}_{\ell s_\ell}) \) and \( \tilde{S}_\ell = \text{diag}(\hat{d}_{\ell 1}, \ldots, \hat{d}_{\ell s_\ell}) \) satisfy \( \hat{c}_{\ell 1} \geq \hat{d}_{\ell 1} \geq \cdots \geq \hat{c}_{\ell s_\ell} \geq \hat{d}_{\ell s_\ell} \), and

3. \( L_\alpha \) is either empty or is a zero block and \( L_b \) is either empty or is a zero block.

**Step 2.2:** Use Lemma 5.4.5 to find \( |x_j\rangle \in \mathbb{R}^{s_j} \) such that the eigenvalues of \( S_j - |x_j\rangle \langle x_j| \) are the eigenvalues of \( \tilde{S}_j \). Similarly, find \( |y_j\rangle \in \mathbb{R}^{t_j} \) such that the eigenvalues of \( \tilde{T}_j - |y_j\rangle \langle y_j| \) are the same as that of \( T_j \).

**Step 2.3:** Let

\[
C_r = U_r \left( (S_1 - |x_1\rangle \langle x_1|) \oplus \cdots \oplus (S_p - |x_p\rangle \langle x_p|) \oplus T_1 \oplus \cdots \oplus T_q \oplus 0 \right) U_r^*
\]

and

\[
\tilde{C}_r = V_r \left( \tilde{S}_1 \oplus \cdots \oplus \tilde{S}_p \oplus (\tilde{T}_1 - |y_1\rangle \langle y_1|) \oplus \cdots \oplus (\tilde{T}_q - |y_q\rangle \langle y_q|) \oplus 0 \right) V_r^*
\]

and set \( A_{r+1} = A_r - C_r \) and \( B_{r+1} = B_r - \tilde{C}_r \). Repeat step 2, taking \( r \leftarrow r + 1 \).

**Step 3:** Suppose the above process stops at \( r = k + 1 \). For \( s = 1, \ldots, k \), find \( \hat{U}_s \) and \( \hat{V}_s \) such that

\[
C_s = \hat{U}_s \text{diag}(\alpha_{s1}, \ldots, \alpha_{sr_s}, 0, \ldots) \hat{U}_s^* \quad \text{and} \quad \tilde{C}_s = \hat{V}_s \text{diag}(\alpha_{s1}, \ldots, \alpha_{sr_s}, 0, \ldots) \hat{V}_s^*
\]

Define \( \rho = |w_1\rangle \langle w_1| + \cdots + |w_k\rangle \langle w_k| \), where \( |w_s\rangle = \sum_{j=0}^{r_s-1} \sqrt{\alpha_{sj}} |\tilde{U}_s| j \rangle \otimes |\tilde{V}_s| j \rangle \).

**Proposition 5.4.7.** The procedures in algorithm 5.4.6 are well-defined and produces \( \rho \in S(\rho_1, \rho_2) \) with \( \text{rank}(\rho) = k \leq \max\{\text{rank}(\rho_1), \text{rank}(\rho_2)\} \) for any given \( \rho_1 \in \mathcal{D}_{n_1} \) and \( \rho_2 \in \mathcal{D}_{n_2} \). More specifically, Step 2 produces \( C_1, \ldots, C_k \in \text{PSD}_{n_1} \) and \( \tilde{C}_1, \ldots, \tilde{C}_k \in \text{PSD}_{n_2} \) such that

1. \( k \leq \max\{\text{rank}(\rho_1), \text{rank}(\rho_2)\} \),

2. \( C_r \) and \( \tilde{C}_r \) are isospectral for \( r = 1, \ldots, k \),

3. \( \rho_1 = C_1 + \cdots + C_k, \rho_2 = \tilde{C}_1 + \cdots + \tilde{C}_k \), and;
4. Suppose \( \text{eig}^k(\rho_1) = (a_1, \ldots, a_{n_1}) \) and \( \text{eig}^k(\rho_2) = (b_1, \ldots, b_{n_2}) \). If we can find distinct indices \( j_1, \ldots, j_s \) and distinct \( \ell_1, \ldots, \ell_s \) such that either

\[
a_{j_1} \geq b_{\ell_1} \geq \cdots \geq a_{j_s} \geq b_{\ell_s} > 0 \quad \text{or} \quad b_{\ell_1} \geq a_{j_1} \geq \cdots \geq b_{\ell_s} \geq a_{j_s} > 0,
\]

then the solution \( \rho \) obtained has rank at most \( \max\{\text{rank}(\rho_1) - s + 1, \text{rank}(\rho_2) - s + 1\} \).

**Proof:** Note that the construction of \( A_r \) and \( B_r \) in step 2.3 of Algorithm 5.4.6, guarantees that for every iteration \( r \), \( A_r \) and \( B_r \) are positive semidefinite and \( \text{tr}(A_r) = \text{tr}(B_r) \). Furthermore,

\[
\text{rank}(A_{r+1}) = \text{rank}(A_r) - \left( \sum_{j=1}^{p} \text{rank}(S_j) \right) - \left( \sum_{\ell=1}^{q} \text{rank}(T_{\ell}) \right) + p \quad (5.36)
\]

\[
\text{rank}(B_{r+1}) = \text{rank}(B_r) - \left( \sum_{j=1}^{p} \text{rank}(S_j) \right) - \left( \sum_{\ell=1}^{q} \text{rank}(T_{\ell}) \right) + q \quad (5.37)
\]

Since \( \text{tr}(A_r) = \text{tr}(B_r) \) and \( A_r, B_r \) are both positive semidefinite, then there exists eigenvalues \( c, \hat{c} \) of \( A_r \) and eigenvalues \( d, \hat{d} \) of \( B_r \) such that \( \hat{c} \geq \hat{d} \) and \( d \geq c \) so that \( p, q \geq 1 \). Hence, \( \text{rank}(A_{r+1}) < \text{rank}(A_r) \) and \( \text{rank}(B_{r+1}) < \text{rank}(B_r) \). This guarantees that the process terminates after finitely many steps. Moreover, for some \( k \leq \max\{\text{rank}(\rho_1), \text{rank}(\rho_2)\} \), we get \( 0 = A_{k+1} = \rho_1 - C_1 - C_2 - \cdots - C_k \) and consequently, \( 0 = B_{k+1} = \rho_2 - \hat{C}_1 - \hat{C}_2 - \cdots - \hat{C}_k \). By Theorem 5.4.5, \( C_j \) and \( \hat{C}_j \) are isospectral and positive semidefinite.

If \( a_{j_1} \geq b_{\ell_1} \geq \cdots \geq a_{j_s} \geq b_{\ell_s} > 0 \) (or \( b_{\ell_1} \geq a_{j_1} \geq \cdots \geq b_{\ell_s} \geq a_{j_s} > 0 \)) for some distinct indices \( j_1, \ldots, j_s \) and distinct \( \ell_1, \ldots, \ell_s \), then \( \rho_1 = C_1 + A_1 \) and \( \rho_2 = \hat{C}_1 + B_1 \) where \( \text{rank}(A_1) \leq \text{rank}(\rho_1) - s \) and \( \text{rank}(B_1) \leq \text{rank}(\rho_2) - s \) using equations (5.36) and (5.37). By Theorem 5.4.5, there is a rank one \( \sigma \in PSD^{n_1 n_2} \) such that \( \text{tr}_1(\sigma) = C_1 \) and \( \text{tr}_2(\sigma) = \hat{C}_1 \). It will also follow from Proposition 5.4.3 that we can find \( \mu \in PSD^{n_1 n_2} \) such that \( \text{tr}_1(\mu) = A_1 \) and \( \text{tr}_2(\sigma) = B_1 \) such that \( \text{rank}(\mu) = \max\{\text{rank}(A_1), \text{rank}(B_1)\} \). Thus \( \rho = \sigma + \mu \in S(\rho_1, \rho_2) \) has rank at most \( \max\{\text{rank}(\rho_1) - s + 1, \text{rank}(\rho_2) - s + 1\} \).

Finally, we present one more scheme to find a low rank solution \( \rho \in S(\rho_1, \rho_2) \). Similar to
Algorithm 5.4.6, we find $\rho$ by first writing

$$
\rho_1 = C_1 + \cdots + C_k \quad \text{and} \quad \rho_2 = \tilde{C}_1 + \cdots + \tilde{C}_k
$$

for $k$ pairs $(C_1, \tilde{C}_1), \ldots, (C_k, \tilde{C}_k)$, of isospectral positive semidefinite matrices such that $k \leq \max \{\text{rank}(\rho_1), \text{rank}(\rho_2)\}$. In fact, these pairs can be chosen so that we can construct a $\rho \in S(\rho_1, \rho_2)$ whose nonzero eigenvalues are given by $\lambda_j = \text{tr}(C_i) = \text{tr}(\tilde{C}_j)$ for $j = 1, \ldots, k$. Furthermore, this solution $\rho$ satisfies

$$
||\rho||_\infty = \max_{\sigma \in S(\rho_1, \rho_2)} ||\sigma||_\infty,
$$

where $|| \cdot ||_\infty$ denotes the operator/spectral norm.

**Algorithm 5.4.8.** Scheme to find $\rho \in S(\rho_1, \rho_2)$ with $\text{rank}(\rho) \leq \max \{\text{rank}(\rho_1), \text{rank}(\rho_2)\}$.

**Step 1:** Suppose $\rho_1 = U \text{diag}(a_1, \ldots, a_{n_1}) U^*$ and $\rho_2 = V \text{diag}(b_1, \ldots, b_{n_2}) V^*$. Set $r = 0$ and define

$$
a_j^{(0)} = a_j \quad \text{for} \quad j = 1, \ldots, n_1 \quad \text{and} \quad b_\ell^{(0)} = b_\ell \quad \text{for} \quad \ell = 1, \ldots, n_2
$$

**Step 2:** If $\sum_{j=1}^{n_1} a_j^{(r)} = 0$, then stop. Otherwise, set $r \leftarrow r + 1$. Find permutations $s_r$ and $\tilde{s}_r$ such that

$$
a_{s_r(1)}^{(r)} \geq \cdots \geq a_{s_r(n_1)}^{(r)} \quad \text{and} \quad b_{\tilde{s}_r(1)}^{(r)} \geq \cdots \geq b_{\tilde{s}_r(n_2)}^{(r)}.
$$

Let $P_r$ and $\tilde{P}_r$ the permutation matrices satisfying

$$
P_r \text{diag}(a_1^{(r)}, \ldots, a_{n_1}^{(r)}) P_r^T = \text{diag}(a_{s_r(1)}^{(r)}, \ldots, a_{s_r(n_1)}^{(r)})
$$

$$
\tilde{P}_r \text{diag}(b_1^{(r)}, \ldots, b_{n_2}^{(r)}) \tilde{P}_r^T = \text{diag}(b_{\tilde{s}_r(1)}^{(r)}, \ldots, b_{\tilde{s}_r(n_2)}^{(r)})
$$

Then, define

$$
C_r = U P_r^T \text{diag}(c_{r_1}, \ldots, c_{r_{n_1}}) P_r U^* \quad \text{and} \quad \tilde{C}_r = V \tilde{P}_r^T \text{diag}(c_{r_1}, \ldots, c_{r_{n_2}}) \tilde{P}_r V^*,
$$

121
where \( c_{rs} = \min\{a_{g_{r}(s)}, b_{h_{r}(s)}\} \) if \( s \in \{1, \ldots, \min\{n_1, n_2\}\} \) and \( c_{rs} = 0 \) otherwise. Then set
\[
a^{(r+1)}_j = a^{(r)}_j - c_{s_j^{(r)},j} \quad \text{for} \quad j = 1, \ldots, n_1 \quad \text{and} \quad b^{(r+1)}_\ell = b^{(r)}_\ell - c_{s_{\nu}(\ell),\ell} \quad \text{for} \quad \ell = 1, \ldots, n_2
\]
and repeat step 2 for \( r \leftarrow r + 1 \).

**Step 3:** Suppose the above process terminates at \( r = k + 1 \). For \( s = 1, \ldots, k \), define
\[
|w_s\rangle = \sum_{j=1}^{\min\{n_1, n_2\}} \sqrt{c_{j}} U|\mathcal{g}_s(j) - 1\rangle \otimes V|\mathcal{h}_s(j) - 1\rangle \quad \text{and} \quad \rho = |w_1\rangle\langle w_1| + \cdots + |w_k\rangle\langle w_k|.
\]

**Proposition 5.4.9.** Let \( \rho_1 \in \mathcal{D}_{n_1} \) and \( \rho_2 \in \mathcal{D}_{n_2} \). The procedures in Algorithm 5.4.8 are well-defined and produces \( \rho \in \mathcal{S}(\rho_1, \rho_2) \). More specifically, the algorithm constructs \( C_1, \ldots, C_k \in \text{PSD}_{n_1} \) and \( \tilde{C}_1, \ldots, \tilde{C}_k \in \text{PSD}_{n_2} \) such that

1. \( k \leq \max\{\text{rank}(\rho_1), \text{rank}(\rho_2)\} \)
2. \( C_j \) and \( \tilde{C}_j \) are isospectral for \( j = 1, \ldots, k \).
3. \( \rho_1 = C_1 + \cdots + C_k \) and \( \rho_2 = \tilde{C}_1 + \cdots + \tilde{C}_k \)
4. If \( |w_1\rangle, \ldots, |w_k\rangle \in \mathbb{C}^{n_1 n_2} \) are the vectors defined in Step 3, then \( \langle w_s|w_t\rangle = \delta_{st}\text{tr}(C_s) \).
5. \( \|\rho\|_{\infty} = \text{tr}(C_1) = \max_{\sigma \in \mathcal{S}(\rho_1, \rho_2)} \|\sigma\|_{\infty} \)

**Proof:** Assume without loss of generality that \( n_1 \leq n_2 \) and
\[
\rho_1 = \text{diag}(a_1, \ldots, a_{n_1}) \quad \text{and} \quad \rho_2 = \text{diag}(b_1, \ldots, b_{n_2}),
\]
where \( a_1 \geq a_2 \geq \cdots \geq a_{n_1} > 0 \) and \( b_1 \geq b_2 \geq \cdots \geq b_{n_2} > 0 \). For any \( j = 1, \ldots, n_1 \), define \( c_j = \min\{a_j, b_j\} \) and \( c_{n_1+1} = \cdots = c_{n_2} = 0 \) and define \( C_1 = \text{diag}(c_1, \ldots, c_{n_1}) \) and \( \tilde{C}_1 = \text{diag}(c_1, \ldots, c_{n_2}) \). Clearly, \( \rho_1 - C_1 \) and \( \rho_2 - \tilde{C}_1 \) are positive semidefinite. Since \( \text{tr}(\rho_1) = \text{tr}(\rho_2) \), there must exists indices \( 1 \leq j_1, j_2 \leq n_1 \) such that \( c_{j_1} = a_{j_1} \) and \( c_{j_2} = b_{j_2} \). This means that \( \text{rank}(\rho_1 - C_1) < \text{rank}(\rho_1) \) and \( \text{rank}(\rho_2 - \tilde{C}_1) < \text{rank}(\rho_2) \). We can replace \( \rho_1 \) and \( \rho_2 \) by \( \rho_1 - C_1 \)
and $\rho_2 - \tilde{C}_1$ and repeat the above process until both matrices become zero. This process will take at most $k = \max\{\text{rank}(\rho_1), \text{rank}(\rho_2)\}$ steps because the rank of $\rho_1$ and $\rho_2$ are reduced by at least one in each step. At the end of this process, we will be able to write $\rho_1$ and $\rho_2$ as $\rho_1 = C_1 + \cdots + C_k$ and $\rho_2 = \tilde{C}_1 + \cdots + \tilde{C}_k$ such that for each $j$,

$$C_j = \text{diag}(c_{j_1}, \ldots, c_{j_n})$$

for some permutation $s_j$. Note that in this scheme, it is true that if $c_{t_j} \neq 0$, either $c_{s_j} = 0$ for all $s \geq t$ or $c_{s_{s_t^{-1}(j)}} = 0$ for all $s \geq t$. That is, $c_{t_j}$ completes the set of nonzero summands for either one of the eigenvalues of $\rho_1$ or one of the eigenvalues of $\rho_2$.

Let $\rho = |w_1\rangle\langle w_1| + \cdots + |w_k\rangle\langle w_k|$, where $w_t = \sum_{j=1}^{n_1} \sqrt{c_{t_j}} |j - 1\rangle \otimes |s^{-1}(j) - 1\rangle$. Now,

$$\langle w_t|w_s\rangle = \sum_{j,\ell=1}^{n_1} \sqrt{c_{t_j} c_{s_{s_{t^{-1}(j)}}}} \langle j - 1| \ell - 1 \rangle \otimes \langle s_{s_t^{-1}(j)} - 1| s_{s_{t^{-1}(j)}}^{-1}(\ell) - 1 \rangle = \sum_{j=1}^{n_1} \sqrt{c_{t_j} c_{s_{s_{t^{-1}(j)}}}}$$

Note that if $s > t$ and $c_{t_j} \neq 0$, then $c_{s_j} = c_{s_{s_{t^{-1}(j)}}} = 0$. If $t > s$ and $c_{s_j} \neq 0$, then $c_{t_j} = c_{s_{s_{t^{-1}(j)}}} = 0$. Thus, $w_1, \ldots, w_k$ form an orthogonal basis. This means that $\lambda_j = \langle w_j|w_j\rangle = c_{j_1} + \cdots + c_{j_n}$, for $j = 1, \ldots, k$ (together with $n_1n_2 - k$ more zeros) are the eigenvalues of $\rho$.

Now, suppose $\sigma \in S(\rho_1, \rho_2)$ has spectral decomposition $\sigma = s_1|x_1\rangle\langle x_1| + \cdots + s_N|x_N\rangle\langle x_N|$. Then

$$\rho_1 = s_1 \text{tr}_2(|x_1\rangle\langle x_1|) + \cdots + s_N \text{tr}_2(|x_N\rangle\langle x_N|)$$

and

$$\rho_2 = s_1 \text{tr}_1(|x_1\rangle\langle x_1|) + \cdots + s_N \text{tr}_1(|x_N\rangle\langle x_N|)$$

Hence $\rho_1 - s_1 \text{tr}_2(|x_1\rangle\langle x_1|)$ and $\rho_2 - s_1 \text{tr}_1(|x_1\rangle\langle x_1|)$ are positive semidefinite. Let $c_1 \geq \cdots \geq c_k$ be the nonzero eigenvalues of $s_1 \text{tr}_2(|x_1\rangle\langle x_1|)$, which are also the nonzero eigenvalues of $s_1 \text{tr}_1(|x_1\rangle\langle x_1|)$. Then using Lidskii’s inequalities, we get $c_j \leq \min\{a_j, b_j\}$ for $j = 1, \ldots, k$. Thus,

$$||\sigma||_\infty = s_1 = \sum_{j=1}^{k} c_j \leq \sum_{j=1}^{k} \min\{a_j, b_j\} \leq \sum_{j=1}^{\min\{n_1, n_2\}} \min\{a_j, b_j\} = ||\rho||_\infty.$$

123
This also follows from Theorem 6.3.1 of [54] using algebraic combinatorics.

Algorithm 5.4.8 can produce a solution $\rho$ that has rank less than $\min\{\text{rank}(\rho_1), \text{rank}(\rho_2)\}$, but usually does not give the minimum rank. Take for example the case

$$
\rho_1 = \text{diag} \left( \frac{7}{10}, \frac{3}{10} \right) \quad \text{and} \quad \rho_2 = \text{diag} \left( \frac{3}{5}, \frac{1}{5}, \frac{1}{5} \right).
$$

There is no $\rho \in S(\rho_1, \rho_2)$ with rank 1, but there is a rank 2 solution given by $\rho = |w_1\rangle\langle w_1| + |w_2\rangle\langle w_2|$, where

$$
|w_1\rangle = \sqrt{\frac{3}{5}} |0\rangle \otimes |0\rangle + \sqrt{\frac{1}{10}} |1\rangle \otimes |1\rangle \quad \text{and} \quad |w_2\rangle = \sqrt{\frac{1}{10}} |0\rangle \otimes |1\rangle + \sqrt{\frac{1}{5}} |1\rangle \otimes |2\rangle
$$

However, Algorithm 5.4.8 will produce a rank 3 solution.

Note that the solutions obtained from Proposition 5.4.3 and Algorithms 5.4.6, 5.4.8 can be utilized as the starting point when implementing Algorithm 5.4.1 to find a solution with lower rank. Here, we note that the solution obtained in Algorithm 5.4.8 has relatively low von Neumann entropy since it has maximal spectral norm, that is, its largest eigenvalue is as close to 1 as possible making it a good pure state approximation. However, as will be seen in the numerical results in the next subsection, it is not guaranteed to have minimal von Neumann entropy.

### 5.4.2 Numerical Experiments

In this subsection, we give some examples to illustrate the effectiveness of Proposition 5.4.3, Algorithms 5.4.1, 5.4.6 and 5.4.8 to construct low rank elements of $S(\rho_1, \rho_2)$. All experiments are performed in MATLAB R2015a on a PC with an Intel Core i7 processor at 2.40GHz with machine precision $\varepsilon = 2.22 \times 10^{-16}$. The programs are available at

http://cklixx.people.wm.edu/mathlib/projection/.

Let $r = \text{rank}(\rho)$ and $err = \max\{||\rho_1 - \text{tr}_2(\rho)||, ||\rho_2 - \text{tr}_1(\rho)||\}$. Denote the maximum and minimum eigenvalues of $\rho$ by $\lambda_M$ and $\lambda_\mu$, respectively; and the Von Neumman entropy of $\rho$ by

124
ent. The following table illustrates the performance of each algorithm.

**Example 5.4.10.** We consider \( \rho_1 \in \mathcal{D}_3 \) and \( \rho_2 \in \mathcal{D}_4 \) with eigenvalues

\[
eig^k(\rho_1) = (0.5951, 0.2341, 0.1708) \quad \text{and} \quad \eig^k(\rho_2) = (0.6124, 0.1926, 0.1654, 0.0296)
\]

<table>
<thead>
<tr>
<th>Alg.</th>
<th>( r )</th>
<th>CPU-time</th>
<th>( \text{err} )</th>
<th>( \lambda_\mu )</th>
<th>( \lambda_M )</th>
<th>( \text{ent} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.4.3</td>
<td>4</td>
<td>0.002s</td>
<td>3.54294e-17</td>
<td>-6.00329e-17</td>
<td>0.399619</td>
<td>1.27929</td>
</tr>
<tr>
<td>5.4.6</td>
<td>3</td>
<td>0.006s</td>
<td>1.11022e-16</td>
<td>-1.48157e-16</td>
<td>0.9313</td>
<td>0.297223</td>
</tr>
<tr>
<td>5.4.8</td>
<td>3</td>
<td>0.004s</td>
<td>1.11022e-16</td>
<td>-4.1612e-17</td>
<td>0.9531</td>
<td>0.215848</td>
</tr>
</tbody>
</table>

TABLE 5.7: Low rank solutions obtained using Algorithms 5.4.3, 5.4.5, and 5.4.8

<table>
<thead>
<tr>
<th>( X_0 )</th>
<th>( # \text{ iter} )</th>
<th>CPU-time</th>
<th>( \text{err} )</th>
<th>( \lambda_\mu )</th>
<th>( \lambda_M )</th>
<th>( \text{ent} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alg. 5.4.6</td>
<td>2</td>
<td>1336</td>
<td>0.54s</td>
<td>9.34747e-16</td>
<td>-4.16498e-17</td>
<td>0.9017</td>
</tr>
<tr>
<td>Alg. 5.4.8</td>
<td>2</td>
<td>3103</td>
<td>1.266s</td>
<td>9.85657e-16</td>
<td>-5.19103e-17</td>
<td>0.9531</td>
</tr>
</tbody>
</table>

TABLE 5.8: Low rank solution from Algorithm 5.4.1 using the solutions from Algorithms 5.4.3 and 5.4.5 as starting point.

Table 5.7 shows the results we get when using Proposition 5.4.3 and Algorithms 5.4.6 and 5.4.8. Using Algorithm 5.4.1, we determine if we can find a solution of rank 2, \ldots, \text{rank}(X_0) - 1, where \( X_0 \) is a solution obtained from one of the algorithms above. The solutions we obtained are shown in Table 5.8.

Note that in this case, the solution obtained by Algorithm 5.4.1 using the solution from Algorithm 5.4.8 as initial point, has minimum entropy in \( S(\rho_1, \rho_2) \). This is because \( \rho \) is rank 2 and the largest eigenvalue of \( \rho \) is the maximum possible eigenvalue of any element of \( S(\rho_1, \rho_2) \).

**Example 5.4.11.** In this example, we consider \( \rho_1 \in \mathcal{D}_6, \rho_2 \in \mathcal{D}_8 \) such that

\[
eig^k(\rho_1) = (0.8213, 0.1234, 0.0553) \quad \text{and} \quad \eig^k(\rho_2) = (0.5720, 0.3068, 0.1000, 0.0189, 0.0020, 0.0003).
\]

**Example 5.4.12.** In this example, we consider \( \rho_1 \in \mathcal{D}_6, \rho_2 \in \mathcal{D}_8 \) such that

\[
eig^k(\rho_1) = (0.2272, 0.2136, 0.1946, 0.1474, 0.1341, 0.0831) \quad \text{and} \quad \eig^k(\rho_2) = (0.2399, 0.1699, 0.1638, 0.1463, 0.1246, 0.0851, 0.0407, 0.0297).
\]

125
<table>
<thead>
<tr>
<th>Alg.</th>
<th>r</th>
<th>CPU-time</th>
<th>err</th>
<th>$\lambda_\mu$</th>
<th>$\lambda_M$</th>
<th>ent</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.4.3</td>
<td>6</td>
<td>0.003s</td>
<td>8.9182e-16</td>
<td>-4.93499e-17</td>
<td>0.469983</td>
<td>1.19924</td>
</tr>
<tr>
<td>5.4.6</td>
<td>4</td>
<td>0.005s</td>
<td>3.31468e-16</td>
<td>-6.27654e-17</td>
<td>0.690947</td>
<td>0.632879</td>
</tr>
<tr>
<td>5.4.8</td>
<td>6</td>
<td>0.004s</td>
<td>2.78333e-16</td>
<td>-5.4791e-17</td>
<td>0.750675</td>
<td>0.755308</td>
</tr>
</tbody>
</table>

TABLE 5.9: Low rank solutions obtained using Proposition 5.4.3 and Algorithms 5.4.5, and 5.4.8

<table>
<thead>
<tr>
<th>$X_0$</th>
<th>r</th>
<th># iter</th>
<th>CPU-time</th>
<th>err</th>
<th>$\lambda_\mu$</th>
<th>$\lambda_M$</th>
<th>ent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alg. 5.4.8</td>
<td>3</td>
<td>76933</td>
<td>44.25s</td>
<td>9.90465e-16</td>
<td>-5.79165e-17</td>
<td>0.729479</td>
<td>0.736448</td>
</tr>
<tr>
<td>Alg. 5.4.6</td>
<td>2</td>
<td>100000</td>
<td>63.5203s</td>
<td>2.26889e-08</td>
<td>-1.44764e-16</td>
<td>0.690947</td>
<td>0.618341</td>
</tr>
<tr>
<td>Alg. 5.4.6</td>
<td>3</td>
<td>6707</td>
<td>4.39s</td>
<td>9.83117e-16</td>
<td>-6.84736e-17</td>
<td>0.690947</td>
<td>0.631907</td>
</tr>
</tbody>
</table>

TABLE 5.10: Low rank solutions obtained Algorithm 5.4.1 utilizing the solutions from Proposition 5.4.3 and Algorithms 5.4.5, and 5.4.8 as starting point.

5.5 Bipartite States with Prescribed Reduced States and Low Entropy **

In this section, we are interested in finding $\rho \in S(\rho_1, \rho_2)$, as defined in Section 5.4, attaining certain extreme functional values for a given scalar function $f$ on states. Our result will cover the case when $f(\rho)$ is the von-Neumann entropy of $\rho$ defined by

$$H(\rho) = -\text{tr}(\rho \log \rho) = - \sum \lambda_j \log(\lambda_j),$$

(5.38)

where $\lambda_j$ are the eigenvalues of $\rho$, and $x \log x = 0$ if $x = 0$, and the Rényi entropy defined by

$$H_\alpha(\rho) = \frac{1}{1 - \alpha} \log \text{tr}(\rho^\alpha) = \frac{1}{1 - \alpha} \log \left( \sum \lambda_j^\alpha \right) \quad \text{for } \alpha \geq 0.$$

(5.39)

Note that $\rho_1 \otimes \rho_2 \in S(\rho_1, \rho_2)$ has maximum von Neumann entropy by the subadditivity property of von Neumann entropy. So, we focus on searching for $\rho \in S(\rho_1, \rho_2)$ with minimum

**The material in this section is also part of [32].
algorithm r CPU-time err $\lambda_{\mu}$ $\lambda_M$ ent
5.4.3 8 0.005s 2.56989e-16 -3.91005e-17 0.151124 2.0642
5.4.6 3 0.014s 4.38087e-16 -1.36117e-16 0.840737 0.515135
5.4.8 4 0.017s 3.08212e-16 -1.05048e-16 0.914875 0.308127

TABLE 5.11: Low rank solutions obtained using Proposition 5.4.3 and Algorithms 5.4.5, and 5.4.8 as starting point.

<table>
<thead>
<tr>
<th>$X_0$</th>
<th>r</th>
<th># iter</th>
<th>CPU-time</th>
<th>err</th>
<th>$\lambda_{\mu}$</th>
<th>$\lambda_M$</th>
<th>ent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alg. 5.4.8</td>
<td>3</td>
<td>26770</td>
<td>45.955s</td>
<td>8.97652e-16</td>
<td>-8.7338e-17</td>
<td>0.914681</td>
<td>0.308847</td>
</tr>
</tbody>
</table>

TABLE 5.12: Low rank solutions obtained Algorithm 5.4.1 utilizing the solutions from Proposition 5.4.3 and Algorithms 5.4.5, and 5.4.8 as starting point.

entropy, that is, we are interested in the following minimization problem

$$\min_{\rho \in \text{PSD}_{n_1n_2} \cap S_2} -\text{tr}(\rho \log \rho), \quad (5.40)$$

where $S_2$ is as define in equation 5.22 for the bipartite case. That is

$$S_2 = \{\rho \in \mathcal{H}_{n_1n_2} : \text{tr}_1(\rho) = \rho_2 \in \mathcal{D}_{n_2} \text{ and } \text{tr}_2(\rho) = \rho_1 \in \mathcal{D}_1\} \quad (5.41)$$

Since $\text{PSD}_{n_1n_2}$ and $S_2$ are closed convex sets, then the set $\text{PSD}_{n_1n_2} \cap S_2$ is also a closed convex set. Now we use the nonmonotone spectral projected gradient (NSPG) method to solve the minimization problem (5.40), which was proposed in Birgin et al [10], on minimizing a continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ on a nonempty closed convex set $M$. As it is quite simple to implement and very effective for large-scale problem, it has been extensively studied in the past years (see [58, 60] and their references for details). The NSPG method has the form $x_{k+1} = x_k + \alpha_k d_k$, where $d_k$ is chosen to be $\text{proj}_M(x_k - t_k \nabla f(x_k)) - x_k$ with $t_k > 0$ a precomputed scalar. The direction $d_k$ is guaranteed to be a descent direction ( [9, Lemma 2.1]) and the step length $\alpha_k$ is selected by a nonmonotone linear search strategy. The key problems to use NSPG method to solve (5.40) are how to compute the gradient of the objective function $f(\rho) = -\text{tr}\rho \log \rho$ and the projection operator $\Phi_{\text{PSD}_{n_1n_2} \cap S_2}(Z)$ of $Z$ onto the set $\text{PSD}_{n_1n_2} \cap S_2$. 127
Such problems is addressed in the following.

For any function \( f : \mathbb{R} \rightarrow \mathbb{R} \), one can extend it to \( f : \mathcal{H}_n \rightarrow \mathcal{H}_n \) such that \( f(A) = \sum f(a_j)P_j \) if \( A \) has spectral decomposition \( A = \sum a_j P_j \). where \( P_j \) is the orthogonal projection of \( \mathbb{C}^n \) onto the kernel of \( A - a_j I \). Furthermore, we can consider the scalar function \( A \mapsto \text{tr} f(A) \). By Theorem 1.1 in [58], we have the following.

**Theorem 5.5.1.** Suppose \( f : [0,1] \rightarrow \mathbb{R} \) is a continuously differentiable concave function with derived function \( f'(x) \). Then the gradient function of the scalar function \( A \mapsto \text{tr} f(A) \) is given by \( f'(A) = \sum f'(a_j)P_j \) if \( A \) has spectral decomposition \( A = \sum a_j P_j \).

Applying the result to the von Neumann entropy and Rényi entropy, we have

**Corollary 5.5.2.** The gradient of the objective function \( H(\rho) = -\text{tr}(\rho \log \rho) \) is

\[
\nabla H(\rho) = -(\log \rho + I_{n_1n_2}).
\]  

The gradient of the objective function \( H_\alpha(\rho) = H_\alpha(\rho) = \frac{1}{1-\alpha} \log \text{tr}(\rho^\alpha) = \frac{1}{1-\alpha} \log \left( \sum \lambda_j^\alpha \right) \) is

\[
\nabla H_\alpha(\rho) = (\text{tr} \rho^\alpha)^{-1} \rho^\alpha - 1.
\]  

In the following, we compute the projection operator \( \text{proj}_{\mathcal{PSD}_{n_1n_2} \cap \mathcal{S}_2}(Z) \). There is no analytic expression of \( \text{proj}_{\mathcal{PSD}_{n_1n_2} \cap \mathcal{S}_2}(Z) \). Fortunately, we can use the Dykstra’s algorithm to derive it, which can be stated in Algorithm 5.5.3. The projection operator \( \text{proj}_{\mathcal{S}_1}(Z) \) is given by Corollary 5.3.6 and the projection operator \( \text{proj}_{\mathcal{PSD}_{n_1n_2} \cap \mathcal{S}_2}(Z) \) has been discussed in 5.3.

**Algorithm 5.5.3.** Alternating Projection Scheme to find \( \rho = \text{proj}_{\mathcal{PSD}_{n_1n_2} \cap \mathcal{S}_2}(Z) \)

**Step 1.** Choose a positive integer \( N \) (iteration limit) and a small positive number \( \delta \) (tolerance).

Set \( X_2^{(0)} = Z \) and do the following steps for \( k = 1, 2, \ldots, N \).

**Step 2.** Compute \( X_1^{(k)} \) and \( X_2^{(k)} \) as follows

\[
X_1^{(k)} = \text{proj}_{\mathcal{S}_2}(X_2^{(k-1)}) \quad \text{and} \quad X_2^{(k)} = \text{proj}_{\mathcal{PSD}_{n_1n_2}}(X_1^{(k)})
\]
Step 3. If $||X_1^{(k)} - X_2^{(k)}||_2 < \delta$, then stop and declare $X_2^{(k)}$ a solution.

By Boyle and Dykstra [14], one can show that the matrix sequences $\{X_1^{(k)}\}$ and $\{X_2^{(k)}\}$ generated by Algorithm 5.5.3 converge to the projection $\text{proj}_{PSD_{n_1n_2}\cap S_2}(Z)$, that is

$$X_1^{(k)} \to \text{proj}_{PSD_{n_1n_2}\cap S_2}(Z), \quad \text{and} \quad X_2^{(k)} \to \text{proj}_{PSD_{n_1n_2}\cap S_2}(Z), \quad k \to +\infty.$$

Thus, Algorithm 5.5.3 will determine projection operator $\text{proj}_{PSD_{n_1n_2}\cap S_2}(Z)$.

Next, we use the nonmonotone spectral projected gradient method (see [9, 10] for more details) to solve the minimization problem (5.40). The algorithm starts with $\rho_0 \in PSD_{n_1n_2} \cap S_2$ and use an integer $M \geq 1$; a small parameter $\alpha_{\text{min}} > 0$; a large parameter $\alpha_{\text{max}} > \alpha_{\text{min}}$; a sufficient decrease parameter $r \in (0, 1)$ and safeguarding parameters $0 < \sigma_1 < \sigma_2 < 1$. Initially, $\alpha_0 \in [\alpha_{\text{min}}, \alpha_{\text{max}}]$ is arbitrary. Given $\rho_t \in PSD_{n_1n_2} \cap S_2$ and $\alpha_t \in [\alpha_{\text{min}}, \alpha_{\text{max}}]$, Algorithm 5.5.4 describes how to obtain $\rho_{t+1}$ and $\alpha_{t+1}$, and when to terminate the process. In the following algorithm, the gradient $\nabla H(\rho)$ is defined in Corollary 5.5.2 and the projection operator $\text{proj}_{PSD_{n_1n_2}\cap S_2}(\cdot)$ is computed by Algorithm 5.5.3.

Algorithm 5.5.4. Scheme to solve minimization problem (5.40)

Step 1. Detect whether the current point is stationary: if $||\text{proj}_{PSD_{n_1n_2}\cap S_2}(\rho_t - \nabla H(\rho_t)) - \rho_t||_F \leq \text{tol}$, then stop and declare that $\rho_t$ is a stationary point.

Step 2. Backtracking

Step 2.1. Compute $d_t = \text{proj}_{PSD_{n_1n_2}\cap S_2}(\rho_t - \alpha_t \nabla H(\rho_t)) - \rho_t$. Set $\lambda \leftarrow 1$.

Step 2.2. Set $\rho_+ = \rho_t + \lambda d_t$.

Step 2.3. If

$$H(\rho_+) \leq \max_{0 \leq j \leq \min\{t,M-1\}} H(\rho_{t-j}) + \gamma \lambda \langle d_t, \nabla H(\rho_t) \rangle,$$

then define $\lambda_t = \lambda$, $\rho_{t+1} = \rho_+$, $s_t = \rho_{t+1} - \rho_t$, $y_t = H(\rho_{t+1}) - H(\rho_t)$, and go to Step 3.
If (5.44) does not hold, define
\[ \lambda_{\text{new}} = \frac{\sigma_1 \lambda + \sigma_2 \lambda}{2} \in [\sigma_1 \lambda, \sigma_2 \lambda], \]
set \( \lambda \leftarrow \lambda_{\text{new}} \), and go to Step 2.2.

**Step 3.** Compute \( b_t = \langle s_t, y_t \rangle \). If \( b_t \leq 0 \), set \( \alpha_{t+1} = \alpha_{\max} \), else, compute \( \alpha_t = \langle s_t, s_t \rangle \) and
\[ \alpha_{t+1} = \min\{\alpha_{\max}, \max\{\alpha_{\min}, \frac{a_t}{b_t}\}\}. \]

By Theorem 2.2 in [58], one can show that the sequence \( \{\rho_t\} \) generated by Algorithm 5.5.4 converges to the solution of the minimization problem (5.40). A computational comment can be made on Algorithm 5.5.4. In order to guarantee the iterative sequence \( \rho_t \in PSD_{n_1, n_2} \cap S_2, t = 0, 1, 2, \ldots \), the initial value \( \rho_0 \) must be in \( PSD_{n_1, n_2} \cap S_2 \). Taking \( \rho_1 \) for example, if \( \rho_0 \in PSD_{n_1, n_2} \cap S_2 \), then \( \rho_1 = \rho_0 + \alpha_1 d_1 \in PSD_{n_1, n_2} \cap S_2 \), because \( d_1 = \text{proj}_{PSD_{n_1, n_2} \cap S_2} (\rho_0 - t_0 \nabla H(\rho_0)) - x_0 \in PSD_{n_1, n_2} \cap S_2 \) and \( \alpha_1 \) is a scalar.

### 5.6 Product of Two Positive Contractions††

It is known that every matrix \( A \in \mathbb{C}^{n \times n} \) with nonnegative determinant can be written as the product of \( k \) positive semidefinite matrices with \( k \leq 5 \); see [3, 22, 93] and their references. Moreover, characterizations are given of matrices that can be written as the product of \( k \) positive semidefinite matrices but not fewer for \( k = 2, \ldots, 5 \). In particular, a matrix \( A \) is the product of two positive semidefinite matrices if it is similar to a diagonal matrix with nonnegative diagonal entries.

In this section, characterizations are given to \( A \in \mathbb{C}^{n \times n} \) which is a product of two positive contractions, i.e., positive semidefinite matrices with norm not larger than one. Evidently, if a

††The material in this section is contained in the paper [65], which is a joint work of C.-K. Li, K.-Z. Wang and the author.
matrix is the product of two positive contractions, then it is a contraction similar to a diagonal matrix with nonnegative diagonal entries. However, the converse is not true. For example, $A = \frac{1}{25} \begin{bmatrix} 9 & 3 \\ 0 & 16 \end{bmatrix}$ is a contraction similar to diag($9, 16$)/25 that is not a product of two positive contractions as shown in [70]. In fact, the result in [70] implies that if $A \in \mathbb{C}^{n \times n}$ is similar to a diagonal matrix with nonzero eigenvalues $a, b \in (0, 1]$ then a necessary and sufficient condition for $A$ to be the product of two positive contractions is:

$\left\{ \|A\|^2 - (a^2 + b^2) + \left(\frac{ab}{\|A\|}\right)^2 \right\}^{1/2} \leq |\sqrt{a} - \sqrt{b}|\sqrt{(1-a)(1-b)}$;

(see Corollary 5.6.6). In particular, a matrix $A = \begin{bmatrix} a & p \\ 0 & b \end{bmatrix} \in M_2$ is the product of two positive contractions if and only if $a, b \in [0, 1]$ and $|p| \leq |\sqrt{a} - \sqrt{b}|\sqrt{(1-a)(1-b)}$.

In Section 5.6.1, we will present several characterizations of a square matrix that can be written as the product of two positive (semidefinite) contractions. In Section 5.6.2, based on one of the characterizations in Section 5.6.1, we use alternating projection method to check the condition and construct the two positive contractions whose product equal to the given matrix if they exist. Some numerical examples generated by Matlab are presented.

### 5.6.1 Characterizations

If $A$ is a product of two positive semidefinite contractions, then $A$ is similar to a diagonal matrix with nonnegative eigenvalues with magnitudes bounded by $\|A\| \leq 1$. We will focus on such matrices in our characterization theorem.

It is known that a matrix $A$ is the product of two orthogonal projections if and only if it is unitarily similar to a matrix which is the direct sum of $I_p \oplus 0_q$ and matrices of the
form
\[
\begin{bmatrix}
a_j & \sqrt{a_j - a_j^2} \\
0 & 0
\end{bmatrix} \in \mathbb{C}^{2 \times 2}, \quad 0 < a_j < 1 \text{ for all } j = 1, \ldots, m;
\]
see [31]. Here we give another characterization which will be useful for our study.

**Proposition 5.6.1.** Suppose \( A \) is similar to \( I_p \oplus 0_q \oplus \text{diag}(a_1, \ldots, a_m) \) with \( a_1, \ldots, a_m \in (0,1) \). Then \( A \) is the product of two orthogonal projections in \( \mathbb{C}^{n \times n} \) if and only if \( A \) is unitarily similar to \( I_p \oplus A_1 \) and there is an \((n-p) \times m\) matrix \( S \) of rank \( m \) such that \( A_1A_1^*S = A_1S = S\text{diag}(a_1, \ldots, a_m) \).

**Proof.** For simplicity, we assume that \( I_p \) is vacuous. Suppose \( A \) is the product of two orthogonal projections in \( \mathbb{C}^{n \times n} \). Let \( D = \text{diag}(a_1, \ldots, a_m) \). We may assume that \( a_1 \geq \cdots \geq a_m \). There is a unitary \( U \) such that \( U^*AU = \begin{pmatrix} D & \sqrt{D-D^2} \\ 0 & 0_m \end{pmatrix} \oplus 0_{q-m}. \) Let \( U = \sum_{j=0}^{n-1} |j\rangle\langle u_{j+1}| \) and \( U_m = \sum_{j=0}^{m-1} |j\rangle\langle u_{j+1}| \in \mathbb{C}^{n \times m} \). Hence, we have \( AA^*S = AS = SD \) with \( S = U_m \).

Conversely, suppose \( S \) satisfies \( AA^*S = AS = S\text{diag}(a_1, \ldots, a_m) \), and has linearly independent columns \(|v_1\rangle, \ldots, |v_m\rangle \). We may assume that \( \langle v_s | v_s \rangle = 1 \) for \( 1 \leq s \leq m \) and \( \langle v_s | v_t \rangle = 0 \) if \( a_s = a_t \) and \( s \neq t \). Since \( AA^* \) is normal and \( v_i \) is an eigenvector of \( AA^* \) corresponding to the eigenvalue \( a_s \), \( \langle v_s | v_i \rangle = 0 \) for \( a_s \neq a_i \). Hence \( S^*S = I_m \).

Now, we can find an orthonormal set \( \{|v_{m+1}\rangle, \ldots, |v_n\rangle \} \) such that \( V = \sum_{j=0}^{n-1} |j\rangle\langle v_{j+1}| \) and \( V^*AA^*V = D \oplus 0_q. \) Then \( V^*AV \) is of the form \( \begin{pmatrix} D & B \\ 0 & 0_q \end{pmatrix} \), where \( B \) is an \( m \times q \) matrix with \( BB^* = D - D^2 \). From the QR factorization, \( B \) can be written as \( RQ \) with \( Q \) unitary and \( R \) lower triangular. Let \( V_1 = I_m \oplus Q^* \). Then \( V_1^*V^*AVV_1 = \begin{pmatrix} D & R \\ 0 & 0_q \end{pmatrix} \) and \( RR^* = BQ^*QB^* = D - D^2 \). Hence \( R = [\sqrt{D-D^2} \ 0_{m,(q-m)}] \), and we see that \( A \) is
unitarily similar to the direct sum of $0_q$ and matrices of the form

$$\begin{bmatrix} a_j & \sqrt{a_j - a_j^2} \\ 0 & 0 \end{bmatrix} \in \mathbb{C}^{2 \times 2}, \quad j = 1, \ldots, m.$$ 

Hence $A$ is the product of two orthogonal projections. \hfill \Box

Recall that $A \in \mathbb{C}^{n \times n}$ has a dilation $B \in \mathbb{C}^{N \times N}$ with $n < N$ if there is a unitary $V \in \mathbb{C}^{N \times N}$ such that $A$ is the leading principal submatrix of $V^*BV$. For two Hermitian matrices $X, Y \in \mathbb{C}^{n \times n}$, we write $X \geq Y$ if $X - Y$ is positive semidefinite. In the next theorem, we present two characterizations for matrices which can be written as the product of two positive contractions in terms of dilation and matrix inequalities. We begin with the following observation.

**Lemma 5.6.2.** Suppose $A \in \mathbb{C}^{n \times n}$ is the product of two positive contractions. Then $A$ is unitarily similar to a matrix of the form

$$I_p \oplus \begin{bmatrix} A_{11} & A_{12} \\ 0 & 0_{n-p-m} \end{bmatrix},$$

where $A_{11} \in \mathbb{C}^{m \times m}$ is similar to a diagonal matrix with the eigenvalues in $(0, 1)$.

**Proof.** Obviously, the eigenvalues of $A$ are in $[0, 1]$. From [2, Proposition 3.1(d)], we have

$$A \cong \begin{pmatrix} I_p & B_1 & B_2 \\ 0 & A_{11} & A_{12} \\ 0 & 0 & 0_{n-p-m} \end{pmatrix},$$

where $A_{11} \in \mathbb{C}^{m \times m}$ is an upper block triangular matrix such that the diagonal blocks are scalar matrices corresponding to distinct scalars, $1 > \lambda_1 > \cdots > \lambda_k > 0$. Since $\|A\| \leq 1$,
$B_1$ and $B_2$ are zero matrices. By [2, Proposition 3.1(c) and (d)], $A_{11}$ is similar to a diagonal matrix, and the desired conclusion follows.

\begin{proof}
\end{proof}

**Theorem 5.6.3.** Suppose $A = I_p \oplus \begin{bmatrix} A_{11} & A_{12} \\ 0 & 0_{n-p-m} \end{bmatrix} \in \mathbb{C}^{n\times n}$ such that $A_{11} \in \mathbb{C}^{m\times m}$ is similar to $D \equiv \text{diag}(a_1, \ldots, a_m)$ with $1 > a_1 \geq \cdots \geq a_m > 0$. The following conditions are equivalent.

(a) $A$ is the product of two positive contractions.

(b) $A$ has a dilation $\tilde{T} \in \mathbb{C}^{(n+2m)\times(n+2m)}$, which is the product of two orthogonal projections and has the same rank and eigenvalues of $A$. Equivalently, there are matrices $R, C \in \mathbb{C}^{m\times m}$ such that

\[
\tilde{T} = I_p \oplus \begin{bmatrix}
A_{11} & A_{12} & 0 & A_{11}C \\
0 & 0_{n-p-m} & 0 & 0 \\
RA_{11} & RA_{12} & 0_m & RA_{11}C \\
0 & 0 & 0 & 0_m \\
\end{bmatrix} \in \mathbb{C}^{(n+2m)\times(n+2m)}
\]

is the product of two orthogonal projections.

(c) There is an invertible contraction $U_{11} \in \mathbb{C}^{m\times m}$ satisfying

\[
A_{11}U_{11} = U_{11}D \quad \text{and} \quad U_{11}DU_{11}^* \geq A_{11}A_{11}^* + A_{12}A_{12}^*.
\]

Moreover, if condition (c) holds, we have $A = (I_p \oplus P)(I_p \oplus Q)$ for the positive contractions

\[
P = \begin{bmatrix}
U_{11}U_{11}^* & 0 \\
0 & 0_{n-p-m} \\
\end{bmatrix} \quad \text{and} \quad Q = \begin{bmatrix}
(U_{11}^*)^{-1}D(U_{11}^*)^{-1} & (U_{11}U_{11}^*)^{-1}A_{12} \\
A_{12}(U_{11}U_{11}^*)^{-1} & A_{12}^*(U_{11}DU_{11}^*)^{-1}A_{12} \\
\end{bmatrix}.
\]

134
**Proof.** For simplicity, we can assume that \( I_p \) is vacuous because the matrix \( A \) is the product of two positive contractions if and only if each of the two positive contractions is a direct sum of \( I_p \) and a positive contraction in \( \mathbb{C}^{(n-p)\times(n-p)} \).

First we establish the equivalence of (a) and (b). If (a) holds, then \( A = PQ \), where \( P, Q \) are two positive contractions. Then

\[
\tilde{P} = \begin{bmatrix}
P & \sqrt{P - P^2} & 0 \\
\sqrt{P - P^2} & I_n - P & 0 \\
0 & 0 & 0_n
\end{bmatrix}
\quad \text{and} \quad
\tilde{Q} = \begin{bmatrix}
Q & 0 & \sqrt{Q - Q^2} \\
0 & 0_n & 0 \\
\sqrt{Q - Q^2} & 0 & I_n - Q
\end{bmatrix}
\]

are orthogonal projections such that

\[
\tilde{P}\tilde{Q} = \begin{bmatrix}
PQ & 0 & P\sqrt{Q - Q^2} \\
\sqrt{P - P^2}Q & 0_n & \sqrt{(P - P^2)(Q - Q^2)} \\
0 & 0 & 0_n
\end{bmatrix}.
\]

Let \( Y = \sqrt{Q^+ - Q^+Q} \) and \( X = \sqrt{P^+ - P^+P} \), where \( P^+, Q^+ \) is the Moore-Penrose inverses of \( P \) and \( Q \). (Recall that for a Hermitian matrix \( H = \sum_{j=1}^{\ell} \lambda_j |\xi_j\rangle\langle\xi_j| \in \mathbb{C}^{n \times n} \) with nonzero eigenvalues \( \lambda_1, \ldots, \lambda_\ell \) and orthonormal eigenvectors \( |\xi_1\rangle, \ldots, |\xi_\ell\rangle \), its Moore-Penrose inverse \( H^+ \) is \( \sum_{j=1}^{\ell} \lambda_j^{-1} |\xi_j\rangle\langle\xi_j| \).) Let

\[
T = \begin{bmatrix}
A & 0 & AY \\
X^*A & 0_n & X^*AY \\
0 & 0 & 0_n
\end{bmatrix}.
\]

The rows of the matrix \( X^*A \) lie in the row space of \([A_{11}A_{12}]\) and the columns of \( AY \) lie in the column space of \( A_{11} \). So, there is unitary matrix of the form \( U = I_n \oplus U_1 \oplus U_2 \) with
Thus, 

\[
\tilde{T} = \begin{bmatrix}
A_{11} & A_{12} & 0 & A_{11}C \\
0 & 0 & 0 & 0 \\
RA_{11} & RA_{12} & 0 & RA_{11}C \\
0 & 0 & 0 & 0 \\
\end{bmatrix} \in \mathbb{C}^{(n+2m)\times(n+2m)}
\]

has the same rank and eigenvalues as the leading submatrix \(A\). Thus, condition \(b\) holds.

Conversely, suppose \(b\) holds. and \(\tilde{T}\) is the product of two orthogonal projections \(\tilde{P} = VV^*\) and \(\tilde{Q} = WW^*\) with \(V \in \mathbb{C}^{(n+2m)\times r}, W \in \mathbb{C}^{(n+2m)\times s}\) such that \(V^*V = I_r\) and \(W^*W = I_s\). Evidently, \(\tilde{T}\) has rank \(m\). So,

\[
V^*W = Y \begin{bmatrix} K & 0 \\ 0 & 0_{(r-m),(s-m)} \end{bmatrix} Z^*
\]

such that \(Y \in \mathbb{C}^{r\times r}, Z \in \mathbb{C}^{s\times s}\) are unitary and \(K \in \mathbb{C}^{m\times m}\) is a diagonal matrix with positive diagonal entries. Let \(Y = [Y_1|Y_2], Z = [Z_1|Z_2]\) be such that \(Y_1 \in \mathbb{C}^{r\times m}, Z_1 \in \mathbb{C}^{s\times m}\). Note that

\[
Y_1^*V^*WZ_1 = Y_1^*[Y_1|Y_2] \begin{bmatrix} K & 0 \\ 0 & 0_{(r-m),(s-m)} \end{bmatrix} [Z_1|Z_2]^*Z_1 = K.
\]
Furthermore,
\[
\tilde{V} = VY_1 = \begin{bmatrix} V_1 \\ V_2 \\ V_3 \end{bmatrix} \quad \text{and} \quad \tilde{W} = WZ_1 = \begin{bmatrix} W_1 \\ W_2 \\ W_3 \end{bmatrix},
\]
where \( V_1, W_1 \) are \( n \times m \), \( V_2, V_3, W_2, W_3 \in \mathbb{C}^{m \times m} \). Then
\[
\tilde{V} \tilde{V}^* \tilde{W} \tilde{W}^* = VY_1 Y_1^* V^* WZ_1 Z_1^* W^* = VY_1 KZ_1^* W^* = VV^* W.W^* = \tilde{T}.
\]

Now, the last \( m \) rows of \( \tilde{T} \) and the \((n + 1)\)st, \( \ldots \), \((n + m)\)th columns of \( \tilde{T} \) are zero. Thus,
\[
V_3 \tilde{V}^* \tilde{W} \tilde{W}^* = V_3 K \tilde{W}^* = 0_{m,(n+2m)} \quad \text{and} \quad \tilde{V} \tilde{V}^* \tilde{W} W_2 = \tilde{V} K W_2^* = 0_{(n+2m),m}.
\]

Because \( K \tilde{W}^* \) has full row rank and \( \tilde{V} K \) has full column rank, we see that \( V_3 = 0_m \) and \( W_2 = 0_m \). Consequently, \( A = V_1 V_1^* W_1 W_1^* \) is the product of two positive contractions \( V_1 V_1^* \) and \( W_1 W_1^* \).

Next, we prove the equivalence of conditions (b) and (c). Suppose (b) holds, and
\[
\tilde{T} = \begin{bmatrix}
A_{11} & A_{12} & 0 & A_{11}C \\
0 & 0 & 0 & 0 \\
RA_{11} & RA_{12} & 0_m & RA_{11}C \\
0 & 0 & 0 & 0_m
\end{bmatrix} \in \mathbb{C}^{(n+2m) \times (n+2m)}
\]
has the same rank and eigenvalues as the leading submatrix \( A \).

Now, assume that \( U = (U_{ij})_{1 \leq i \leq 4, 1 \leq j \leq 3} \in \mathbb{C}^{(n+2m) \times (n+2m)} \) is unitary with \( U_{11}, U_{12} \in \mathbb{C}^{n \times m} \).
\( C^{m \times m}, U_{13} \in C^{m \times n} \) and \( U_{31}, U_{41} \in C^{m \times m}, U_{21} \in C^{n-m \times m} \) such that
\[
U^* \tilde{T} U = \begin{bmatrix} D & \sqrt{D - D^2} & 0 \\ 0 & 0_m & 0 \\ 0 & 0 & 0_n \end{bmatrix}.
\]

Now,
\[
\begin{bmatrix} A_{11} U_{11} + A_{12} U_{21} \\ 0_{n-m,m} \\ R A_{11} U_{11} + R A_{12} U_{21} \\ 0_m \end{bmatrix} = \tilde{T} \begin{bmatrix} U_{11} \\ U_{21} \\ U_{31} \\ U_{41} \end{bmatrix} = D.
\]

It follows that \( U_{21}, U_{41} \) are zero matrices. Furthermore,
\[
A_{11} U_{11} = U_{11} D, \quad R A_{11} U_{11} = U_{31} D.
\]

Thus, \( R U_{11} D = U_{31} D \) so that \( R U_{11} = U_{31} \). If \( x \in C^m \) satisfies \( U_{11} x = 0 \), then
\[
x = (U^*_{11} \ U^*_{31}) \begin{bmatrix} U_{11} \\ U_{31} \end{bmatrix} x = U^*_{11} (I_m + R^* R) U_{11} x = 0.
\]

Hence, \( U_{11} \in C^{m \times m} \) has linearly independent columns, i.e., \( U_{11} \) is invertible.

Next, observe that
\[
\tilde{T} \tilde{T}^* U = U \begin{bmatrix} D & 0 & 0 \\ 0 & 0_m & 0 \\ 0 & 0 & 0_n \end{bmatrix}.
\]

So,
\[
(A_{11} A^*_{11} + A_{12} A^*_{12} + A_{11} CC^* A^*_{11})(I_m + R^* R) U_{11} = U_{11} D,
\]
and hence
\[(A_{11}A_{11}^* + A_{12}A_{12}^* + A_{11}CC^*A_{11}^*) = U_{11}DU_{11}^*, \tag{5.45}\]
because
\[I_m = U_{11}^*U_{11} + U_{31}^*U_{31} = U_{11}^*(I_m + R^*R)U_{11} = (I_m + R^*R)U_{11}U_{11}^*. \tag{5.46}\]

So, \(R\) and \(C\) exist if and only if there is a contraction \(U_{11} \in \mathbb{C}^{m \times m}\) satisfying
\[A_{11}U_{11} = U_{11}D \quad \text{and} \quad U_{11}DU_{11}^* \geq A_{11}A_{11}^* + A_{12}A_{12}^*.\]

Conversely, suppose \((c)\) holds. Then there exist \(R\) and \(C\) satisfying (5.45) and (5.46).

Let
\[\tilde{U} = \begin{bmatrix}
U_{11} \\
0_{n-m,m} \\
RU_{11} \\
0_m
\end{bmatrix}.
\]

Then \(\tilde{U}\) has rank \(m\) and the matrix \(\tilde{T}\) in condition \((b)\) satisfies \(\tilde{T}\tilde{T}^*\tilde{U} = \tilde{T}\tilde{U} = \tilde{U}D\). By Proposition 5.6.1, we see that \(\tilde{T}\) is the product of two orthogonal projections.

To verify the last statement, note that \(A_{11}U_{11} = U_{11}D\) so that \(A_{11} = U_{11}DU_{11}^{-1}\). Hence,
\[PQ = \begin{bmatrix}
U_{11}DU_{11}^{-1} & A_{12} \\
0 & 0_{n-m}
\end{bmatrix} = \begin{bmatrix}
A_{11} & A_{12} \\
0 & 0_{n-m}
\end{bmatrix}.
\]
and \( Q = ZZ^* \) with 
\[
Z = \begin{bmatrix}
(U^*)^{-1/2}D^{1/2} \\
A_{12}^*(U^*)^{-1/2}D^{-1/2}
\end{bmatrix}
\]
so that
\[
Z^*Z = D^{1/2}U_{11}^{-1}(U^*)^{-1}D^{1/2} + D^{-1/2}U_{11}^{-1}A_{12}A_{12}^*(U^*)^{-1}D^{-1/2}
\]
\[
= D^{-1/2}U_{11}^{-1}(A_{11}A_{11}^* + A_{12}A_{12}^*)(U^*)^{-1}D^{-1/2}
\]
\[
\leq D^{-1/2}U_{11}^{-1}(U_{11}DU_{11}^*)(U^*)^{-1}D^{-1/2} = I_m.
\]

This shows that \( Z \) is a contraction and hence so is \( Q \). \( \square \)

As pointed out by the referee, from Theorem 5.6.3 one can deduce the following corollary, which can be viewed as a 2-variable generalization of the fact that every positive contraction can be dilated to an orthogonal projection; see [47, Problem 222(b)].

**Corollary 5.6.4.** If \( A \in \mathbb{C}^{n \times n} \) is the product of two positive contractions, then \( A \) can be dilated to a product of two projections on \( \mathbb{C}^{n+2m} \), where \( m \) equals the number of eigenvalues of \( A \) which are not equal to 0 or 1.

It is not easy to check the existence of the matrices \( R, C \in \mathbb{C}^{m \times m} \) in condition (b), and the existence of \( U_{11} \) in condition (c) of Theorem 5.6.3. We refine condition (c) to get Theorem 5.6.5 below so that one can use computational techniques such as positive semidefinite programming or alternating projection methods to check the condition. In Section 3, we will develop Matlab programs using an alternating projection method based on Theorem 5.6.5 to check whether a matrix can be written as the product of two positive semidefinite contractions, and construct them if they exist.

**Theorem 5.6.5.** Let \( A \in \mathbb{C}^{n \times n} \) be unitarily similar to \( I_p \oplus 0_q \oplus \begin{bmatrix} A_{11} & A_{12} \\ 0 & 0_{n-p-q-m} \end{bmatrix} \), where \( A_{11} \in \mathbb{C}^{m \times m} \) such that \( A_{11} \) is diagonalizable with distinct eigenvalues \( \alpha_1 > \cdots > \alpha_k \) in \((0, 1)\) with multiplicities \( m_1, \ldots, m_k \), respectively. Suppose \( V = [V_1 \cdots V_k] \in \mathbb{C}^{m \times m} \) is an
invertible matrix such that the columns of the $n \times m_j$ matrix $V_j$ form an orthonormal basis for the null space of $A_{11} - \alpha_j I_m$, for $j = 1, \ldots, k$, i.e., $A_{11} V = V D$, where $D = \alpha_1 I_{m_1} \oplus \cdots \oplus \alpha_k I_{m_k}$ and $V_j^* V_j = I_{m_j}$ for $j = 1, \ldots, k$. Then $A$ is the product of two positive contractions if and only if there is a block diagonal matrix $\Gamma = \Gamma_1 \oplus \cdots \oplus \Gamma_k \in \mathbb{C}^{m_1 \times m_1} \oplus \cdots \oplus \mathbb{C}^{m_k \times m_k}$ satisfying

$$D^{1/2} V^* (A_{11} A_{11}^* + A_{12} A_{12}^*)^{-1} V D^{1/2} \geq \Gamma \geq V^* V. \quad (5.47)$$

**Proof.** Suppose $A_{11} V = V D$ as asserted. Then $U$ satisfies $A_{11} U = U D$ if and only if $U = V L$ for some block matrix $L = L_1 \oplus \cdots \oplus L_k \in \mathbb{C}^{m_1 \times m_1} \oplus \cdots \oplus \mathbb{C}^{m_k \times m_k}$. One readily checks that condition (c) in Theorem 5.6.3 reduces to the existence of $\Gamma = (L L^*)^{-1}$. \qed

By Theorem 5.6.5, we can deduce the following corollary. The first part of the corollary was obtained in [70, Lemma 2.1] by some rather involved arguments. The second part of the corollary is a proof of a comment in our introduction.

**Corollary 5.6.6.** Let $A = \begin{bmatrix} a & p \\ 0 & b \end{bmatrix}$ with $a, b \in [0, 1]$. Then $A$ is the product of two positive contractions if and only if

$$|p| \leq |\sqrt{a} - \sqrt{b}| \sqrt{(1 - a)(1 - b)}. \quad (5.48)$$

Consequently, if $B \in \mathbb{C}^{n \times n}$ is similar to a diagonal matrix with nonzero eigenvalues $a, b \in (0, 1]$ then a necessary and sufficient condition for $A$ to be the product of two positive contractions is:

$$\left\{ \|B\|^2 - (a^2 + b^2) + (ab/\|B\|^2) \right\}^{1/2} \leq |\sqrt{a} - \sqrt{b}| \sqrt{(1 - a)(1 - b)}. \quad (5.48)$$

**Proof.** Case 1. $a = b$. If $A$ is the product of two positive contractions, then $A$ is similar to a diagonal matrix so that $p = 0$, and inequality (5.48) holds. If inequality (5.48)
holds, then $p = 0$, and $A = aI_2$ is the product of positive contractions $I_2$ and $aI_2$.

Case 2. $a \neq b$. We focus on the non-trivial case that $a, b \in (0, 1)$, $a \neq b$ and $p \neq 0$. One sees that $V$ in Theorem 5.6.5 can be chosen to be

$$V = \begin{bmatrix} 1 & p/\gamma \\ 0 & (b-a)/\gamma \end{bmatrix}$$

with $\gamma = \sqrt{(a-b)^2 + p^2}$ so that up to diagonal congruence we have

$$V^*V = \begin{bmatrix} 1 & p/\gamma \\ p/\gamma & 1 \end{bmatrix}.$$ 

We need to find a diagonal matrix $\Gamma = \text{diag}(d_1, d_2)$ with $d_1, d_2 \geq 0$ such that $\Gamma - V^*V \geq 0$ and $VV^* - \text{diag}(ad_1, bd_2) \geq 0$. Thus, we want

$$(d_1 - 1)(d_2 - 1) \geq p^2/\gamma^2, \quad (1 - d_1a)(1 - d_2b) \geq p^2/\gamma^2.$$ 

We consider the maximum values for

$$f(d_1, d_2) = (d_1 - 1)(d_2 - 1)$$

subject to the condition of

$$g(d_1, d_2) = (d_1 - 1)(d_2 - 1) - (1 - d_1a)(1 - d_2b) = 0.$$ 

Consider the Lagrangian function $L(d_1, d_2, \mu) = f(d_1, d_2) - \mu g(d_1, d_2)$.

$$0 = L_{d_1}(d_1, d_2, \mu) = (d_2 - 1) - \mu[(d_2 - 1) + a(1 - d_2b)]$$

and

$$0 = L_{d_2}(d_1, d_2, \mu) = (d_1 - 1) - \mu[(d_1 - 1) + b(1 - d_1a)].$$
Thus,

\[(1 - \mu)^2(d_1 - 1)(d_2 - 1) = \mu^2 ab(1 - d_1 a)(1 - d_2 b).\]

Because \((d_1 - 1)(d_2 - 1) = (1 - d_1 a)(1 - d_2 b)\), we see that \((1 - \mu)^2 = \mu^2 ab\), and thus, \(\mu = (1 + \sqrt{ab})^{-1}\). Here, we use the root satisfying \(1 - \mu > 0\). Solving \(d_1\) and \(d_2\), we get

\[(d_1 - 1)(d_2 - 1) = (1 - a)(1 - b)/(1 + \sqrt{ab})^2.\]

Furthermore, \((d_1 - 1)(d_2 - 1) \geq p^2/\gamma^2\) if and only if

\[p^2 \leq (a - b)^2(1 - a)(1 - b)/(\sqrt{a} + \sqrt{b})^2 = (\sqrt{a} - \sqrt{b})^2(1 - a)(1 - b).\]

For the last assertion, note that if \(B\) satisfies the given assumption, then \((B - aI)(B - bI) = 0\), and \(B\) is unitarily similar to the direct sum of \(aI_p \oplus bI_l\) and matrices of the form

\[B_j = \begin{bmatrix} a & p_j \\ 0 & b \end{bmatrix},\]

where \(p_1 \geq \cdots \geq p_k > 0\), for \(j = 1, \ldots, k\). By Theorem 1.1 in [70], \(B\) is a product of two positive contractions if and only if

\[\|\text{diag}(p_1, \ldots, p_k)\| = |p_1| \leq |\sqrt{a} - \sqrt{b}| \sqrt{(1 - a)(1 - b)}.\]

It is easy to check that \(\|B\| = \|B_1\|\) and

\[\|B_1\|^2 + (ab/\|B_1\|)^2 - (a^2 + b^2) = \text{tr}(B_1^*B_1) - (a^2 + b^2) = p_1^2.\]

The assertion follows.
5.6.2 Alternating projections and numerical examples

In Theorem 5.6.5, if $A_{11}$ has distinct eigenvalues, then one only needs to search for a diagonal matrix satisfying the condition. However, there is no guarantee that there is a diagonal matrix $\Gamma$ satisfying the condition in general as shown in the following example.

**Example 5.6.7.** Let $D = \text{diag}(0.15, 0.15, 0.2)$, $A = \begin{bmatrix} A_{11} & A_{12} \\ 0_3 & 0_3 \end{bmatrix}$ with

$$A_{11} = \begin{bmatrix} 0.1500 & 0 & 0 \\ 0 & 0.1500 & 0.0375 \\ 0 & 0 & 0.2000 \end{bmatrix},$$

and

$$A_{12} = \{UDU^* - A_{11}A_{11}^*\}^{1/2} = \begin{bmatrix} 0.3571 & 0 & 0 \\ 0 & 0.3215 & 0.1070 \\ 0 & 0.1070 & 0.1689 \end{bmatrix},$$

where

$$U = VR = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 5/\sqrt{40} & 3/\sqrt{40} \\ 0 & 0 & 4/\sqrt{40} \end{bmatrix},$$

with

$$V = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 1/\sqrt{2} & -1/\sqrt{2} & 3/5 \\ 0 & 0 & 4/5 \end{bmatrix} \quad \text{and} \quad R = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 5/\sqrt{40} & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 5/\sqrt{40} \end{bmatrix}.$$
$A_{12}A_{12}^*$. There is no $\Gamma = \text{diag}(\mu_1, \mu_2, \mu_3)$ such that

$$M = D^{1/2}V^*(A_{11}A_{11}^* + A_{12}A_{12}^*)^{-1}VD^{1/2} = \begin{bmatrix} 1.3 & -0.3 & 0 \\ -0.3 & 1.3 & 0 \\ 0 & 0 & 1.6 \end{bmatrix} \geq \Gamma$$

and

$$\Gamma \geq V^*V = \begin{bmatrix} 1.000 & 0 & 0.4243 \\ 0 & 1.000 & -0.4243 \\ 0.4243 & -0.4243 & 1.00 \end{bmatrix}$$

because $\mu_1, \mu_2 \in (1, 1.3)$ so that the leading $2 \times 2$ principal submatrix $M - \Gamma$ cannot be positive semidefinite. Hence, $A$ is not the product of two positive contractions. \hfill \square

To check whether there exists $\Gamma$ satisfying (5.47), we turn to alternating projection method. Suppose $A \in \mathbb{C}^{n \times n}$ is a contraction matrix unitarily similar to $I_p \oplus 0_q \oplus \begin{bmatrix} A_{11} & A_{12} \\ 0 & 0_{n-p-q-m} \end{bmatrix}$ and $V \in \mathbb{C}^{m \times m}$ is an invertible matrix with unit columns $v_1, \ldots, v_m$ satisfying $A_{11}V = VD$ with $D = \alpha_1 I_{m_1} \oplus \cdots \oplus \alpha_k I_{m_k}$ with $\alpha_1 > \cdots > \alpha_k > 0$ the distinct eigenvalues of $A_{11}$. Define the convex sets

$$\mathcal{S}_1 = \{ \Gamma = \Gamma_1 \oplus \cdots \oplus \Gamma_k \in \mathbb{C}^{m_1 \times m_1} \oplus \cdots \oplus \mathbb{C}^{m_k \times m_k} : \Gamma \text{ is positive semidefinite} \},$$

$$\mathcal{S}_2 = \{ \Gamma \in \mathbb{C}^{m \times m} : D^{1/2}V^*(A_{11}A_{11}^* + A_{12}A_{12}^*)^{-1}VD^{1/2} \geq \Gamma \geq 0 \},$$

and

$$\mathcal{S}_3 = \{ \Gamma \in \mathbb{C}^{m \times m} : \Gamma \geq V^*V \}.$$

The following proposition can be readily verified. Here we use the notation $X^+$ for the positive semidefinite part of a Hermitian matrix $X$, i.e., $X^+ = (X + \sqrt{X^2})/2$. 

145
Proposition 5.6.8. Let $G = [G_{st}]$ be a Hermitian matrix, where $G_s \in \mathbb{C}^{m_s \times m_s}$.

1. $\text{proj}_{S_1}(G) = G_{11}^+ \oplus \cdots \oplus G_{kk}^+$.

2. $\text{proj}_{S_2}(G) = M - (M - G)^+$, where $M = D^{1/2}V^*(A_{11}A_{11}^* + A_{12}A_{12}^*)^{-1}VD^{1/2}$.

3. $\text{proj}_{S_3}(G) = (G - V^*V)^+ + V^*V$.

In the following algorithm, we create a sequence

$$\Gamma_0 \rightarrow \hat{\Gamma}_1 \rightarrow \Gamma_1 \rightarrow \hat{\Gamma}_2 \rightarrow \Gamma_2 \rightarrow \cdots$$

where $\Gamma_k \in S_1$, $\hat{\Gamma}_{2k-1} \in S_2$ and $\hat{\Gamma}_{2k} \in S_3$ for all $k \geq 1$. This sequence converges to a solution $\Gamma \in S_1 \cap S_2 \cap S_3$, provided $S_1 \cap S_2 \cap S_3 \neq \emptyset$; see [14].

Algorithm 5.6.9. For checking the existence of $\Gamma \in S_1 \cap S_2 \cap S_3$.

Step 0. Set $k = 0$. Let $X = D^{1/2}V^*(A_{11}A_{11}^* + A_{12}A_{12}^*)^{-1}VD^{1/2}$ and $Y = V^*V$.

Partition $X$ into $[X_{pq}]$ and $Y$ into $[Y_{pq}]$, both conformed to $D$.

Set $\Gamma_0 = \frac{1}{2} \left( (X_{11} + Y_{11}) \oplus \cdots \oplus (X_{kk} + Y_{kk}) \right)$. Go to Step 1.

Step 1. Change $k$ to $k + 1$, and set

$$\hat{\Gamma}_k = \begin{cases} 
X - (X - \Gamma_{k-1})^+ & \text{if } k \text{ is odd,} \\
(\Gamma_{k-1} - Y)^+ + Y & \text{if } k \text{ is even,}
\end{cases}$$

where $M_+$ denotes the positive part of $M$.

Partition $\hat{\Gamma}_k$ into $[G_{ij}]$ conformed to $D$ and let $\Gamma_k = G_{11}^+ \oplus \cdots \oplus G_{kk}^+$.

If error $= \max(0, -\lambda_{\min}(\Gamma_k - Y)) + \max(0, -\lambda_{\min}(X - \Gamma_k)) \approx 0$, stop.

Otherwise, go to step 1.
Once we have $\Gamma$, we can set $U = V\Gamma^{-1/2}$, and construct the two projections as shown in Theorem 5.6.3. In particular, we can set $A = (I_p \oplus P)(I_p \oplus Q)$ with

$$P = \begin{bmatrix} UU^* & 0 \\ 0 & 0_{n-p-m} \end{bmatrix} \quad \text{and} \quad Q = \begin{bmatrix} (U^*)^{-1}DU^{-1} & (UU^*)^{-1}A_{12} \\ A_{12}^*(UU^*)^{-1} & A_{12}^*(UDU^*)^{-1}A_{12} \end{bmatrix}.$$ (5.49)

We illustrate our Matlab program (see http://cklixx.wm.edu/mathlib/Twoposcon.txt) for checking whether a given matrix $A \in \mathbb{C}^{n \times n}$ is the product of two positive contractions in the following. Note that all numerical experiments were performed using Matlab 2015a on a Intel(R) Core(TM) i7-5500U CPU @ 2.4GHz with 8GB RAM and a 64-bit OS.

**Example 5.6.10.** Suppose $A = \begin{bmatrix} A_{11} & A_{12} \\ 0_5 & 0_5 \end{bmatrix}$, where

$$A_{11} = \begin{bmatrix} 0.125 & 0.0126 & 0.0033 & 0.024 & -0.006 \\ 0 & 0.0625 & 0 & 0.012 & 0.0152 \\ 0 & 0 & 0.0625 & 0.0025 & 0.0453 \\ 0 & 0 & 0 & 0.2 & 0 \\ 0 & 0 & 0 & 0 & 0.2 \end{bmatrix}$$

and

$$A_{12} = \begin{bmatrix} 0.0658 & 0.0218 & 0.0031 & 0.05 & -0.0033 \\ 0.0218 & 0.113 & -0.0107 & -0.0120 & 0.0098 \\ 0.0031 & -0.0107 & 0.0418 & 0.0048 & -0.0409 \\ 0.0500 & -0.012 & 0.0048 & 0.1103 & 0.0037 \\ -0.0033 & 0.0098 & -0.0409 & 0.0037 & 0.128 \end{bmatrix}.$$
We set

\[ V \approx \begin{bmatrix}
1 & -0.1976 & -0.0507 & -0.3169 & -0.0169 \\
0 & 0.9803 & -0.0102 & -0.0824 & -0.1026 \\
0 & 0 & 0.9987 & -0.0172 & -0.3108 \\
0 & 0 & 0 & -0.9447 & 0.0203 \\
0 & 0 & 0 & 0 & -0.9445
\end{bmatrix}, \]

which has unit columns and satisfies \( A_{11}V = V \text{diag}(0.125, 0.0625, 0.0625, 0.2, 0.2) \); the second and third columns of \( V \) are orthogonal and the fourth and fifth columns are orthogonal.

Using our Matlab program, we obtain \( U = V \Gamma^{-\frac{1}{2}} \), where

\[ \Gamma = \begin{bmatrix}
3.4737 & 0 & 0 & 0 & 0 \\
0 & 2.3344 & 0.0216 & 0 & 0 \\
0 & 0.0216 & 2.9472 & 0 & 0 \\
0 & 0 & 0 & 2.1257 & -0.2132 \\
0 & 0 & 0 & -0.2132 & 1.6425
\end{bmatrix}. \]

Defining \( P \) and \( Q \) as in equation (5.49), we get that \( \lambda_1(P) = s_1^2(U) = 0.7024 \) and \( \lambda_1(Q) = 1 \). Note that \( \Gamma \) is obtained using alternating projection method after 79 iterations done in approximately 0.085 seconds with error \( = \|PQ - A\| = 4.3774 \times 10^{-14} \).
Example 5.6.11. Suppose

\[
A_{11} = \begin{bmatrix}
0.1 & 0.0244 & 0.026 & 0.0167 & 0.0114 & 0.0014 & 0.0674 \\
0 & 0.2 & 0.0176 & 0.0251 & 0.0345 & 0.0122 & 0.0088 \\
0 & 0 & 0.3 & 0 & 0.0072 & 0.0119 & 0.0166 \\
0 & 0 & 0 & 0 & 0.4 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0.4 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0.4
\end{bmatrix}
\]

and

\[
A_{12} = \begin{bmatrix}
0.098 & 0.0157 & -0.0315 & 0.0033 & -0.04 & -0.0196 & 0.0171 \\
0.0157 & 0.0545 & -0.0366 & 0.0302 & 0.0081 & 0.0003 & 0.004 \\
-0.0315 & -0.0366 & 0.1246 & -0.0449 & -0.0005 & 0.0232 & -0.0047 \\
0.0033 & 0.0302 & -0.0449 & 0.1025 & -0.0193 & -0.031 & 0.0191 \\
-0.04 & 0.0081 & -0.0005 & -0.0193 & 0.1285 & 0.0038 & -0.0504 \\
-0.0196 & 0.0003 & 0.0232 & -0.031 & 0.0038 & 0.0779 & -0.0192 \\
0.0171 & 0.004 & -0.0047 & 0.0191 & -0.0504 & -0.0192 & 0.0895
\end{bmatrix}
\]

We let

\[
V = \begin{bmatrix}
1 & -0.2373 & -0.1475 & -0.1015 & -0.0632 & -0.0196 & -0.2348 \\
0 & -0.9714 & -0.1713 & -0.2329 & -0.1858 & -0.0673 & -0.0569 \\
0 & 0 & -0.9741 & 0.0563 & -0.0702 & -0.1162 & -0.1512 \\
0 & 0 & 0 & -0.9656 & -0.0910 & -0.0052 & -0.0896 \\
0 & 0 & 0 & 0 & -0.9738 & 0.023 & 0.0454 \\
0 & 0 & 0 & 0 & 0 & -0.9905 & 0.0278 \\
0 & 0 & 0 & 0 & 0 & 0 & -0.9528
\end{bmatrix}
\]
Using our Matlab program, we obtain

\[
\Gamma = [2.9099] \oplus [2.592] \oplus \begin{bmatrix} 1.9048 & 0.1063 \\ 0.1063 & 1.866 \end{bmatrix} \oplus \begin{bmatrix} 1.6447 & 0.0046 & 0.0768 \\ 0.0046 & 1.6923 & 0.0215 \\ 0.0768 & 0.0215 & 1.5846 \end{bmatrix}
\]

after 59 iterations (approximately 0.075 seconds) with a $1.227 \times 10^{-16}$ error. The positive semidefinite matrices $P$ and $Q$ defined in equation (5.49) will have largest eigenvalues 0.8309 and 1, respectively.

**Example 5.6.12.** Let $A =$ \[
\begin{bmatrix} A_{11} & 0 \\ 0 & 0 \end{bmatrix}
\]
and $B =$ \[
\begin{bmatrix} B_{11} & 0 \\ 0 & 0 \end{bmatrix},
\]
where

\[
A_{11} = \begin{bmatrix} 0.5 & 0.09429 \\ 0 & 0.3 \end{bmatrix} \quad \text{and} \quad B_{11} = \begin{bmatrix} 0.5 & 0.0943 \\ 0 & 0.3 \end{bmatrix}.
\]

It follows from [70] that $A$ is a product of two contractions and $B$ is not. Notice that $A$ and $B$ are very close to each other. For $A$, we ran the alternating projection algorithm and obtained $\Gamma = \text{diag}(1.2759, 1.6591)$ after 66321 iterations (48.26 seconds). We also get $||PQ - A|| \approx 1.4778 \times 10^{-16}$ and $\lambda_1(Q), \lambda_1(P) \approx 1$. Meanwhile, for $B$, after running 100,000 iterations (69.06 seconds) of the algorithm, we see that the values $\max(0, -\min(\text{eig}(M - \Gamma)))$ and $\max(0, -\min(\text{eig}(\Gamma - V^*V)))$ starts to alternate back and forth from $8.5 \times 10^{-5}$ to $8.52925 \times 10^{-5}$.

### 5.7 Conclusion

In this section 5.2, we studied the basic problem of constructing a quantum channel that maps between given sets of quantum states. We have used the Choi matrix represen-
tation for completely positive maps to show that the construction is equivalent to solving a Hermitian positive semidefinite linear feasibility problem. This feasibility problem has special structure that can be exploited. We have shown the efficiency of using alternating projection and Douglas-Rachford projection/reflection algorithms for accurately solving large scale problems to high accuracy. This included finding trace preserving completely positive, TPCP, maps with high rank, as well as the nonconvex problems of finding TPCP maps with low rank.

In section 5.3, 5.4 and 5.5, we use projection methods to construct (global) quantum states with prescribed reduced (marginal) states, and specific ranks and possibly extreme Von Neumann or Renyi entropy. Using convex analysis, optimization techniques on matrix manifolds, we obtained numerical algorithms based on alternative projection methods to solve the problem. Matlab programs were written based on these algorithms, and numerical examples of low dimension cases were demonstrated. In our study, we have theoretical results ensuring convergence in some of the problems, and there are only numerical results supporting the efficiency of our schemes. It would be interesting to obtain convergence results for the latter cases. It is interesting to note that there are other projection methods such as the Douglas-Rachford reflection method. It is interesting to note that even if the convergence theory of such methods are not so well-developed, but the performance of the schemes often lead to optimal solutions.

In connection to our study, there are many follow up problems deserving further investigations. We mention some specific questions in the following.

1. We have only demonstrated our algorithms with low dimension examples. It is interesting to improve the algorithm so that it can deal with practical problems (of large sizes).

2. Besides the alternating projection methods, it is interesting to study other schemes
such as the Douglas-Rachford reflection method (for example, see [28, 89, 80]) to solve our problems.

3. Prove or disprove that Algorithm 5.4.1 will always converge to a global state with rank at most $k$ if such a state exists. More generally, derive a convergent algorithm for finding a minimum rank or low rank global states with prescribed partial states in a multipartite system.

Finally, in section 5.6, we gave a characterization of a product of two positive semidefinite contractions that can be formulated as a problem of existence of an element in the intersection of two convex sets. This in turn, can be solved using alternating projections. It is of great interest to find a characterization of two positive semidefinite contractions that is easier to check. The set of matrices that can be written as a product of a finite number of positive semidefinite matrices has been completely characterized but the set of matrices that can be written as a product of a finite number of positive contractions is not yet completely understood. This is a possible future research direction one can look into.
CHAPTER 6

Minkowski product of convex sets
and product numerical range*

6.1 Introduction

Let $K_1, K_2$ be compact convex sets in $\mathbb{C}$. We study the Minkowski product of the sets defined and denoted by

$$K_1 K_2 = \{ab : a \in K_1, b \in K_2\}.$$

This topic arises naturally in many branches of research. For example, in numerical analysis, computations are subject to errors caused by the precision of the machines and round-off errors. Sometimes measurement errors in the raw data may also affect the accuracy. So, when two real numbers $a$ and $b$ are multiplied, the actual answer may actually be the product of numbers in two intervals containing $a$ and $b$; when two complex numbers $a$ and $b$ are multiplied, the actual answer may actually be the product of numbers from

*The material in this chapter is contained in the paper [63], which is a joint work of C.K. Li, Y.T. Poon, K.Z. Wang and the author.
two regions in the complex plane. The study of the product set also has applications in computer-aided design, reflection and refraction of wavefronts in geometrical optics, stability characterization of multi-parameter control systems, and the shape analysis and procedural generation of two-dimensional domains. For more discussion about these topics, see [37] and the references therein. Another application comes from the study of quantum information science. For a complex $n \times n$ matrix $A$, its numerical range is defined and denoted by

$$W(A) = \{ \langle x|A|x \rangle x \in \mathbb{C}^n, \langle x|x \rangle = 1 \}.$$ 

The numerical range of a matrix is always a compact convex set and carries a lot of information about the matrix, e.g., see [50].

Denote by $X \otimes Y$ the Kronecker product of two matrices or vectors. Then the product numerical range of $T \in \mathbb{C}^{m \times m} \otimes \mathbb{C}^{n \times n} \equiv \mathbb{C}^{mn \times mn}$ is defined by

$$W^{\otimes}(T) = \{ \langle x| \otimes \langle y|T(|x| \otimes |y|) : |x\rangle \in \mathbb{C}^m, |y\rangle \in \mathbb{C}^n, \langle x|x \rangle = \langle y|y \rangle = 1 \},$$

which is a subset of $W(T)$. In the context of quantum information science, this set corresponds to the collection of $\langle T, P \otimes Q \rangle$, where $P \in \mathbb{C}^{m \times m}, Q \in \mathbb{C}^{n \times n}$ are pure states (i.e., rank one orthogonal projections). In particular, if $T = A \otimes B$ with $(A, B) \in \mathbb{C}^{m \times m} \times \mathbb{C}^{n \times n}$, then $W^{\otimes}(A \otimes B) = \{ \langle x| \otimes \langle y|(A \otimes B)|x\rangle \otimes |y\rangle : x \in \mathbb{C}^m, y \in \mathbb{C}^n, \langle x|x \rangle = \langle y|y \rangle = 1 \} = W(A)W(B)$. So, the set $W^{\otimes}(A \otimes B)$ is just the Minkowski product of the two compact convex sets $W(A)$ and $W(B)$. In particular, the following was proved in [81]. (Their proofs concern the product numerical range that can be easily adapted to general compact convex sets.)

**Proposition 6.1.1.** Suppose $K_1, K_2$ are compact convex sets in $\mathbb{C}$.

(a) The set $K_1K_2$ is simply connected.
(b) If $0 \in K_1 \cup K_2$, then $K_1K_2$ is star-shaped with 0 as a star center.

It was conjectured in [81] that the set $K_1K_2$ is always star-shaped. In this paper, we will show that the conjecture is not true in general (Section 6.3.1). The proof depends on a detailed analysis of the product sets of two closed line segments (Section 6.2). Then we obtain some conditions under which the product set of two convex polygons is star-shaped (Sections 6.3.2). Furthermore, we show that $K_1K_2$ is star-shaped for any compact convex set $K_2$ if $K_1$ is a closed line segment or a closed circular disk in Sections 6.4 and 6.5. Some additional results and open problems are mentioned in Section 6.6. In particular, in Theorem 6.6.2, we will improve the following result, which is a consequence of the simply connectedness of $K_1K_2$ [81, Proposition 1].

**Proposition 6.1.2.** Suppose $K_1$, $K_2$ are compact convex sets in $\mathbb{C}$ and $p \in K_1K_2$. Then $K_1K_2$ is star-shaped with $p$ as a star center if and only if $K_1K_2$ contains the line segment joining $p$ to $ab$ for any $a \in \partial K_1$ and $b \in \partial K_2$.

In our discussion, the convex hull of the set $\{z_1, \ldots, z_m\} \subseteq \mathbb{C}$ will be denoted by $\text{Co}(z_1, z_2, \ldots, z_m)$. In particular, $\text{Co}(z_1, z_2)$ is the line segment in $\mathbb{C}$ joining $z_1, z_2$. Also, if $K_1 = \{\alpha\}$, we write $K_1K_2 = \alpha K_2$.

### 6.2 The product set of two segments

We first give a complete description of the set $K_1K_2$ when $K_1 = \text{Co}(\alpha_1, \alpha_2)$ and $K_2 = \text{Co}(\beta_1, \beta_2)$ are two line segments. McAllister has plotted some examples in [77] but the analysis is not complete. In the context of product numerical range, it is known, see for example, [61, Theorem 4.3], that $W(T)$ is a line segment if and only if $T$ is normal with collinear eigenvalues. In such a case, $W(T) = W(T_0)$ for a normal matrix $T_0 \in \mathbb{C}^{2 \times 2}$ having the two endpoints of $W(T)$ as its eigenvalues. Thus, the study of $K_1K_2$ when
$K_1, K_2$ are close line segments corresponds to the study of $W^\otimes (A \otimes B) = W(A)W(B)$ for $A \in \mathbb{C}^{m \times m}, B \in \mathbb{C}^{n \times n}$ with special structure, and $W^\otimes (A \otimes B) = W^\otimes (A_0 \otimes B_0)$ for some normal matrices $A_0, B_0 \in \mathbb{C}^{2 \times 2}$. We have the following result.

**Theorem 6.2.1.** Let $K_1 = \text{Co}(\alpha_1, \alpha_2)$ and $K_2 = \text{Co}(\beta_1, \beta_2)$ be two line segments in $\mathbb{C}$. Then $K_1 K_2$ is a star-shaped subset of $\text{Co}(\alpha_1 \beta_1, \alpha_1 \beta_2, \alpha_2 \beta_1, \alpha_2 \beta_2)$.

In general, $\text{Co}(\alpha_1, \ldots, \alpha_n) \text{Co}(\beta_1, \ldots, \beta_m) \subseteq \text{Co}(\alpha_1 \beta_1, \alpha_1 \beta_2, \ldots, \alpha_i \beta_j, \ldots, \alpha_n \beta_m)$ because

$$
\left( \sum_i p_i \alpha_i \right) \left( \sum_j q_j \beta_j \right) = \left( \sum_{i,j} p_i q_j \alpha_i \beta_j \right)
$$

and $\sum_i p_i = 1$ and $\sum_j q_j = 1$ imply that $\sum_{i,j} p_i q_j = 1$. The key point of Theorem 6.2.1 is the star-shapedness of the product of two line segments in $\mathbb{C}$.

We will give a complete description of the set $K_1 K_2$ in the following. If one or both of the line segments $K_1, K_2$ lie(s) in a line passing through origin, the description is relatively easy as shown in the following.

**Proposition 6.2.2.** Let $K_1 = \text{Co}(\alpha_1, \alpha_2)$ and $K_2 = \text{Co}(\beta_1, \beta_2)$ be two line segments in $\mathbb{C}$.

1. If both $\text{Co}(0, \alpha_1, \alpha_2)$ and $\text{Co}(0, \beta_1, \beta_2)$ are line segments, then $K_1 K_2$ is the line segment $\text{Co}(\alpha_1 \beta_1, \alpha_1 \beta_2, \alpha_2 \beta_1, \alpha_2 \beta_2)$.

2. Suppose $\text{Co}(0, \alpha_1, \alpha_2)$ is a line segment and $\text{Co}(0, \beta_1, \beta_2)$ is not.

   (2.a) If $0 \in \text{Co}(\alpha_1, \alpha_2)$, then $K_1 K_2 = \text{Co}(0, \alpha_1 \beta_1, \alpha_1 \beta_2) \cup \text{Co}(0, \alpha_2 \beta_1, \alpha_2 \beta_2)$ is the union of two triangles (one of them may degenerate to \{0\}) meeting at 0, which is the star center of $K_1 K_2$.

   (2.b) If $0 \notin \text{Co}(\alpha_1, \alpha_2)$ then $K_1 K_2 = \text{Co}(\alpha_1 \beta_1, \alpha_1 \beta_2, \alpha_2 \beta_1, \alpha_2 \beta_2)$.
Proof.

1. There exist \( \alpha, \beta, a_1, a_2, b_1, b_2 \in \mathbb{R} \) such that \( K_1 = \{ re^{i\alpha} : a_1 \leq r \leq b_1 \} \) and \( K_2 = \{ re^{i\beta} : a_2 \leq r \leq b_2 \} \). So, we have

\[
K_1 K_2 = \{ re^{i(\alpha+\beta)} : a_3 \leq r \leq b_3 \}
\]

for some \( a_3, b_3 \in \mathbb{R} \).

(2.a) Evidently, \( K_1 K_2 = Co(0, \alpha_1) K_2 \cup Co(0, \alpha_2) K_2 \) and \( Co(0, \alpha_i) K_2 \subseteq Co(0, \alpha_i \beta_1, \alpha_i \beta_2) \) for \( i = 1, 2 \). We are going to show that \( Co(0, \alpha_i) Co(\beta_1, \beta_2) = Co(0, \alpha_i \beta_1, \alpha_i \beta_2) \) for \( i = 1, 2 \).

Clearly, \( 0 \in Co(0, \alpha_i) Co(\beta_1, \beta_2) \). If \( x \in Co(0, \alpha_i \beta_1, \alpha_i \beta_2) \setminus \{0\} \), then there exist \( s, t \geq 0 \) with \( 0 < s + t \leq 1 \) such that \( x = sa_i \beta_1 + t\alpha_i \beta_2 \). Therefore, \( x = ab \), where

\[
a = (s + t)\alpha_i \in Co(0, \alpha_i) \quad \text{and} \quad b = \frac{s}{s+t} \beta_1 + \frac{t}{s+t} \beta_2 \in Co(\beta_1, \beta_2)
\]

Thus,

\[
Co(0, \alpha_i) Co(\beta_1, \beta_2) = Co(0, \alpha_i \beta_1, \alpha_i \beta_2)
\]

and

\[
K_1 K_2 = Co(0, \alpha_2 \beta_1, \alpha_2 \beta_2) \cup Co(0, \alpha_1 \beta_1, \alpha_1 \beta_2).
\]

(2.b) Let \( x \in Co(\alpha_1 \beta_1, \alpha_1 \beta_2, \alpha_2 \beta_1, \alpha_2 \beta_2) \). Then \( x = sa_1 \beta_1 + t\alpha_1 \beta_2 + u\alpha_2 \beta_1 + v\alpha_2 \beta_2 \) for some \( s, t, u, v \geq 0 \) with \( s + t + u + v = 1 \). Since \( 0 \notin Co(\alpha_1, \alpha_2) \), \( \alpha_2 = k\alpha_1 \) for some \( k > 0 \), then \( x = (pa_1 + (1-p)\alpha_2)(q\beta_1 + (1-q)\beta_2) \), where

\[
p = s + t, \quad q = \frac{s + uk}{s + t + k(u + v)} \in [0, 1].
\]

\( \square \)

The situation is more involved if neither \( Co(0, \alpha_1, \alpha_2) \) nor \( Co(0, \beta_1, \beta_2) \) is a line seg-
ment. To describe the shape of $K_1K_2$ in such a case, we put the two segments in a certain “canonical” position. More specifically, the next proposition shows that we can find $\alpha_0$ and $\beta_0 \in \mathbb{C}$ such that $\alpha_0^{-1}K_1$ and $\beta_0^{-1}K_2$ lie in the vertical line $\{z \in \mathbb{C} : \text{Re}(z) = 1\}$.

**Proposition 6.2.3.** Let $K_1 = \text{Co}(\alpha_1, \alpha_2)$ and $K_2 = \text{Co}(\beta_1, \beta_2)$ be two line segments in $\mathbb{C}$ such that neither $\text{Co}(0, \alpha_1, \alpha_2)$ nor $\text{Co}(0, \beta_1, \beta_2)$ is a line segment. Let

$$
\alpha_0 = \frac{\alpha_1 \alpha_2 - \alpha_2 \alpha_1}{2(\alpha_2 - \alpha_1)} \quad \text{and} \quad \beta_0 = \frac{\beta_1 \beta_2 - \beta_2 \beta_1}{2(\beta_2 - \beta_1)} \quad (6.1)
$$

Then $\alpha_0$ (respectively, $\beta_0$) is the point on the line passing through $\alpha_1$ and $\alpha_2$ (respectively, $\beta_1$ and $\beta_2$) closest to 0. We have

$$
\frac{\alpha_1}{\alpha_0} = 1 + a_1i, \quad \frac{\alpha_2}{\alpha_0} = 1 + a_2i, \quad \frac{\beta_1}{\beta_0} = 1 + b_1i, \quad \frac{\beta_2}{\beta_0} = 1 + b_2i \quad (6.2)
$$

for some $a_1$, $a_2$, $b_1$ and $b_2 \in \mathbb{R}$.

**Proof.** The line passing through $\alpha_1$ and $\alpha_2$ is given by the parametric equation $r(t) = \alpha_1 + t(\alpha_1 - \alpha_2)$, $t \in \mathbb{R}$. $\alpha_0$ in (6.1) is obtained by minimizing $|r(t)|^2$. Similarly, we have $\beta_0$. By direct calculation we have (6.2) with

$$
a_1 = \frac{\alpha_1 \overline{\alpha_2} + \alpha_2 \overline{\alpha_1} - 2|\alpha_1|^2}{i(\alpha_1 \overline{\alpha_2} - \alpha_2 \overline{\alpha_1})}, \quad a_2 = \frac{\alpha_1 \overline{\alpha_2} + \alpha_2 \overline{\alpha_1} - 2|\alpha_2|^2}{i(\alpha_2 \overline{\alpha_1} - \alpha_1 \overline{\alpha_2})},
$$

$$
b_1 = \frac{\beta_1 \overline{\beta_2} + \beta_2 \overline{\beta_1} - 2|\beta_1|^2}{i(\beta_1 \overline{\beta_2} - \beta_2 \overline{\beta_1})}, \quad b_2 = \frac{\beta_1 \overline{\beta_2} + \beta_2 \overline{\beta_1} - 2|\beta_2|^2}{i(\beta_2 \overline{\beta_1} - \beta_1 \overline{\beta_2})} \quad \square
$$

We can now describe $K_1K_2$ for two line segments $K_1 = \text{Co}(\alpha_1, \alpha_2)$ and $K_2 = \text{Co}(\beta_1, \beta_2)$ in the “canonical” position. In the following theorem, because $\text{Co}(\alpha_1, \alpha_2)\text{Co}(\beta_1, \beta_2)$ is a simply connected set, we focus on the description of the boundary and the set of star centers of $K_1K_2$. 

158
Theorem 6.2.4. Let \( K_1 = \text{Co}(\alpha_1, \alpha_2) \) and \( K_2 = \text{Co}(\beta_1, \beta_2) \) with \( \alpha_1 = 1 + ia_1, \alpha_2 = 1 + ia_2, \beta_1 = 1 + ib_1, \beta_2 = 1 + ib_2 \) such that \( a_1 < a_2 \) and \( b_1 < b_2 \). Assume \( a_1 \leq b_1 \); otherwise, interchange the roles of \( K_1 \) and \( K_2 \). Define \( C = \{(1 + si)^2 : s \in \mathbb{R}\} \). Then one of the following holds.

(a) \( a_1 < a_2 \leq b_1 < b_2 \). Then \( K_1 K_2 \) is the convex quadrilateral \( \text{Co}(\alpha_1 \beta_1, \alpha_1 \beta_2, \alpha_2 \beta_1, \alpha_2 \beta_2) \), which will degenerate to the triangle \( \text{Co}(\alpha_1 \beta_1, \alpha_1 \beta_2, \alpha_2 \beta_2) \) if \( a_2 = b_1 \); see Figure 6.1a.

(b) \( a_1 \leq b_1 < a_2 \leq b_2 \). Then \( K_1 K_2 \subseteq \text{Co}(\alpha_1 \beta_1, \alpha_1 \beta_2, \alpha_2 \beta_2) \), and the boundary of \( K_1 K_2 \) consists of the line segments \( \text{Co}(\alpha_2^2, \alpha \beta_2), \text{Co}(\alpha \beta_2, \alpha_1 \beta_2), \text{Co}(\alpha_1 \beta_2, \alpha_1 \beta_1), \text{Co}(\alpha_1 \beta_1, \beta_1^2) \), and the curve \( E = \{(1 + si)^2 : s \in [b_1, a_2]\} \subseteq C \). Here, \( \text{Co}(\alpha_2^2, \alpha_2 \beta_2) \) lies on the tangent line of the curve \( E \) at \( \alpha_2^2 \), and \( \text{Co}(\beta_1^2, \alpha_1 \beta_1) \) lies on the tangent line of the curve \( E \) at \( \beta_1^2 \). The set of star centers equals \( \text{Co}(\alpha_1, \beta_1)\text{Co}(\alpha_2, \beta_2) \), which may be a quadrilateral, a line or a point; see Figure 6.1b.

(c) Suppose \( a_1 < b_1 < b_2 < a_2 \). Then the boundary of \( K_1 K_2 \) consists of the line segments \( \text{Co}(\beta_2^2, \alpha_2 \beta_2), \text{Co}(\alpha_2 \beta_2, \alpha_2 \beta_1), \text{Co}(\alpha_2 \beta_1, \beta_1 \beta_2), \text{Co}(\beta_1 \beta_2, \alpha_1 \beta_2), \text{Co}(\alpha_1 \beta_2, \alpha_1 \beta_1), \text{Co}(\beta_1^2, \alpha_1 \beta_1) \) and the curve segment \( \{(1 + si)^2 : s \in [b_1, b_2]\} \subseteq C \). Here, \( \text{Co}(\beta_2^2, \alpha_2 \beta_2) \) lies on the tangent line of the curve \( C \) at \( \beta_2^2 \), and \( \text{Co}(\beta_1^2, \alpha_1 \beta_1) \) lies on the tangent line of the curve \( C \) at \( \beta_1^2 \). The unique star center is \( \beta_1 \beta_2 \); see Figure 6.1c.

To prove Theorem 6.2.4, we need the following lemma that treat some special cases of the theorem. It turns out that these special cases are the building blocks for the general case.

Lemma 6.2.5. Let \( a_1 < a_2 \leq b_1 < b_2 \). Then

(a) \( \text{Co}(1 + a_1 i, 1 + a_2 i)\text{Co}(1 + b_1 i, 1 + b_2 i) \) is the quadrilateral (or triangle if \( a_2 = b_1 \)),

\[
K = \text{Co} \left( (1 + a_1 i)(1 + b_1 i), (1 + a_1 i)(1 + b_2 i), (1 + a_2 i)(1 + b_1 i), (1 + a_2 i)(1 + b_2 i) \right).
\]
(a) \( a_1 < a_2 \leq b_1 < b_2 \)

(b) \( b_1 \leq a_1 < a_2 \leq b_2 \)

(c) \( a_1 < b_1 < b_2 < a_2 \)

FIG. 6.1: Three cases of the Minkowski product of two lines described in Theorem 6.2.4.
(b) $Co(1 + a_1i, 1 + a_2i)Co(1 + a_1i, 1 + a_2i)$ is the simply connected region bounded by the line segments

$$L_1 = Co\left((1 + a_1i)^2, (1 + a_1i)(1 + a_2i)\right), \quad L_2 = Co\left((1 + a_2i)^2, (1 + a_1i)(1 + a_2i)\right),$$

and the curve $E = \{(1 + si)^2 : s \in [a_1, a_2]\}$. The set $L_1$ is a segment of the tangent line of $E$ at $(1 + a_1i)^2$, and $L_2$ is a segment of the tangent line of $E$ at $(1 + a_2i)^2$.

Proof. (a) Suppose $a_j = 1 + a_ji$ and $\beta_j = 1 + b_ji$ for $j = 1, 2$ are such that $a_1 < a_2 \leq b_1 < b_2$. Let $K_1 = Co(\alpha_1, \alpha_2)$ and $K_2 = Co(\beta_1, \beta_2)$. It suffices to show that the union of the line segments

$$\ell_1 = \beta_2K_1, \quad \ell_2 = \beta_1K_1, \quad \ell_3 = \alpha_2K_2, \quad \ell_4 = \alpha_1K_2$$

forms the boundary of the quadrilateral (or triangle) $K$, that is, the union is a simple closed curve. By simply connectedness and the fact that $K_1K_2$ is a subset of $K$, we get the desired conclusion. For the convenience of discussion, we will identify $x + iy \in \mathbb{C}$ with $(x, y) \in \mathbb{R}^2$ and $(x, y, 0) \in \mathbb{R}^3$. Note that since $\arg(\alpha_1\beta_1) < \arg(\alpha_2\beta_1), \arg(\alpha_1\beta_2) < \arg(\alpha_2\beta_2)$, it suffices to show that $\alpha_1\beta_2$ and $\alpha_2\beta_1$ are on opposite sides of the line $\ell$ passing through $\alpha_1\beta_1$ and $\alpha_2\beta_2$. This is true if and only if the cross product $(\alpha_2\beta_1 - \alpha_2\beta_2) \times (\alpha_1\beta_1 - \alpha_2\beta_2)$ and $(\alpha_1\beta_2 - \alpha_2\beta_2) \times (\alpha_1\beta_1 - \alpha_2\beta_2)$ are pointing in opposite directions, that is

$$\det \begin{bmatrix}
\text{Re}(\alpha_2\beta_1 - \alpha_2\beta_2) & \text{Re}(\alpha_1\beta_1 - \alpha_2\beta_2) \\
\text{Im}(\alpha_2\beta_1 - \alpha_2\beta_2) & \text{Im}(\alpha_1\beta_1 - \alpha_2\beta_2)
\end{bmatrix} \cdot \det \begin{bmatrix}
\text{Re}(\alpha_1\beta_2 - \alpha_2\beta_2) & \text{Re}(\alpha_1\beta_1 - \alpha_2\beta_2) \\
\text{Im}(\alpha_1\beta_2 - \alpha_2\beta_2) & \text{Im}(\alpha_1\beta_1 - \alpha_2\beta_2)
\end{bmatrix} \leq 0$$

The expression on the left hand side is

$$[(b_1 - b_2)(a_2 - a_1)(a_2 - b_1)] \cdot [(b_1 - b_2)(a_2 - a_1)(b_2 - a_1)] = (b_1 - b_2)^2(a_2 - a_1)^2(a_2 - b_1)(b_2 - a_1)$$

Since $a_2 \leq b_1$ and $b_2 > a_1$, then we are done.
To prove (b), first note that $L_1, L_2$ and $E$ are clearly in $K_1$. Direct calculation shows that $L_1$ with equation $x = 1 - a_1(y - a_1)$ and $L_2$ with equation $x = 1 - a_2(y - a_2)$ are tangent to the parabola $E$ with equation $x = 1 - \frac{y^2}{4}$ at the points $(1 - a_1^2, 2a_1)$ and $(1 - a_2^2, 2a_2)$ respectively.

Since $K_1$ is simply connected, the region

$$S = \left\{ x + iy : \frac{y^2}{4} \leq x \leq 1 - a_1(y - a_1), 1 - a_2(y - a_2) \right\}, \quad (6.3)$$

which is the region enclosed by $L_1, L_2$ and $E$ is a subset of $K_1$. Now, suppose $x + iy \in K_1$. Then there exist $r$ and $s$ with $a_1 \leq r, s \leq a_2$ such that

$$x + iy = (1 + ir)(1 + is) = 1 - rs + i(r + s).$$

Note that

$$x = 1 - rs \geq 1 - \frac{1}{4}(r + s)^2 = 1 - \frac{y^2}{4}$$

always holds. Also, if $a \leq t \leq b$, then $(a + b - t)t \geq ab$. Since

$$a_1 \leq r \leq s + r - a_1 \text{ and } s + r - a_2 \leq r \leq a_2,$$

we have $rs \geq a_1(s + r - a_1), a_2(s + r - a_2)$. Hence,

$$x = 1 - rs \leq 1 - a_1(r + s - a_1) = 1 - a_1(y - a_1), \quad \text{and}$$

$$x = 1 - rs \leq 1 - a_2(r + s - a_2) = 1 - a_2(y - a_2).$$

This shows that $K_1$ lies inside $S$. Thus $K_1 = S$. \qed

Proof of Theorem 6.2.4. Suppose $K_1 = \text{Co}(1 + i a_1, 1 + i a_2)$ and $K_2 = \text{Co}(1 + i b_1, a + i b_2)$.
such that \( a_1 \leq a_2, b_1 \leq b_2 \). We show that if \( K_1K_2 \) can be written as the union of subsets of the form in Lemma 6.2.5. In fact, if \([a_1, a_2] \cap [b_1, b_2] = [c_1, c_2] \), then

\[
K_1K_2 = (\alpha_0 \beta_0) [(AC) \cup (AB) \cup (CC) \cup (CB)],
\]

where \( C = \text{Co}(1+c_1i, 1+c_2i) \), \( B = \text{Co}(1+b_1i, 1+b_2i) \setminus C \) and \( A = \text{Co}(1+a_1i, 1+a_2i) \setminus C \).

By Lemma 6.2.5, we get the conclusion.

By Theorem 6.2.4, we have the following corollary giving information about the star center of the product of two line segments without putting them in the “canonical” position.

**Corollary 6.2.6.** Let \( K_1 = \text{Co}(\alpha_1, \alpha_2) \) and \( K_2 = \text{Co}(\beta_1, \beta_2) \), where \( \alpha_1, \alpha_2, \beta_1, \beta_2 \in \mathbb{C} \) such that \( \arg(\alpha_1) < \arg(\alpha_2) < \arg(\alpha_1) + \pi \) and \( \arg(\beta_1) < \arg(\beta_2) < \arg(\beta_1) + \pi \). Then \( K_1K_2 \) is star-shaped and one of the following holds.

(a) There exists \( \xi \in \mathbb{C} \) such that \( \xi K_1 \subseteq K_2 \). Equivalently, the segments \( \text{Co}(\alpha_1 \beta_1, \alpha_1 \beta_2) \) and \( \text{Co}(\alpha_2 \beta_1, \alpha_2 \beta_2) \) intersect at \( \xi \alpha_1 \alpha_2 \). In this case, \( \xi \alpha_1 \alpha_2 \) is the unique star-center of \( K_1K_2 \).

(b) There exists \( \xi \in \mathbb{C} \) such that \( \xi K_2 \subseteq K_1 \). Equivalently, the segments \( \text{Co}(\alpha_1 \beta_1, \alpha_2 \beta_1) \) and \( \text{Co}(\alpha_1 \beta_2, \alpha_2 \beta_2) \) intersect at \( \xi \beta_1 \beta_2 \). In this case, \( \xi \beta_1 \beta_2 \) is the unique star-center of \( K_1K_2 \).

(c) Condition (a) and (b) do not hold, and every point in \( \text{Co}(\beta_1 \alpha_2, \beta_2 \alpha_1) \) is a star center of \( K_1K_2 \).
6.3 The product set of two convex polygons

In this section, we study the product set of two convex polygons (including interior). It is known that for every convex polygon $K_1$ with vertices $\mu_1, \ldots, \mu_n$, then $K_1 = W(T)$ for $T = \text{diag}(\mu_1, \ldots, \mu_n) \in \mathbb{C}^{n \times n}$. In Section 6.3.1, we will show that the product set of two convex polygons may not be star-shaped. In particular, we have a product set of two triangles that are not star-shaped. This gives a negative answer to the conjecture in [81].

6.3.1 Products of polygons that are not star-shaped

In this subsection, we show that there are examples $K_1$ and $K_2$ such that $K_1 \ast K_2$ is not star-shaped. The first example has the form $K_1 = K_2 = \text{Co}(\alpha_1, \overline{\alpha}_1, \alpha_2)$, where $\alpha_2 \notin \mathbb{R}$. One can regard $K_1 = W(T)$ with $T = \text{diag}(\alpha_1, \overline{\alpha}_1, \alpha_2) \in \mathbb{C}^{3 \times 3}$ so that the set $W^\otimes(T \otimes T) = W(T)W(T)$ is not star-shaped. We can construct another example of the form $K_1 = K_2 = \text{Co}(\alpha_1, \overline{\alpha}_1, \alpha_2, \overline{\alpha}_2)$, which is symmetric about the real axis, such that $K_1 \ast K_2$ is not star-shaped. One can regard $K_1 = W(A)$ for a real normal matrix $A \in \mathbb{C}^{4 \times 4}$ with eigenvalues $\alpha_1, \overline{\alpha}_1, \alpha_2, \overline{\alpha}_2$ so that $W^\otimes(A \otimes A)$ is not star-shaped.

Example 6.3.1. Let $K_1 = \text{Co}(e^{i\pi/3}, e^{-i\pi/3}, 0.95e^{i\pi/4})$. Then $K_1 \ast K_1$ is not star-shaped.

Proof. Let $\alpha_1 = e^{i\pi/3}$ and $\alpha_2 = 0.95e^{i\pi/4}$, then $1 = \alpha_1 \overline{\alpha}_1$, $0.95^2 i = \alpha_2^2 \in K_1 K_1$. We are going to show that a) if $s$ is a star center of $K_1 \ast K_1$, then $s = 1$ and b) $(1 - t) + t0.95^2 i \notin K_1 \ast K_1$ for all $t \in (0, 1)$.

Let $S$ be a closed and bounded subset of $\mathbb{C}$, with $0 \notin S$. Suppose $t \in \mathbb{R}$ and $S \cap \{re^{it} : r > 0\} \neq \emptyset$. Let $\rho_0^S(t) = \min\{r > 0 : re^{it} \in S\}$ and $\rho_0^S(t) = \max\{r > 0 : re^{it} \in S\}$.

Let $L_1 = \text{Co}(\alpha_1, \overline{\alpha}_1)$, $S_1 = K_1 K_1$ and $S_2 = L_1 L_1$. Since $\rho_0^{K_1}(\theta) = \rho_0^{L_1}(\theta)$ for $-\pi/3 \leq \theta \leq \pi/3$, it follows that $\rho_0^{S_1}(\theta) = \rho_0^{S_2}(\theta)$ for $-2\pi/3 \leq \theta \leq 2\pi/3$.

Note that $x + iy \in S_2$ if and only if $4(x + iy) \in (2L_1)(2L_1)$. Then, applying Lemma
6.2.5 (b) to \(2L_1 = \text{Co}(1 - i\sqrt{3}, 1 + i\sqrt{3})\), we have

\[
S_2 = \{x + iy : 1 - 4y^2 \leq 4x \leq 1 - \sqrt{3}(4y - \sqrt{3}), \ 1 + \sqrt{3}(4y + \sqrt{3})\}
\]

![Diagram showing the plot of \(S_2 = L_1 L_1\), where \(L_1 = \text{Co}(e^{\frac{\pi}{3}}, e^{-i\frac{\pi}{3}})\).]

**FIG. 6.2:** Plot of \(S_2 = L_1 L_1\), where \(L_1 = \text{Co}(e^{\frac{\pi}{3}}, e^{-i\frac{\pi}{3}})\).

a) Note that \(\{\rho_{S_1}^z(\theta) : \theta \in [-2\pi/3, 2\pi/3]\} = \{\rho_{S_2}^z(\theta) : \theta \in [-2\pi/3, 2\pi/3]\} = \{z^2 : z \in L_1\}\). This means that the curve \(\{z^2 : z \in L_1\}\) is a boundary curve of \(S_2\). By Proposition 1.2, if \(s\) were a star-center of \(S_2\), then the segment \(\text{Co}(s, z^2)\) must be in \(S_2\) for any \(z \in L_1\).

If \(s = x + iy\) is a star center of \(S_1\), then we must have

\[
4x \geq 1 - \sqrt{3}(4y - \sqrt{3}), \ 1 + \sqrt{3}(4y + \sqrt{3}) \Rightarrow x \geq 1
\]

Since \(|z| \leq 1\) for all \(z \in S_1\), we have \(s = 1\).

b) Let \(L_2 = \text{Co}(\alpha_1, \alpha_2), L_3 = \text{Co}(\bar{\alpha}_1, \alpha_2)\). Then the boundary of the simply connected set \(S_1 = K_1K_1\) is a subset of \(\bigcup_{1 \leq i \leq j \leq 3} L_i L_j\).

Suppose \(0 < \theta < \frac{\pi}{2}\) and \(\rho_{S_1}^z(t) = r\). Then \(re^{i\theta} \in L_2 L_3 \cup L_3 L_3\). Direct calculation using Lemma 6.2.5 and Proposition 6.2.3 shows that \(\rho_{L_2L_3}(\theta), \rho_{L_3L_3}(\theta) < \rho_{\text{Co}(1, \alpha_2)}^z(\theta)\); see
We conclude that $K_1K_1$ is not star-shaped. \hfill $\Box$

Next, we modify Example 6.3.1 to Example 6.3.2 so that $\tilde{K}_1 = K_1(\alpha_1,\alpha_2,\bar{\alpha}_1,\bar{\alpha}_2)$ with $\alpha_1 = e^{i\pi/3}$ and $\alpha_2 = 0.95e^{i\pi/4}$. In this case, one can regard $K_1 = W(A)$ for some real symmetric $A \in \mathbb{C}^{4 \times 4}$. The product set $K_1K_2$ will be larger than the product set considered in Example 6.3.1. Never-the-less, we can analyze the product of the sets $L_iL_j$ for $i, j = 1, 2, 3, 4$, where $L_1 = \text{Co}(\alpha_1,\bar{\alpha}_1)$, $L_2 = \text{Co}(\alpha_1,\alpha_2)$, $L_3 = \text{Co}(\alpha_2,\bar{\alpha}_2)$, $L_4 = \text{Co}(\bar{\alpha}_2,\bar{\alpha}_1)$ so that $\bigcup_{1 \leq i < j \leq 4} L_iL_j$ contains the boundary of the simply connected set $K_1K_1$. Again one can show that the part of the boundary $\{z^2 : z \in \text{Co}(\alpha_1,\bar{\alpha}_1)\}$ of $L_1L_1$ is also part of the boundary of $K_1K_1$ so that $1 = \alpha_1\bar{\alpha}_1 \in K_1K_1$ is the only possible candidate to serve as a star-center for $K_1K_1$. However, none of the set $L_iL_j$ contains the set $\{t + (1 - t)0.95^2i : 0 < t < 1/3\}$. Thus, the line segment joining 1 and $\alpha_2^2 = 0.95^2i$ is not in $K_1K_1$. Hence, 1 is not the star center of $K_1K_1$, and $K_1K_1$ is not star-shaped.

**Example 6.3.2.** Let $K_1 = \text{Co}(e^{i\pi/4},e^{-i\pi/4},0.95e^{i\pi/2},0.95e^{-i\pi/2})$. Then $K_1$ is symmetric about the $x$-axis but $P = K_1K_1$ is not star-shaped (see Figure 6.4).
6.3.2 A necessary and sufficient condition

In the following result, we establish a necessary and sufficient condition for the product of two polygons to be a star-shaped set.

**Theorem 6.3.3.** Let $K_1 = \text{Co}(a_1, \ldots, a_n)$ and $K_2 = \text{Co}(b_1, \ldots, b_m)$. Then $K_1K_2$ is star-shaped if and only if there is $p \in K_1K_2$ such that $\text{Co}(p, a_ib_j) \subseteq K_1K_2$ for all $1 \leq i \leq n$ and $1 \leq j \leq m$.

**Proof.** Assume that $K_1 = \text{Co}(\alpha_1, \ldots, \alpha_n)$ and $K_2 = \text{Co}(\beta_1, \ldots, \beta_m)$. From Proposition 6.1.1 (a), we only need to prove that given any $1 \leq i_1, i_2 \leq n$ and $1 \leq j_1, j_2 \leq m$, $\text{Co}(p, q) \subseteq K_1K_2$ for all $q \in \text{Co}(\alpha_{i_1}, \alpha_{i_2})\text{Co}(\beta_{j_1}, \beta_{j_2})$. Without loss of generality, we may assume that for $r = 1, 2$, $i_r = j_r = r$, $\alpha_r = 1 + ia_r$ and $\beta_r = 1 + ib_r$ satisfy one of the conditions (a), (b) or (c) in Theorem 6.2.4.

Since $\text{Co}(p, \alpha_r \beta_t) \subseteq K_1K_2$ for $r, t = 1, 2$, by the fact that $K_1K_2$ is simply connected, we see that

$$K = \text{Co}(p, \alpha_1 \beta_1, \alpha_1 \beta_2) \cup \text{Co}(p, \alpha_2 \beta_1, \alpha_2 \beta_2) \cup \text{Co}(p, \alpha_1 \beta_1, \alpha_2 \beta_1) \cup \text{Co}(p, \alpha_1 \beta_2, \alpha_2 \beta_2) \subseteq K_1K_2.$$
If Co\((\alpha_1, \alpha_2)\)Co\((\beta_1, \beta_2)\) is convex, then Co\((p, q) \subseteq K\) for all \(q \in \text{Co}(\alpha_1, \alpha_2)\text{Co}(\beta_1, \beta_2)\).

If Co\((\alpha_1, \alpha_2)\)Co\((\beta_1, \beta_2)\) is not convex, then \(a_1, a_2, b_1\) and \(b_2\) satisfy conditions (b) or (c) in Theorem 6.2.4. Let \([a_1, a_2] \cap [b_1, b_2] = [c_1, c_2], \ C = \text{Co}(1+c_1i, 1+c_2i), \ B = \text{Co}(1+b_1i, 1+b_2i) \setminus C \) and \(A = \text{Co}(1+a_1i, 1+a_2i) \setminus C\). Since \(K_1K_2 = (AC) \cup (AB) \cup (CC) \cup (CB)\), and previous argument shows that Co\((p, q) \subseteq K_1K_2\) for all \(q \in (AC) \cup (AB) \cup (CB)\), it remains to show that Co\((p, q) \subseteq K_1K_2\) for all \(q \in \partial(CC)\). Let

\[
V = (1+c_1i)\text{Co}(1+c_1i, 1+c_2i) \cup (1+c_2i)\text{Co}(1+c_1i, 1+c_2i) \text{ and } U = \{(1+si)^2 : s \in (c_1, c_2)\}.
\]

Note that \(\partial(CC) = V \cup U\) and \(V \subseteq \text{Co}(\alpha_1\beta_1, \alpha_1\beta_2) \cup \text{Co}(\alpha_2\beta_1, \alpha_2\beta_2) \cup \text{Co}(\alpha_1\beta_1, \alpha_2\beta_1) \cup \text{Co}(\alpha_1\beta_2, \alpha_2\beta_2)\). So it remains to show that Co\((p, q) \subseteq K_1K_2\) for all \(q \in E^o = \{(1+si)^2 : s \in (c_1, c_2)\}\).

Suppose \(q \in E^o\). Let \(L\) be the tangent line to \(E^o\) at \(q\) and \(H\) the open half plane determined by \(L\) and contains 0 (see Figure 6.5).

Consider the following three cases:

**Case 1** If \(p \in H\), then there exists \(t > 1\) such that \(s = p + t(q-p) \in V\). Therefore, Co\((p, q) \subseteq \text{Co}(p, s) \subseteq K_1K_2\).

**Case 2** If \(p \in (C \setminus H) \cap (CC)\), then Co\((p, q) \subseteq (CC) \subseteq K_1K_2\) because \((C \setminus H) \cap (CC)\) is a triangular region containing \(q\).
Case 3 If \( p \in \mathbb{C} \setminus (H \cup (CC)) \), then there exists \( 0 < t < 1 \) such that \( s = p + t(q - p) \in V \). Therefore, \( \text{Co}(p, q) = \text{Co}(p, s) \cup \text{Co}(s, q) \subseteq K_1K_2 \). \( \square \)

We have the following consequence of Theorem 6.3.3.

**Corollary 6.3.4.** Let \( K_1 \) be a triangle set with \( K_1 = \overline{K}_1 \). Then \( K_1 = \text{Co}(r, a, \overline{a}) \) for some \( r \in \mathbb{R} \) and \( a \in \mathbb{C} \). The product set \( P = K_1K_1 \) is a star-shaped set with \( |a|^2 \) as a star center.

**Proof.** By Theorem 6.3.3, it suffices to show that for \( q \in \{r^2, ra, r\overline{a}, a^2, \overline{a}^2\} \), we have \( \text{Co}(|a|^2, q) \in P \).

1. For \( 0 \leq t \leq 1 \), let \( f(t) = (tr + (1 - t)a)(tr + (1 - t)\overline{a}) \in P \). Since \( f(0) = |a|^2 \) and \( f(1) = r^2 \), we have \( \text{Co}(|a|^2, r^2) \in P \).

2. \( \text{Co}(|a|^2, ra) = a \cdot \text{Co}(\overline{a}, r) \subseteq P \).

3. \( \text{Co}(|a|^2, r\overline{a}) = \overline{a} \cdot \text{Co}(a, r) \subseteq P \).

4. \( \text{Co}(|a|^2, a^2) = a \cdot \text{Co}(\overline{a}, a) \subseteq P \).

5. \( \text{Co}(|a|^2, \overline{a}^2) = \overline{a} \cdot \text{Co}(a, \overline{a}) \subseteq P \). \( \square \)

Suppose \( A \in \mathbb{C}^{n \times n} \) is a real matrix. Then \( W(A) \) is symmetric about the real axis. By Corollary 6.3.4, if \( A \in \mathbb{C}^{3 \times 3} \) is a real normal matrix, then \( W(A)W(A) \) is star-shaped. In fact, if \( A \) is Hermitian, then \( W(A)W(A) \) is convex; otherwise, \( |a|^2 \) is a star center, where \( a, \overline{a} \) are the complex eigenvalues of \( A \).

### 6.4 A line and a convex set

In this section, we consider the product of a line segment and a convex set. In the context of numerical range, we consider \( W(A)W(B) \), where \( A \) is a normal matrix with collinear eigenvalues, and \( B \) is a general matrix.
Theorem 6.4.1. Let $K_1 = \text{Co}(\alpha, \beta)$ for some $\alpha, \beta \in \mathbb{C}$ and $K_2$ be a compact convex sets in $\mathbb{C}$. Then $K_1K_2$ is star-shaped.

We begin with the following easy cases.

Proposition 6.4.2. Suppose that $K_1 = \text{Co}(\alpha, \beta)$ is a line segment and that $K_2$ is a (not necessarily compact) convex set.

(1) If $0 \in K_1 \cup K_2$, then $K_1K_2$ is star-shaped with 0 as a star center.

(2) If there is a nonzero $\xi_1 \in \mathbb{C}$ such that $\xi_1K_1 \subseteq (0, \infty)$, then $K_1K_2$ is convex.

(3) If there is a nonzero $\xi_1 \in \mathbb{C}$ such that $\xi_1K_1 \subseteq K_2$, then $K_1K_2$ is star-shaped with $\xi_1\alpha\beta$ as a star center.

Proof. (1) follows from Proposition 6.1.1 (b). For (2), we may assume that $\xi_1 = 1$. Then $K_1K_2 = \cup_{\alpha \leq t \leq \beta} tK_2$ is convex. Similarly for (3), we assume $\xi_1 = 1$. For every $p \in K_1$ and $q \in K_2$, we will show that

$$\text{Co}(\alpha\beta, pq) \subseteq \text{Co}(\alpha, \beta)\text{Co}(\alpha, \beta, q) \subseteq K_1K_2.$$ 

To this end, note that

$$\text{Co}(\alpha\beta, \alpha^2) = \alpha\text{Co}(\alpha, \beta) \quad \text{Co}(\alpha\beta, \beta^2) = \beta\text{Co}(\alpha, \beta)$$

$$\text{Co}(\alpha\beta, \alpha q) = \alpha\text{Co}(\beta, q) \quad \text{Co}(\alpha\beta, \beta q) = \beta\text{Co}(\alpha, q)$$

So, we have $\text{Co}(\alpha\beta, v) \in \text{Co}(\alpha, \beta)\text{Co}(\alpha, \beta, q)$ for any $v \in \{\alpha^2, \alpha\beta, \alpha q, \beta^2, \beta q\}$, which is the set of the product of vertexes of $\text{Co}(\alpha, \beta)$ and $\text{Co}(\alpha, \beta, q)$. By Theorem 6.3.3, $\text{Co}(\alpha, \beta)\text{Co}(\alpha, \beta, q)$ is star-shaped with $\alpha\beta$ as a star center. Thus,

$$\text{Co}(\alpha\beta, pq) \subseteq \text{Co}(\alpha, \beta)\text{Co}(\alpha, \beta, q) \subseteq K_1K_2.$$
If $\xi_1 \neq 1$, then $(\xi_1 \alpha)(\xi_1 \beta)$ is a star center of $(\xi_1 K_1)K_2 = \xi_1 K_1 K_2$ by the above argument. Thus, $\xi_1(\alpha \beta)$ is a star center of $K_1 K_2$.  

From now on, we will focus on convex sets $K_1$ and $K_2$ that do not satisfy the hypotheses in Proposition 6.4.2 (1) – (3). In particular, we may find $\xi_1$ and $\xi_2$ so that $\xi_1 K_1 = \text{Co}(\hat{a}, \hat{b})$ and $\xi_2 K_2$ is a compact convex set containing $\hat{c}, \hat{d}$ and lying in the cone $C = \{t_1 \hat{c} + t_2 \hat{d} : t_1, t_2 \geq 0\}$, where $\hat{a} = 1 + ia, \hat{b} = 1 + ib, \hat{c} = 1 + ic, \hat{d} = 1 + id$ with $a \leq b, c \leq d$. There could be five different configurations of the two sets $\xi_1 K_1$ and $\xi_2 K_2$ as illustrated in Figure 6.6. (Here, we assume that Proposition 6.4.2 (3) does not hold so that we do not have the case $c \leq a < b \leq d$.) If $K_1, K_2$ are put in these “canonical” positions, we can describe the star centers of $K_1 K_2$ in the next theorem.

FIG. 6.6: The following figures illustrate the canonical representations of a line segment $K_1 = \text{Co}(\hat{a}, \hat{b})$ and a convex set $K_2$ described in Theorem 6.4.3.
Theorem 6.4.3. Let \( \hat{a} = 1 + ia, \hat{b} = 1 + ib, \hat{c} = 1 + ic, \hat{d} = 1 + id \) with \( a \leq b, c \leq d \). Suppose \( K_1 = \text{Co}(\hat{a}, \hat{b}) \) and \( K_2 \) be a compact convex set containing \( \hat{c}, \hat{d} \) and lying in the cone

\[
C = \{ t_1\hat{c} + t_2\hat{d} : t_1, t_2 \geq 0 \}
\]
such that the hypotheses of Proposition 6.4.2 (1) – (3) do not hold. Then \( K_1 K_2 \) is star-shaped and one of the following holds.

(a) If \( a \leq b \leq c \leq d \), then \( \hat{b} \hat{c} \) is a star center.
(b) If \( a \leq c \leq b \leq d \), then \( \hat{b} \hat{c} \) is a star center.
(c) If \( a \leq c \leq d \leq b \), then \( \hat{c} \hat{d} \) is a star center.
(d) If \( c \leq a \leq d \leq b \), then \( \hat{a} \hat{d} \) is a star center.
(e) If \( c \leq d \leq a \leq b \), then \( \hat{a} \hat{d} \) is a star center.

We need some lemmas to prove Theorem 6.4.3.

Lemma 6.4.4. Suppose \( C = 1 + i\tan \theta_C, \ D = 1 + i\tan \theta_D \) and \( P = re^{i\theta_P} \) with \( r > 0 \), \( -\frac{\pi}{2} < \theta_C < \theta_P < \theta_D < \frac{\pi}{2} \). Let

\[
\frac{-i(P - C)}{|P - C|} = e^{i\theta_1} \quad \text{and} \quad \frac{i(P - D)}{|P - D|} = e^{i\theta_2} \quad \text{with} \quad -\frac{\pi}{2} < \theta_1, \theta_2 < \frac{\pi}{2}.
\]

Then there exists \( \xi_1, \xi_2 \) such that \( \xi_1 C = 1 + i\tan(\theta_C - \theta_1) \) and \( \xi_1 P = 1 + i\tan(\theta_P - \theta_1) \), \( \xi_2 D = 1 + i\tan(\theta_D - \theta_2) \) and \( \xi_2 P = 1 + i\tan(\theta_P - \theta_2) \).

Consequently, we have

1. If \( \text{Re} \ (P) \leq 1 \), then \( \theta_2 \leq 0 \leq \theta_1 \) and \( \theta_C - \theta_1 \leq \theta_P - \theta_1 \leq \theta_P - \theta_2 \leq \theta_D - \theta_2 \).
2. If \( \text{Re} \ (P) \geq 1 \), then \( \theta_1 \leq 0 \leq \theta_2 \) and \( \theta_C - \theta_1 \leq \theta_P - \theta_1 \) and \( \theta_P - \theta_2 \leq \theta_D - \theta_2 \).

Proof. First consider \( C \) and \( P \). Then \( \theta_1 \) is the angle from \( \overrightarrow{CD} \) to \( \overrightarrow{CP} \). Then the result follows from simple geometry.
On one also can calculate directly with $\xi_1 = \frac{\cos \theta_C}{\cos(\theta_C - \theta_1)} e^{-i\theta_1}$.

For the second statement, apply the above result on $\bar{D}$ and $\bar{P}$, the complex conjugate of $D$ and $P$.

Lemma 6.4.5. Suppose $a \leq c \leq d$, $p = t_1(1+ic) + t_2(1+id)$ is nonzero for some $t_1, t_2 \geq 0$, $K_1 = \text{Co}(1+ia, 1+id)$, and $K_2 = \text{Co}(1+ic, 1+id, p)$. Then $K_1K_2$ is star-shaped with $(1+ic)(1+id)$ as a star center.

Proof. Let $\hat{a} = 1+ia$, $\hat{c} = 1+ic$, $\hat{d} = 1+id$. By Theorem 6.3.3, it suffices to show that $\text{Co}(\hat{c}\hat{d}, uv) \subseteq K_1K_2$ for each pair of elements $(u, v)$ in $\{\hat{a}, \hat{d}\} \times \{\hat{c}, \hat{d}, p\}$. If $u = \hat{d}$, then $\text{Co}(\hat{c}\hat{d}, \hat{d}v) = \hat{d} \cdot \text{Co}(\hat{c}, v) \subseteq K_1K_2$. Similarly, if $u = \hat{c}$, then $\text{Co}(\hat{c}\hat{d}, \hat{c}v) = \hat{c} \cdot \text{Co}(\hat{d}, v) \subseteq K_1K_2$. Thus, the only nontrivial case is when $(u, v) = (\hat{a}, p)$.

By continuity, we may assume that $t_1, t_2 > 0$. We consider two cases.

Case 1 Suppose $\text{Re}(p) \leq 1$. Then by Lemma 6.4.4 and Theorem 6.2.4, $\text{Co}(\hat{a}, \hat{c})\text{Co}(p, \hat{d})$ is convex. So

$$\text{Co}(\hat{c}\hat{d}, \hat{a}p) \subseteq \text{Co}(\hat{a}, \hat{c})\text{Co}(p, \hat{d}) \subseteq K_1K_2.$$
and $\alpha_0 p = 1 + p_1 i$ such that $c_1 > c$. By Theorem 6.2.4, if $p_1 \geq d$, then $\hat{c}\hat{d}$ is a star center of $\text{Co}(\hat{a}, \hat{d})\text{Co}(\hat{c}, p)$. If $p_1 < d$, then $\text{Co}(\hat{a}\hat{c}, \hat{d}\hat{c})$ intersects $\text{Co}(\hat{a}\hat{p}, \hat{d}\hat{p})$ and $\hat{c}\hat{d}$ lies inside the triangle with vertices $\hat{a}\hat{p}$, $\hat{d}\hat{p}$, $\hat{d}\hat{a}$ (see Figure 6.8). Thus, $\text{Co}(\hat{c}\hat{d}, \hat{a}\hat{p})$ is in the interior of the region enclosed by $\text{Co}(\hat{d}\hat{p}, \hat{c}\hat{d}) \cup \text{Co}(\hat{c}\hat{d}, \hat{a}\hat{d}) \cup \text{Co}(\hat{a}\hat{d}, \hat{a}\hat{p}) \cup \text{Co}(\hat{a}\hat{p}, \hat{c}\hat{a}) \subseteq K_1 K_2$.

In both cases, we have $\text{Co}(\hat{c}\hat{d}, \hat{a}\hat{p}) \subseteq K_1 K_2$.

**Lemma 6.4.6.** Suppose $a < b \leq c < d$, $p = t_1(1 + ic) + t_2(1 + id)$ is nonzero for some $t_1, t_2 \geq 0$ and $K_1 = \text{Co}(1 + ia, 1 + ib)$, and $K_2 = \text{Co}(1 + ic, 1 + id, p)$. Assume also that there is no $\xi \in \mathbb{C}$ such that $\xi K_1 \subseteq K_2$. Then $K_1 K_2$ is star-shaped and $(1 + bi)(1 + ci)$ is a star center.

*Proof.* Let $\hat{a} = 1 + ia$, $\hat{c} = 1 + ic$, $\hat{d} = 1 + id$. Similar to the previous lemma, it is enough to show that $\text{Co}(\hat{b}\hat{c}, \hat{a}\hat{p}) \subseteq K_1 K_2$ for any $p = t_1\hat{c} + t_2\hat{d}$ such that $t_1, t_2 \geq 0$.

Let $\xi \in \mathbb{C}$ such that $\xi \text{Co}(\hat{c}, p)$ is a vertical line segment with real part 1. If $\xi \text{Co}(\hat{c}, p) \not\subseteq \text{Co}(\hat{a}, \hat{b})$, then by Corollary 6.2.6, $\hat{b}\hat{c}$ is a star-center of $K_1 \text{Co}(\hat{c}, p)$ and hence $\text{Co}(\hat{b}\hat{c}, \hat{a}\hat{p}) \subseteq K_1 K_2$. Otherwise, we have $\xi \text{Co}(\hat{c}, p) \subseteq \text{Co}(\hat{a}, \hat{b})$ and $K_1 \text{Co}(\hat{c}, p)$ is as shown in Figure 6.9c. This will only happen if $\text{Re}(p) < 1$. Since $\hat{a}\hat{p} = t_1(\hat{c}\hat{a}) + t_2\hat{d}\hat{a}$ for some $t_1, t_2 \geq 0$ such that $t_1 + t_2 < 1$, then $\hat{a}\hat{p} \in \text{Co}(0, \hat{c}\hat{a}, \hat{d}\hat{a})$ and $\hat{b}\hat{p} \in \text{Co}(0, \hat{c}\hat{b}, \hat{d}\hat{b})$. Note also that 0 and $p\hat{a}$ are separated by the line segment $\text{Co}(\hat{c}\hat{b}, \hat{c}\hat{a})$. Hence, $p\hat{a}$ is in the quadrilateral $K_1 \text{Co}(\hat{c}, \hat{d})$ and therefore $\text{Co}(\hat{a}\hat{p}, \hat{c}\hat{b}) \subseteq K_1 K_2$. This finishes the proof that $\hat{c}\hat{b}$ is a star center for $K_1 K_2$. \qed
**Proof of Theorem 6.4.3:** Note that (d) follows from (b) by considering $\bar{K}_1 \bar{K}_2$. Similarly, (e) follows from (a). Thus, we only need to prove (a)-(c).

To prove that $s$ is a star center of $K_1 K_2$, we show that for any $p \in K_2$, $s$ is a star center of $K_1 \text{Co}(\hat{c}, \hat{d}, p)$. To accomplish this, it is enough to show that $\text{Co}(s, \hat{u} \hat{v}) \subseteq K_1 K_2$ for all pairs $(u, v) \in \{\hat{b}, \hat{a}\} \times \{\hat{c}, \hat{d}, p\}$ by Theorem 6.3.3, where $p = t_1 \hat{c} + t_2 \hat{d}$ for some $t_1, t_2 \geq 0$.

For (a), the conclusion follows directly from Lemma 6.4.6.

To prove (c), the only nontrivial cases to consider are when $(u, v) = (\hat{a}, p)$ or $(u, v) = (\hat{b}, p)$. By Lemma 6.4.5, $\text{Co}(\hat{c} \hat{d}, \hat{a} p) \subseteq \text{Co}(\hat{a}, \hat{d}) \text{Co}(\hat{c}, \hat{d}, p) \subseteq K_1 K_2$. By Lemma 6.4.5 again, the product $\text{Co}(\hat{b}, \hat{c}) \text{Co}(\hat{c}, \hat{d}, p)$, has $\hat{c} \hat{d}$ as a star center. Thus, $\hat{c} \hat{d}$ is a star center of $\text{Co}(\hat{b}, \hat{c}) \text{Co}(\hat{c}, \hat{d}, p)$ and thus $\text{Co}(\hat{c} \hat{d}, \hat{b} p) \subseteq \text{Co}(\hat{b}, \hat{c}) \text{Co}(\hat{c}, \hat{d}, \hat{p}) \subseteq K_1 K_2$.

To prove (b), it is enough to show that $\text{Co}(\hat{c} \hat{b}, \hat{a} p) \subseteq K_1 K_2$ for all $p \in K_2$. We consider two cases,

1. Suppose $p = t_1 \hat{d} + t_2 \hat{b}$ for some $t_1, t_2 \geq 0$. Then by Lemma 6.4.6, $\hat{b} \hat{c}$ is a star-center of $\text{Co}(\hat{a}, \hat{c}) \text{Co}(\hat{b}, \hat{d}, p)$. Thus $\text{Co}(\hat{b} \hat{c}, \hat{a} p) \subseteq \text{Co}(\hat{a}, \hat{c}) \text{Co}(\hat{b}, \hat{d}, p) \subseteq K_1 K_2$.

2. Suppose $p = t_1 \hat{b} + t_2 \hat{c}$ for some $t_1, t_2 \geq 0$. Then by Lemma 6.4.5, $\hat{b} \hat{c}$ is a star-center of $\text{Co}(\hat{a}, \hat{b}) \text{Co}(\hat{b}, \hat{c}, p)$. Thus $\text{Co}(\hat{b} \hat{c}, \hat{a} p) \subseteq \text{Co}(\hat{a}, \hat{b}) \text{Co}(\hat{b}, \hat{c}, p) \subseteq K_1 K_2$. 

---

**FIG. 6.9**
In both cases, $\tilde{b}\hat{c}$ is a star-center for $K_1K_2$. 

It is clear that Theorem 6.4.1 follows from Proposition 6.4.2 and Theorem 6.4.3.

6.5 A circular disk and a closed set

It is known that the product of two circular disks is star-shaped [37, 38, 77, 81]. In this section, we will prove some unexpected results that if $K_1$ is a circular disk, then for many closed sets $K_2$, the product set is star-shaped. We will use $D(\mu, R)$ to denote the closed disk with center $\mu \in \mathbb{C}$ and radius $R \geq 0$.

Note that if $0 \in K_1$, then for every non-empty set $K_2$, $K_1K_2$ is star-shaped with 0 as star center. Suppose $0 \notin K_1$, we can always scale $K_1$ so that it is a circular disk centered at 1 with radius $r < 1$.

We have the following results showing that the product set of a circular disk and another set would be star-shaped under some very general conditions. We begin with the following observation.

Lemma 6.5.1. Suppose $r \in (0, 1]$ and $b \in D(1, r)$. Then the product $D(1, r)\{b\}$ is a disk containing $1 - r^2$.

Proof. Let $b \in D(1, r)$. Then $bD(1, r) = D(b, |b|r)$.

$$|b - (1 - r^2)|^2 = (b - (1 - r^2))(\overline{b} - (1 - r^2))$$

$$= |b|^2 - (b + \overline{b})(1 - r^2) + (1 - r^2)^2$$

$$= |b|^2 r^2 - (1 - r^2)(-|b|^2 + (b + \overline{b}) - (1 - r^2))$$

$$= |b|^2 r^2 - (1 - r^2)(r^2 - (b - 1)(\overline{b} - 1))$$

$$\leq |b|^2 r^2 \quad \text{because } |b - 1| \leq r \leq 1.$$ 

$\Box$
From the above simple proposition, we get the following.

**Theorem 6.5.2.** Suppose $K_1 = D(\mu, R)$ does not contain 0. For every nonempty subset $S$ of $K_1$, the product set $K_1 S$ is star shaped with star center $\mu^2(1 - r^2)$, where $r = |\mu^{-1} R|$. 

In the numerical range context, for every circular disk $K_1$, there is $A \in \mathbb{C}^{2 \times 2}$ such that $A - (\text{tr} A) I / 2$ is nilpotent and $W(A) = K_1$. Moreover, $B \in \mathbb{C}^{n \times n}$ satisfies $W(B) \subseteq W(A)$ if and only if $B$ admits a dilation of the form $I \otimes A$; see [1, 21]. By Theorem 6.5.2, if $A \in \mathbb{C}^{2 \times 2}$ such that $(A - \text{tr} A) I / 2$ is nilpotent, then $W(A) W(B)$ is star-shaped for any $B \in \mathbb{C}^{n \times n}$ satisfying $W(B) \subseteq W(A)$.

Next, we have the following.

**Theorem 6.5.3.** Suppose $r \in (0, 1]$ and $b \in \mathbb{C}$ with $\text{Re} (b) \geq 1$. Then the product $\text{Co}(1, b) D(1, r)$ is star-shaped with 1 as star center.

**Proof.** Suppose $b = \text{Re} i \theta$ with $R \geq 0$ and $-\pi / 2 \leq \theta \leq \pi / 2$. Let $c \in \text{Co}(1, b)$. Then $c = 1 + s \text{Re} i \theta$ for some $0 \leq s \leq 1$. $c K_1 = D(c, |c| r)$. Therefore, $\text{Co}(1, b) D(1, r) = \cup \{ D(c, |c| r) : c \in \text{Co}(1, b) \}$. Let $z \in \text{Co}(1, b) D(1, r)$. Then $|z - (1 + s \text{Re} i \theta)| \leq |1 + s \text{Re} i \theta| r$ for some $0 \leq s \leq 1$. Let $0 \leq t \leq 1$. We have

\[
|t z + (1 - t) - (1 + ts \text{Re} i \theta)|^2 \\
= |t(z - (1 + s \text{Re} i \theta))|^2 \\
\leq t^2 |1 + s \text{Re} i \theta|^2 r^2 \\
= ((1 + ts R \cos \theta)^2 + (ts R \sin \theta)^2) r^2 \\
= ((1 + ts R \cos \theta)^2 + (ts R \sin \theta)^2 - (1 - t)(1 + t + 2ts R \cos \theta)) r^2 \\
\leq ((1 + ts R \cos \theta)^2 + (ts R \sin \theta)^2) r^2 \\
= |1 + ts \text{Re} i \theta|^2 r^2.
\]
Therefore, $tz + (1 - t) \in D(1 + tsRe^{i\theta}, |1 + tsRe^{i\theta}|r) \subseteq Co(1, b)D(1, r)$. \hfill \Box

**Theorem 6.5.4.** Suppose $S$ is a star-shaped subset of $\mathbb{C}$ with star center $s$ such that $|s| \leq |z|$ for every $z \in S$. Then $D(a, r)S$ is star-shaped for every circular disk $D(a, r)$. In particular, if $S$ is convex, then $D(a, r)S$ is star-shaped for every circular disk $D(a, r)$.

**Proof.** If either $S$ or $D(a, r)$ contains 0, the result holds. So we may assume that $0 \notin S \cup D(a, r)$.

We may assume that $s = 1$ and $D(a, r) = D(1, r)$ with $0 \leq r \leq 1$. Then for every $z \in S$, $z = 1 + Re^{i\theta}$ for some $-\frac{\pi}{2} \leq \theta \leq \frac{\pi}{2}$. By Theorem 6.5.3, the product $Co(1, z)D(1, r)$ is star shaped with star center 1. Hence, $SD(1, r)$ is also star shaped with star center 1. \hfill \Box

Apart from the nice results above, there are some limitations about the star-shapedness of the product set of a circular disk and another set in $\mathbb{C}$ as shown in the following.

**Example 6.5.5.** Let $S = Co(1, 2e^{i\frac{11\pi}{12}}) \cup Co(1, 2e^{-i\frac{11\pi}{12}})$. Then $S$ is star-shaped with 1 as star center. Let $D(1, \frac{1}{2})$ be the disk centered at 1 with radius $\frac{1}{2}$. Then the product set $SD(1, \frac{1}{2})$ is not simply connected (see Figure 6.10.)

![Image of the product set](image-url)

**FIG. 6.10:** The product set $(Co(1, 2e^{i\frac{11\pi}{12}}) \cup Co(1, 2e^{-i\frac{11\pi}{12}})) \cdot D(1, \frac{1}{2})$ is not simply connected.
6.6 Additional results and further research

We have to assume compactness in most of our results. One may wonder what happen if we relax this assumption. The following example shows that without the end points, the product of two line segments may not be star-shaped.

**Example 6.6.1.** Let \( K_1 = K_2 \) be the line segment joining \( 1+i \) and \( 1-i \) without the end points. Then \( K_1K_2 \) has no star center.

**Verification.** Note that the closure of \( K_1K_2 \) equals \( S = \text{Co}(1+i, 1-i)\text{Co}(1+i, 1-i) \) has a unique star-center 2. The set \( K_1K_2 \) is obtained from \( S \) by removing the line segments \( \text{Co}(2, 2i) \) and \( \text{Co}(2, -2i) \). The only point in the closure can reach all the points in \( K_1K_2 \) is 2, but it is not in \( K_1K_2 \). So, \( K_1K_2 \) is not star-shaped.

Recall that an extreme point of a compact convex set \( S \subseteq \mathbb{C} \) is an element in \( S \) that cannot be written as the mid-point of two different elements in \( S \). If \( S \) is a polygon (with interior) then its vertexes are the extreme points. We can extend Theorem 6.3.3 to the following.

**Theorem 6.6.2.** Let \( K_1, K_2 \subseteq \mathbb{C} \) be compact convex sets. Then \( K_1K_2 \) is star-shaped if and only if there is \( p \in K_1K_2 \) such that \( \text{Co}(p, ab) \subseteq K_1K_2 \) for any extreme points \( a \in K_1 \) and \( b \in K_2 \).

**Proof.** If \( K_1K_2 \) is star-shaped, then a star center \( p \in K_1K_2 \) satisfies \( \text{Co}(p, c) \subseteq K_1K_2 \) for any \( c \in K_1K_2 \). Now, suppose there is \( p \in K_1K_2 \) satisfying \( \text{Co}(p, ab) \subseteq K_1K_2 \) for any extreme points \( a \in K_1 \) and \( b \in K_2 \). Let \( \mu = \mu_1\mu_2 \) with \( \mu_1 \in K_1, \mu_2 \in K_2 \). By the Carathéodory Theorem \( \mu_1 \in \text{Co}(a_1, a_2, a_3) \) and \( \mu_2 \in \text{Co}(b_1, b_2, b_3) \) for some extreme points \( a_1, a_2, a_3 \in K_1 \) and \( b_1, b_2, b_3 \in K_2 \). (Some of the \( a_i \)'s may be the same, and also some of the \( b_i \)'s may be the same.) Suppose \( p = p_1p_2 \) with \( p_1 \in K_1 \) and \( p_2 \in K_2 \). Then \( p_1 \in \text{Co}(a_4, a_5, a_6) \) and \( p_2 \in \text{Co}(b_4, b_5, b_6) \) for some extreme points \( a_4, a_5, a_6 \in K_1 \) and
\(b_4, b_5, b_6 \in K_2\). By Theorem 6.3.3, \(\text{Co}(p, \mu_1 \mu_2) \subseteq \text{Co}(a_1, \ldots, a_6)\text{Co}(b_1, \ldots, b_6) \subseteq K_1K_2\).

Thus, \(p\) is a star center of \(K_1K_2\). 

Another observation is the following extension of Proposition 6.1.1(b). Note that we do not need to impose compactness conditions on \(K_1\) or \(K_2\).

**Proposition 6.6.3.** Suppose \(K_1 \subseteq \mathbb{C}\) is star-shaped with 0 as a star center. Then for any non-empty subset \(K_2 \subseteq \mathbb{C}\), the set \(K_1K_2\) is star-shaped with 0 as a star center.

**Proof.** Let \(p = p_1p_2 \in K_1K_2\) with \(p_1 \in K_1, p_2 \in K_2\). Then \(\text{Co}(0, p) = \text{Co}(0, p_1)\{p_2\} \subseteq K_1K_2\). 

There are other interesting questions deserve further research. We mention a few of them in the following.

**P1** Find necessary and sufficient conditions on \(K_1\) and \(K_2\) so that \(K_1K_2\) is convex or star-shaped.

In the context of numerical range if \(A \in \mathbb{C}^{2 \times 2}\), then \(W(A)\) is an elliptical disk. So, it is also of interest to study the following.

**P2** Let \(K_1, K_2\) be two elliptical disks. Determine conditions on \(K_1, K_2\) so that \(K_1K_2\) is star-shaped or convex.

One may also consider the following.

**P3** Characterize those elliptical disks \(K_1\) such that \(K_1K_2\) is star-shaped for all compact convex set \(K_2\).

More generally, one may consider the following.

**P4** Characterize those compact convex sets \(K_1\) such that \(K_1K_2\) is convex or star-shaped for any compact convex set \(K_2\).
In connection to Problem P4, we have shown that if $K_1$ is a close line segment or a close circular disk, then $K_1 K_2$ is star-shaped for any compact convex set $K_2$. These results are also connected to Problem P3 because a line segment and a circular disk can be viewed as elliptical disks.

It is also interesting to study the Minkowski product of $s$ (convex) sets $K_1, \ldots, K_s$. The study will be more challenging. As pointed out in [81], the set $K_1 \cdots K_s$ may not be simply connected in general. Nevertheless, our results in Section 6.5 and Proposition 6.6.2 imply the following.

**Proposition 6.6.4.** Suppose $K_1, \ldots, K_s \subseteq \mathbb{C}$.

1. If any one of the sets $K_1, \ldots, K_s$ is star-shaped with 0 as a star center, then $K_1 \cdots K_s$ is star-shaped with 0 as a star center.

2. Suppose there is a nonzero number $\mu_1$ such that $\mu_1 K_1$ is a circular disk center at 1 with radius $r < 1$.

   (2.a) If there is $\mu \in \mathbb{C}$ such that $\mu K_2 \cdots K_s \subseteq \mu_1 K_1$, then $K_1 \cdots K_r$ is star-shaped with $(\mu_1 \mu)^{-1}(1 - r^2)$ as a star center.

   (2.b) If there is $\mu \in \mathbb{C}$ such that $\mu K_2 \cdots K_s \subseteq \{z \in \mathbb{C} : \text{Re}(z) \geq 1\}$, then $K_1 \cdots K_r$ is star-shaped with $(\mu_1 \mu)^{-1}$ as a star center.

It is also interesting to study the following problem.

**P5** Characterize those compact (convex) sets $K$ such that $K^2$ is convex or star-shaped.
CHAPTER 7

Summary and Concluding Remarks

The problems presented in this dissertation have been in keeping with the spirit of the main goals of the field of quantum information. In [78], M.A. Neilsen stated that quantum information theory is concerned with (a) determining the theoretical extent and limitations in carrying out information processing tasks using quantum mechanical laws and; (b) to provide constructive means for achieving these tasks.

The first problem provided an algorithmic way to efficiently break down a general $n$-qubit quantum operation, on a closed system, into simpler operations with associated costs. This result is more in line with (b). It remains an open question if the decomposition scheme presented in chapter 2 is optimal in a sense that for any other decomposition scheme, there exists a general $n$-qubit quantum operation for which the cost of applying our scheme is less than the other scheme. This challenging problem is more aligned with (a).

In the second problem, we found theoretical bounds on certain classes of functions on two known quantum states, where one is presumed to go through an unknown quantum channel that belongs to a specific set of quantum operations. The theoretical results
in chapter 3 help give insight to the limitations of information that can be harnessed after quantum mechanical processes have taken place. As a matter of fact, obtaining the theoretical bounds was a consequence of determining the instances in which these bounds are attained. It is the authors’ hope that this knowledge can potentially improve the design of some experiments. Or perhaps help identify a working quantum computer.

In chapter 4, we addressed a very specific problem involving bipartite qubit-qudit states with maximally mixed qudit reduced states and found some interesting general observations. A general answer to the problem presented has been evasive but answers for relatively small matrix dimensions have been found. It would be delightful to find simple general patterns for the list of necessary and sufficient conditions for something to be an element of $E_n$. But keep in mind that it has often been the case that the dimensions of systems considered in experimental quantum physics are relatively small.

In chapter 6, we considered the shape of Minkowski products of convex sets. From a purely mathematical perspective, this problem is challenging and exciting. The problem itself and the definitions necessary to define it are easy to understand. But it requires some ingenuity in proving the results. In the context of quantum information theory, these results are important to describe the product numerical range of a product state, which in turn have been used in the study of positive maps, minimum output entropy of a channel, local discrimination of unitary operators and quantum error-correction among other things [45].

Several theoretical results for some basic problems have been presented but do not necessarily give constructive means to utilize these results [67]. In chapter 5, we used numerical methods to aid with this. With the help of technology and powerful computers, one hopefully gets a better intuition about these problems and ultimately find a solution.

Quantum information theory is a very active area of research and there is a vast array of research topics in the field. We have touched on some of them in this dissertation. For
some problems that we have solved, new and more challenging questions arise and some solutions and techniques have also sparked questions that have not been considered before. There have been stumbling blocks in completely solving some of them and we will continue to look in different directions to find the right tool we need until we reach the limit.
APPENDIX A
Matlab Scripts

A.1 Implementation of Partial trace Maps

The Matlab script `bitriPT.m` computes the reduced state bipartite or a tripartite system whose global state is $A$. The vector $w$ contains the dimensions of the constituent systems and `pos` (takes either 'first' or 'mid' or 'last') indicates the position of the system(s) to be traced out. For example, if $A$ is an $18 \times 18$ density matrix, the command `bitriPT('mid',[2,3,3],A)` produces a $6 \times 6$ density matrix which is $\text{tr}_2(A)$.

```matlab
function B1=bitriPT(pos,w,A)
    if strcmp(pos,'first');
        m=w(2); B1=zeros(m);
        for ii=1:w(1)
            B1=B1+A(1+(ii-1)*m:m*ii,1+(ii-1)*m:m*ii);
        end
    elseif strcmp(pos,'last');
        m=w(1); n=w(2); mn=m*n; B1=zeros(m);
        for ii=1:n
            B1=B1+A(ii:n:mn,ii:n:mn);
        end
    else strcmp(pos,'mid');
        B1=zeros(w(1)*w(3));
        for ii=1:w(2)
            r1=ones(1,w(1));
            r3=ones(1,w(3));
            s=(ii-1)*w(3)+1:ii*w(3);
            t=0:w(2)*w(3):(w(1)-1)*w(2)*w(3);
            inds=kron(r1,s)+kron(t,r3);
            B1=B1+A(inds,inds);
        end
    end
end
```
The Matlab script `parttrace.m` computes any reduced state $\text{tr}_J(A)$ of any $k-$partite quantum system with global state $A$. The vector $v_1$ contains the sizes of the subsystems of $A$ and $v_2$ is a binary vector. The zeros in $v_2$ indicate the systems to be traced out. For example, if $A$ is a $48 \times 48$ density matrix, then `parttrace([3,2,2,4],[1,0,1,0],A)` produces $\text{tr}_{24}(A)$, which is a $6 \times 6$ density matrix.

```matlab
function B=parttrace(v1,v2,A)
    v1=v1(:)'; v2=v2(:)'; k=size(v2,2);
    while size(v1(v2==0),2)>0
        i0=0; j0=1;
        while (i0<k)&(v2(i0+1)==1)
            i0=i0+1; j0=j0*v1(i0);
        end
        i1=i0; j1=1;
        while (i1<k)&(v2(i1+1)==0)
            i1=i1+1; j1=j1*v1(i1);
        end
        i2=i1; j2=1;
        while i2<k
            i2=i2+1; j2=j2*v1(i2);
        end
        if j2>1
            if j1>1
                pos='mid';
                w=[j0,j1,j2];
            else
                pos='first';
                w=[j0,j1];
            end
        else
            pos='last';
            w=[j0,j1];
        end
        A=bitriPT(pos,w,A);
        v1(i0+1:i1)=[];
        v2(i0+1:i1)=[];
        k=k-i1+i0;
    end
    B=A;
end
```
A.2 Unitary Gate Decomposition

The following Matlab script implements the decomposition scheme for a unitary matrix \( U \in U_{2^n} \) into a product of controlled gates described in chapter 2. The input \( U \) is optional and will be assigned randomly if not provided by the user. Output \((x,y)\) will display the order in which the entries are to be annihilated, while \( A \) displays the representation of the controlled gates \( c_n c_{n-1} \cdots c_1 \in \{0,1,*V\}^n \) used. The matrix \( W_s \in U_2 \) used by the \( j^{th} \) controlled gate is given by the \( s^{th} \) row of \( V \). That is, \( W_s = \begin{bmatrix} V(s,1) & V(s,2) \\ V(s,3) & V(s,4) \end{bmatrix} \). The outputs \( \text{num} \) and \( \text{controls} \) are positive numbers that count the number of nontrivial gates (i.e., not equal to \( I_{2^n} \)), and the total number of controls used in the decomposition.

```matlab
function \([A,x,y,controls,\text{num},V]=\text{decomposition}(n,U)\)

\([x,y,A]=\text{schemetable}(n); \%\text{see subroutine below}\)

if \!exist('U',\'var\') \%IF U IS NOT SPECIFIED
    U=\text{randomunitary}(n); \%\text{see subroutine below}
end

N=2^n;
d=N*(N-1)/2;
V=zeros(d,4);
Y=U;
controls=0;
num=d;

\%COMPUTES \text{num} AND \text{controls} AND GENERATES V
for j=1:d
    \[D,K,c]=\text{ithgate}(A(j,:),x(j,1),y(j,1),Y,n); \%\text{see subroutine below}\n    V(j,:)=K;
    if V(j,:)==[1,0,0,1]
        num=num-1;
    else
        Y=D*Y;
        controls=controls+c;
    end
end

\%FUNCTION \text{randomunitary}(n)
function \[W]=\text{randomunitary}(n)\n\%\text{THIS FUNCTION GENERATES A RANDOM 2}^n \text{ by 2}^n \text{ UNITARY MATRIX}
W=rand(2^n)+1i*rand(2^n);
H=0.5*(W+W');
W=expm(1i*H);
```

187
function [x, y, A] = schemetable(n)
% THIS FUNCTION GENERATES THE SCHEME TABLE FOR n
N = 2^n; % dimension of matrix
d = N*(N-1)/2; % number of gates used
% ORDER OF ANNIHILATION, COLUMN INDICES
y = zeros(d, 1);
temp1 = 0;
for j = 1:N-1;
y(temp1+1:temp1+N-j, 1) = j;
temp1 = temp1 + 2^n - j;
end
% ORDER OF ANNIHILATION, ROW INDICES AND GATE-TYPES
x = zeros(d, 1);
A = repmat('T', d, n);
x(1, 1) = 2;
A(1, n) = 'T';
for k = 2:n % loop index signifies leading 2^k by 2^k subblock
    temp2 = 2^(k-1);
    % COLUMN 1 (ENTRIES AND GATES)
x(temp2:2*temp2-1, 1) = [x(1:temp2-1, 1) + temp2; temp2 + 1]; % ROWS 2^(k-1)+1--2^k
    A(temp2:2*temp2-1, n-k+1:n) = [A(1:temp2-1, n-k+1:n);...
        repmat('T', 1, k-1)];
    for i = 1:k-1
        A(temp2 + 2^i - 2, n-k+1) = '1'; % 1G, where G gate that annihilates 2^i+1
    end
    % COLUMNS 2-2^k
    temp3 = 2^n-1; % temp3 counts number of columns left
    for j = 2:temp2 % FOR LOWER LEFT OF 2^k subblock
        x(temp3+temp2-j+1:temp3+2*temp2-j+1, 1) = Fell(k, j, x(temp2:2*temp2-1, 1)); % see subroutine below
    end
    A(temp3+temp2-j+1:temp3+2*temp2-j, n-k+1:n) = ...% see subroutine below
        Fell(A(temp2:2*temp2-1, n-k+1:n), j);
    temp3 = temp3 + N-j;
end
for jj = 1:temp2-1 % FOR UPPER LEFT/LOWER RIGHT OF 2^k SUBBLOCK
    bb = (jj-1)*(N-jj/2);
    x(temp3+1:temp3+temp2-jj, 1) = x(bb+1:bb+temp2-jj, 1) + temp2;
    A(temp3+1:temp3+temp2-jj, n-k+1:n) = ...% see subroutine below
        repmat('1', temp2-jj, 1), A(bb+1:bb+temp2-jj, n-k+2:n);
    temp3 = temp3 + N-temp2-jj;
end
function [D,K,c]=ithgate(Ai,xi,yi,U,n)
%This function generates the controlled gate D of the form described in
%Ai that annihilates the (xi,yi) entry of the unitary matrix U of size 2^n
c=0;
D=1;
if yi==2^n-1
   D=U';
   K=[conj(U(2^n-1,2^n-1));conj(U(2^n,2^n-1))];
   c=c+n-1;
else
   for k=n:-1:1
      if isequal(Ai(k),'0')==1
         D=kron([1,0;0,0],D);
         c=c+1;
      elseif isequal(Ai(k),'1')==1
         D=kron([0,0;0,1],D);
         c=c+1;
      elseif isequal(Ai(k),'T')==1 && U(xi,yi)==0
         K=[1,0,0,1];
         D=kron(zeros(2,2),D);
      elseif isequal(Ai(k),'T')==1 && (bitget(yi-1,n-k+1)==0)
         a=U(Fell(n,2^(n-k)+1,xi),yi);
         b=U(xi,yi);
         z=sqrt(abs(a)^2+abs(b)^2);
         K=(1/z)*[a,-b,conj(b),conj(a)];
         D=kron(-1*eye(2,2)+[K(1,1:2);K(1,3:4)],D);
      elseif isequal(Ai(k),'T')==1 && (bitget(yi-1,n-k+1)==1)
         a=U(Fell(n,2^(n-k)+1,xi),yi);
         b=U(xi,yi);
         z=sqrt(abs(a)^2+abs(b)^2);
         K=(1/z)*[conj(a),conj(b),-b,a];
         D=kron(-1*eye(2,2)+[K(1,1:2);K(1,3:4)],D);
      else
         D=kron(eye(2,2),D);
      end
   end
   D=D+eye(2^n,2^n);
end

function [Y]=Gell(X,l)
%THIS IS THE FUNCTION G_l IN PROCEDURE 2.1
%l is an integer from 1 to 2^k-1; X must be p by k; Y=G_l(X)
Y=X;
[p,k]=size(X);
C=repmat('1',1,k);
s=dec2bin(l-1);
r=size(s,2);
Y(p,1)=’T’;
for m=1:r
    if bitget(l-1,m)==1
        for t=1:p-1
            if X(t,k-m+1)==’1’
                Y(t,k-m+1)=’0’;
            end
        end
        Y(p,k-m+1)=’1’;
    end
end
for t=1:p-1
    if size(intersect(X(t,1:k-r),C),2)==0
        Y(t,1)=’1’;
    end
end

function [v] = Fell(n,r,u)
%Fell takes a vector of integers u and send it to the vector of integers v, 
%the binary representation of u(i,j) and v(i,j) differ precisely in places 
%where the binary digit of r (in a word of length n) is 1
ub=dec2bin(u(:,1)-1,n);
rv=r*ones(size(u));
rbin=dec2bin(rv(:,1)-1,n);
flip=mod(ub+rb,2);
fbits=cellstr(num2str(flip));
v=bin2dec(fbits(:,1))+1;

The Matlab script gatecount.m was used to generate Figure fig:costcomp. Given n, it plots
the difference $T_2(k) - T_1(k)$ for $k = 1, \ldots, n$. The output $w$ is a $2 \times n$ array wuch that the $(j,k)$
entry is $T_j(k)$.

function [w]=gatecount(n)
G=zeros(n,n); %no. of r-qbit gates w/ k-1 controls (Pelejo, Li)
H=zeros(n,n); %no. of r-qbit gates with k-1 controls (Vartiainen et al)
W=zeros(n,n); %weight matrix column k of G*W is column k of G times (k-1)

w=zeros(2,n);
for r=1:n
W(r,r)=r-1;
G(r,1)=r;
H(r,1)=2^-(r-1);
if r>1
G(r,2)=r*(r-1)*(2^-(r-2)+1);
for k=2:r
H(r,k)=H(r-1,k)+H(r-1,k-1)+ max([2^-(r-2),2^-(k-1)])+ 2^-(2*r-k-1)-2^-(r-2)
end
end
if r>2
G(r,3)=(1/3)*(4^r-4)-(2^r)*(r-1)+r*(r-1)*(r-2)/2;
end
if r>3
for k=4:r
G(r,k)=G(r-1,k)+G(r-1,k-1)+ nchoosek(r-1,k-1);
end
end
V=G*W;
X=H*W;
w(1,r)=sum(V(r,:));
w(2,r)=sum(X(r,:));
end

x=1:n;
plot(x,w(2,:)-w(1,:),'k','LineWidth', 2);
ylabel('log(T2(n)-T1(n)) base 10');
xlabel('n');

A.3 Optimal Values of $F(\rho_1, \Phi(\rho_2))$ and $H(\rho_1||\Phi(\rho_2))$

The Matlab script `maxFidvN` carries out the steps in Algorithm 3.3.4 to generate the matrix C such that the fidelity C is majorized by B (i.e. there is a mixed unitary/unital channel sending B to C) and the fidelity $F(A, C)$ is maximum and $H(A||C)$ is minimum. It also outputs $fmin, fmax$, which are the minimum and maximum values, respectively, of $F(A, \Phi(B))$ over all mixed unitary (or over all mixed unital) channels. Similarly, $rvnmin, rvnmax$, are the minimum and maximum values of the quantum relative entropy $H(A||\Phi(B))$ over all mixed unitary (or over all mixed unital) channels. The subroutines Fid and RvN computes the fidelity and the quantum relative entropy of two density matrices. Another subroutine `ismajorized` returns 1 if $x/\sum(x)$ is majorized by $y/\sum(y)$ and 0 otherwise.

```matlab
function [C,fmin,fmax,rvnmin,rvnmax]=maxFidvN(A,B,n)
A=(A+A')/2;
A=A/trace(A);
B=(B+B')/2;
```
B = B/trace(A);
[Us,Ds] = eig(A);
a = diag(Ds);
[as,Is] = sort(a, 'descend');
Us = Us(:, Is);
b = eig(B);
b = sort(b, 'descend');
c = zeros(n, 1);
if min(a) < 0 | min(b) < 0
    C = zeros(n);
    fprintf('ERROR: Your A and B are not positive semidefinite');
else
    indf = 1;
    while indf <= n
        if indf == n
            c(n) = b(n);
        elseif a(indf) == 0
            c(indf:n) = b(indf:n);
            indf = n + 1;
        else
            indl = indf;
            while indl < n
                if ismajorized(a(indf:indl+1), b(indf:indl+1)) == 1
                    indl = indl + 1;
                else
                    break;
                end
            end
            sa = sum(a(indf:indl));
            sb = sum(b(indf:indl));
            c(indf:indl) = sb*a(indf:indl)/sa;
            indf = indl + 1;
        end
    end
    bdwn = sort(b, 'ascend');
    C = Us*diag(c)*Us';
    fmin = Fid(diag(a), diag(bdwn));
    fmax = Fid(A, C);
    rvnmin = RvN(A, C);
    rvnmax = RvN(diag(a), diag(bdwn));
end

function l = ismajorized(x, y)
x = x/sum(x);
y = y/sum(y);
\begin{verbatim}
x=sort(x,'descend');
y=sort(y,'descend');
l=1;
k=1;
n=size(x,1);
while l==1 & k<n
    if (sum(x(1:k)))>(sum(y(1:k)))
        l=0;
        break;
    else
        k=k+1;
    end
end
function f=Fid(X,Y)
    sqX=X^(0.5);
    sqXY=(sqX*Y*sqX)^(0.5);
    f=trace(sqXY);
function g=RvN(V,W)
    temp = [V, W];
    if rank(temp)>rank(W) \%means that Col(V) is not a subset of Col(W)
        g = inf;
    else \%other case when supp(V) is contained in supp(W)
        [U1,D1] = eig(V);
        [U2,D2] = eig(W);
        L1 = D1;
        L2 = D2;
        for ii=1:size(L1,2)
            if D1(ii,ii)>0 \%we take log 0 to be 0
                L1(ii,ii)=log(D1(ii,ii));
            end
            if D2(ii,ii)>0 \%we take log 0 to be 0
                L2(ii,ii)=log(D2(ii,ii));
            end
        end
        g = trace(V*(U1*L1*U1'-U2*L2*U2'));
    end
end
\end{verbatim}

A.4 On Finding Extreme Points of $E^5$

The following script, named n5EXT.m, was used to generate the extreme points of $E^5$. 

193
The following script can be used to find a $\rho \in S_2(\frac{1}{5}I_5)$ that is permutationally similar to a direct sum of $2 \times 2$ matrices and whose eigenvalue is given by one of the extreme points listed above.

```matlab
function [J,feassimp,rho]=Findnicesol(a)
a(a<0)=0;
a=a/sum(a);
a=a(:);
a=sort(a,'descend');
a=sort(a,'descend');
a=sort(a,'descend');
a=sort(a,'descend');
```
I=[1,10,2,6,4,9,3,8,7,5]’;
fessimp=0;
N=factorial(10);
j=0;
err=10^(-15);
while j<N && fessimp==0
    j=j+1;
    J=nthperm(I,j);
    l1=logical((a(J(1))+a(J(10)))>= 0.2-err);
    l2=logical((a(J(2))+a(J(10)))<= 0.2+err);
    l3=logical((a(J(1))+a(J(10))+a(J(2))+a(J(3)))>= 0.4-err);
    l4=logical((a(J(1))+a(J(10))+a(J(2))+a(J(4)))<= 0.4+err);
    l5=logical((a(J(5))+a(J(7))+a(J(8))+a(J(9))) >= 0.4-err);
    l6=logical((a(J(6))+a(J(7))+a(J(8))+a(J(9))) <= 0.4+err);
    l7=logical((a(J(7))+a(J(9)))>= 0.2-err);
    l8=logical((a(J(8))+a(J(9)))<= 0.2+err);
    fessimp=l1*l2*l3*l4*l5*l6*l7*l8;
end

if feassimp==0
    J=zeros(size(I));
    fprintf('EIGS %g %g %g %g %g %g %g %g %g %g nosimplesol \n’, ... 
a(1),a(2),a(3),a(4),a(5),a(6),a(7),a(8),a(9), a(10));
else
    aj=zeros(1,10);
    for i=1:10
        aj(J(i))=a(i);
    end
    x1=1/5-aj(10);
    x2=2/5-aj(10)-aj(1)-aj(2);
    x3=3/5-aj(10)-aj(1)-aj(2)-aj(3)-aj(4);
    x4=4/5-aj(10)-aj(1)-aj(2)-aj(3)-aj(4)-aj(5)-aj(6);
    x5=aj(9);
    y1=sqrt(x1*(1/5-x2)-aj(1)*aj(2));
    y2=sqrt(x2*(1/5-x3)-aj(3)*aj(4));
    y3=sqrt(x3*(1/5-x4)-aj(5)*aj(6));
    y4=sqrt(x4*(1/5-x5)-aj(7)*aj(8));
    D=diag([x1,x2,x3,x4,x5]);
    X= [zeros(4,1),diag([y1,y2,y3,y4]);zeros(1,5)];
    rho=[D, X; X’, 1/5-D];
end
APPENDIX B

Extreme Points of $\mathcal{E}_5$ and $\mathcal{E}_6$

B.1 Extreme points of $\mathcal{E}_5$

Here is the list of extreme points of $\mathcal{E}_5$.

$$
\left(\frac{2}{5}, \frac{1}{5}, 0, \ldots, 0\right), \left(\frac{2}{3}, \frac{2}{3}, 0, \ldots, 0\right), \left(\frac{2}{3}, \frac{3}{10}, 0, \ldots, 0\right), \left(\frac{4}{5}, \frac{2}{5}, 0, \ldots, 0\right), \left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, 0, \ldots, 0\right), \\
\left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, 0, \ldots, 0\right), \left(\frac{3}{5}, \frac{1}{5}, 0, \ldots, 0\right), \left(\frac{2}{5}, \frac{3}{10}, 0, \ldots, 0\right), \left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, 0, \ldots, 0\right), \\
\left(\frac{3}{5}, \frac{1}{5}, 0, \ldots, 0\right), \left(\frac{2}{5}, \frac{3}{10}, 0, \ldots, 0\right), \left(\frac{3}{5}, \frac{1}{5}, 0, \ldots, 0\right), \left(\frac{2}{5}, \frac{3}{10}, 0, \ldots, 0\right), \left(\frac{3}{5}, \frac{1}{5}, 0, \ldots, 0\right), \\
\left(\frac{2}{5}, \frac{3}{10}, 0, \ldots, 0\right), \left(\frac{3}{5}, \frac{1}{5}, 0, \ldots, 0\right), \left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, 0, \ldots, 0\right), \left(\frac{3}{5}, \frac{1}{5}, 0, \ldots, 0\right), \\
\left(\frac{2}{5}, \frac{3}{10}, 0, \ldots, 0\right), \left(\frac{3}{5}, \frac{1}{5}, 0, \ldots, 0\right), \left(\frac{2}{5}, \frac{3}{10}, 0, \ldots, 0\right), \left(\frac{3}{5}, \frac{1}{5}, 0, \ldots, 0\right).
$$

196
B.2 Extreme points of $\mathcal{E}_6$

The following are the extreme points of $\mathcal{E}_6$.

\[
\left(\frac{1}{3}, \frac{1}{8}, 0, \ldots, 0\right), \left(\frac{1}{3}, \frac{1}{8}, \frac{1}{4}, 0, \ldots, 0\right), \left(\frac{1}{3}, \frac{1}{8}, \frac{1}{6}, 0, \ldots, 0\right), \left(\frac{2}{3}, \frac{3}{8}, \frac{2}{6}, 0, \ldots, 0\right), \\
\left(\frac{1}{3}, \frac{1}{8}, \frac{1}{6}, 0, \ldots, 0, \frac{1}{3}\right), \left(\frac{1}{3}, \frac{1}{8}, \frac{1}{6}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, 0, \ldots, 0\right), \\
\left(\frac{1}{3}, \frac{1}{8}, \frac{1}{6}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, 0, \ldots, 0\right), \\
\left(\frac{1}{3}, \frac{1}{8}, \frac{1}{6}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, 0, \ldots, 0\right), \\
\left(\frac{1}{3}, \frac{1}{8}, \frac{1}{6}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, 0, \ldots, 0\right), \\
\left(\frac{1}{3}, \frac{1}{8}, \frac{1}{6}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, 0, \ldots, 0\right), \\
\left(\frac{1}{3}, \frac{1}{8}, \frac{1}{6}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, 0, \ldots, 0\right), \\
\left(\frac{1}{3}, \frac{1}{8}, \frac{1}{6}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, 0, \ldots, 0\right), \\
\left(\frac{1}{3}, \frac{1}{8}, \frac{1}{6}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, 0, \ldots, 0\right), \\
\left(\frac{1}{3}, \frac{1}{8}, \frac{1}{6}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, 0, \ldots, 0\right), \\
\left(\frac{1}{3}, \frac{1}{8}, \frac{1}{6}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, 0, \ldots, 0\right), \\
\left(\frac{1}{3}, \frac{1}{8}, \frac{1}{6}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, 0, \ldots, 0\right), \\
\left(\frac{1}{3}, \frac{1}{8}, \frac{1}{6}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, 0, \ldots, 0\right), \\
\left(\frac{1}{3}, \frac{1}{8}, \frac{1}{6}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, 0, \ldots, 0\right), \\
\left(\frac{1}{3}, \frac{1}{8}, \frac{1}{6}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, 0, \ldots, 0\right), \\
\left(\frac{1}{3}, \frac{1}{8}, \frac{1}{6}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, 0, \ldots, 0\right), \\
\left(\frac{1}{3}, \frac{1}{8}, \frac{1}{6}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, 0, \ldots, 0\right), \\
\left(\frac{1}{3}, \frac{1}{8}, \frac{1}{6}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, 0, \ldots, 0\right), \\
\left(\frac{1}{3}, \frac{1}{8}, \frac{1}{6}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, 0, \ldots, 0\right), \\
\left(\frac{1}{3}, \frac{1}{8}, \frac{1}{6}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, 0, \ldots, 0\right), \\
\left(\frac{1}{3}, \frac{1}{8}, \frac{1}{6}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, 0, \ldots, 0\right), \\
\left(\frac{1}{3}, \frac{1}{8}, \frac{1}{6}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, 0, \ldots, 0\right), \\
\left(\frac{1}{3}, \frac{1}{8}, \frac{1}{6}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, 0, \ldots, 0\right), \\
\left(\frac{1}{3}, \frac{1}{8}, \frac{1}{6}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, 0, \ldots, 0\right), \\
\left(\frac{1}{3}, \frac{1}{8}, \frac{1}{6}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, 0, \ldots, 0\right), \\
\left(\frac{1}{3}, \frac{1}{8}, \frac{1}{6}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, 0, \ldots, 0\right), \\
\left(\frac{1}{3}, \frac{1}{8}, \frac{1}{6}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, 0, \ldots, 0\right), \\
\left(\frac{1}{3}, \frac{1}{8}, \frac{1}{6}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, 0, \ldots, 0\right). \\
\right.

197
APPENDIX C

Proof of Theorem 5.3.5

Note that the condition $\text{tr}_{J_s}(\rho) = \rho_{J_s}$ can be written as a set of linear constraint of the form $A_s x = b_s$ by vectorizing $\rho$ into $x \in \mathbb{R}^n$ and $\rho_{J_s}$ into $b_s \in \mathbb{R}^m$. Thus, Theorem 5.3.5 will follow from proposition 5.3.4 and the following theorem.

**Theorem C.0.1.** Let $A_j \in M_{n_j,N}$ and $b_i \in M_{n_j}$ for $j = 1, \ldots, m$. For any $\{j_1, \ldots, j_r\} \subseteq \{1, \ldots, m\}$, denote by $A_{\{j_1, \ldots, j_r\}}$ the matrix whose row space is $\bigcap_{s=1}^{r} \text{Row}(A_j)$. The set

$$L = \{x \mid A_s x = b_s \text{ for } s = 1, \ldots, m\}$$

is nonempty if and only if for any subset $\{j_1, \ldots, j_r\}$ of $\{1, \ldots, m\}$, the projection of $b_{j_s}$ onto $\bigcap_{\ell=1}^{r} \text{Row}(A_{j_\ell})$ is constant for all $s = 1, \ldots, r$. In this case, denote this projection by $b_{\{j_1, \ldots, j_r\}}$. Then the least square projection of $z \in \mathbb{C}^N$ onto $L$ is given by

$$\tilde{z} = z + \sum_{s=1}^{m} (-1)^r \sum_{\{j_1, \ldots, j_r\} \subseteq \{1, \ldots, m\}} A_{\{j_1, \ldots, j_r\}}^T \left(A_{\{j_1, \ldots, j_r\}} x - b_{\{j_1, \ldots, j_r\}} \right)$$

**Proof:**

We will prove this theorem by induction.

First, we consider the case when $m = 2$. Let $V = (V_1^T \ V_2^T \ V_3^T)^T$ such that the rows of $V_1$ form an orthonormal basis for $\text{Row}(A_1) \cap \text{Row}(A_2)^\perp$, the rows of $V_2$ form an orthonormal basis for $\text{Row}(A_1) \cap \text{Row}(A_2)$ and the rows of $V_3$ form an orthonormal basis for $\text{Row}(A_2) \cap \text{Row}(A_1)^\perp$. Then for some unitary $U_1 = \begin{pmatrix} U_{11} \\ U_{21} \end{pmatrix} \in M_{n_1}$ and $U_2 = \begin{pmatrix} U_{12} \\ U_{22} \end{pmatrix} \in M_{n_2}$, we have

$$\begin{pmatrix} A_1 \\ A_2 \end{pmatrix} = (U_1^* \oplus U_2^*) \begin{pmatrix} C_1 & 0 & 0 \\ 0 & C_2 & 0 \\ 0 & 0 & C_3 \end{pmatrix} V$$

198
Thus
\[
\begin{pmatrix}
A_1 \\
A_2
\end{pmatrix}^+ = V^* \begin{pmatrix}
C_1^\dagger & 0 & 0 & 0 \\
0 & C_2^\dagger & C_2^\dagger & 0 \\
0 & 0 & 0 & C_3^\dagger
\end{pmatrix} (U_1 \oplus U_2)
\]
\[
= \begin{pmatrix}
A_1^\dagger & A_1^\dagger \\
A_2^\dagger
\end{pmatrix} - \frac{1}{2} \begin{pmatrix}
A_1^\dagger P_1^* & A_2^\dagger P_2^*
\end{pmatrix}
\]
where \( P_1 = U_1^* \begin{pmatrix}
0 & 0 \\
0 & I
\end{pmatrix} \) \( U_1 \) is the projection from \( \text{Row}(A_1) \) to \( \text{Row}(A_1) \cap \text{Row}(A_2) \) and \( P_2 = U_2^* \begin{pmatrix}
I & 0 \\
0 & 0
\end{pmatrix} \) \( U_2 \) is the projection from \( \text{Row}(A_2) \) to \( \text{Row}(A_1) \cap \text{Row}(A_2) \). Note that
\[
A_1^\dagger P_1^* A_1 = V \begin{pmatrix}
0 & 0 \\
0 & C_2^\dagger
\end{pmatrix} \begin{pmatrix}
C_1 & 0 & 0 \\
0 & C_2 & 0
\end{pmatrix} V^* = V \begin{pmatrix}
0 & 0 \\
0 & C_2 & 0
\end{pmatrix} \begin{pmatrix}
0 & 0 \\
0 & C_3
\end{pmatrix} V^* = A_2^\dagger P_2^* A_2 := A_{[1,2]};
\]
If \( L \neq \emptyset \), then there must be \( \tilde{x} \) such that \( A_1 \tilde{x} = b_1 \) and \( A_2 \tilde{x} = b_2 \). Thus \( A_1^\dagger P_1^* b_1 = A_2^\dagger P_1^* A_1 \tilde{x} = A_2^\dagger P_2^* A_2 \tilde{x} = A_2^\dagger P_2^* b_2 : b_{[1,2]} \). Thus, the least square approximation of a given \( x \in \mathbb{R}^n \) on the set \( L \) is given by
\[
\tilde{x} = x - \begin{pmatrix}
A_1 \\
A_2
\end{pmatrix}^\dagger \begin{pmatrix}
A_1 \\
A_2
\end{pmatrix} x - \begin{pmatrix}
b_1 \\
b_2
\end{pmatrix}
\]
\[
= x - A_1^\dagger (A_1 x - b_1) - A_2^\dagger (A_2 x - b_2) + \frac{1}{2} A_1^\dagger P_1^* (A_1 x - b_1) + \frac{1}{2} A_2^\dagger P_2^* (A_2 x - b_2)
\]
\[
= x - A_1^\dagger (A_1 x - b_1) - A_2^\dagger (A_2 x - b_2) + A_{[1,2]}^\dagger (A_{[1,2]} x - b_{[1,2]})
\]
This proves the theorem for the case \( m = 2 \).

Now, suppose it is true for \( m = 2, \ldots, s - 1 \). The least square approximation of a given \( x \in \mathbb{R}^N \) on \( L \) is given by
\[
\tilde{x} = x - \begin{pmatrix}
A_1 \\
A_2 \\
\vdots \\
A_s
\end{pmatrix}^\dagger \begin{pmatrix}
A_1 \\
A_2 \\
\vdots \\
A_s
\end{pmatrix} x - \begin{pmatrix}
b_1 \\
b_2 \\
\vdots \\
b_s
\end{pmatrix}
\]
From the \( m = 2 \) case, we have
\[
\tilde{x} = x - \begin{pmatrix}
A_1 \\
\vdots \\
A_{s-1}
\end{pmatrix}^\dagger \begin{pmatrix}
A_1 \\
\vdots \\
A_{s-1}
\end{pmatrix} x - \begin{pmatrix}
b_1 \\
\vdots \\
b_{s-1}
\end{pmatrix} - A_1^\dagger (A_s x - b_s) + \begin{pmatrix}
A_{[1,s]} \\
\vdots \\
A_{[s-1,s]}
\end{pmatrix}^\dagger \begin{pmatrix}
A_{[1,s]} \\
\vdots \\
A_{[s-1,s]}
\end{pmatrix} x - \begin{pmatrix}
b_{[1,s]} \\
\vdots \\
b_{[s-1,s]}
\end{pmatrix},
\]
Apply the induction hypothesis to get

\[
y_1 = x - \begin{pmatrix} A_1 \\ \vdots \\ A_{s-1} \end{pmatrix}^\dagger \begin{pmatrix} (A_1) \\ \vdots \\ (A_{s-1}) \end{pmatrix} x - \begin{pmatrix} b_1 \\ \vdots \\ b_{s-1} \end{pmatrix}
\]

\[
= x + \sum_{r=1}^{s-1} (-1)^r \sum_{\{i_1, \ldots, i_r\} \subseteq \{1, \ldots, s-1\}} A_{[i_1, \ldots, i_r]}^\dagger (A_{[i_1, \ldots, i_r]} x - b_{[i_1, \ldots, i_r]})
\]

\[
y_2 = x - \begin{pmatrix} A_{[1,s]} \\ \vdots \\ A_{[s-1,s]} \end{pmatrix}^\dagger \begin{pmatrix} (A_{[1,s]}) \\ \vdots \\ (A_{[s-1,s]}) \end{pmatrix} x - \begin{pmatrix} b_{[1,s]} \\ \vdots \\ b_{[s-1,s]} \end{pmatrix}
\]

\[
= x + \sum_{r=1}^{s-1} (-1)^r \sum_{\{j_1, \ldots, j_r\} \subseteq \{1, \ldots, s-1\}} A_{[j_1, \ldots, j_r,s]}^\dagger (A_{[j_1, \ldots, j_r,s]} x - b_{[j_1, \ldots, j_r,s]})
\]

Then \( \hat{x} = y_1 - y_2 + x - A_s^\dagger (A_s x - b_s) \), which gives the desired equation.


VITA

Diane Christine Pelejo

Diane Pelejo was born on June 26, 1988 in the town of Rodriguez, Rizal in the Philippines. After hearing several stories about school from her older brother, she became really eager to start going to school. When she was six years old, she attended a private kindergarten and started to learn how to read and write. In 1995, she attended the Eulogio Rodriguez Elementary School (ERES) – a public elementary school in her town where she graduated valedictorian. In 2001, she earned a full scholarship to attend a private high school called Roosevelt College, where she graduated first honorable mention in her class. In 2005, she was accepted in the B.S. Mathematics program of the University of the Philippines Diliman (UPD). She graduated cum laude in 2009 and received the ‘Best Undergraduate Thesis in Mathematics’ award for her research on the $\Phi_J$-polar decomposition of matrices with rank 4. Her undergraduate thesis was published the following year in the journal Linear Algebra and Its Applications. She went on to teach Mathematics in UPD while working on her M.S. Mathematics degree. She obtained her Master’s degree in 2011 and decided that she wants to go to the USA for her doctoral degree. She loved studying matrices and linear algebra. She reached out to Dr. Chi-Kwong Li of the Mathematics Department of the College of William and Mary, who is an expert in the field. In 2013, she started working with Dr. Li on matrix-related problems in quantum information theory. After graduation, Diane will be returning to the Philippines, where an assistant professor position is waiting for her in UPD. She aims to contribute to research and development in Mathematics in her country.