ERROR BOUNDS AND PENALTY METHODS FOR OPTIMIZATION PROBLEMS OVER THE SIGN-CONSTRAINED STIEFEL MANIFOLD

YIFAN HE

PhD

The Hong Kong Polytechnic University

2025

The Hong Kong Polytechnic University

Department of Applied Mathematics

Error Bounds and Penalty Methods for Optimization Problems over the Sign-constrained Stiefel Manifold

Yifan He

A thesis submitted in partial fulfilment of the requirements

for the degree of Doctor of Philosophy

December, 2024

Certificate of Originality

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____(Signed)

Yifan He (Name of student)

To those who love me

Abstract

Optimization over sign-constrained Stiefel manifold requires that certain columns of variables in the Stiefel manifold are nonnegative and the remaining are nonpositive. When all columns are nonnegative, optimization problems with nonnegative orthogonal constraints arise, which have wide applications in fields such as signal and image processing. The properties of the constraints endow these models' physical meanings but also make the optimization problems hard to solve due to the combinatory property, for example, the quadratic assignment problem. One way to handle the difficulties is to seek help from penalty methods. The error bounds are commonly used to prove the exact penalty property, but the existence of the error bounds on the sign-constrained Stiefel manifold is unknown. In this thesis, we investigate the error bounds on the sign-constrained Stiefel manifold and design an effective algorithm called the smoothing proximal reweighted method (SPR) to solve the penalty problems.

In the first part, the sign-constrained Stiefel manifold in $\mathbb{R}^{n\times r}$ is a segment of the Stiefel manifold with fixed signs (nonnegative or nonpositive) for some entries of the matrices. We begin with the special case, the nonnegative Stiefel manifold, to discuss the error bounds, then extend the results to the sign-constrained Stiefel manifold. We present global and local error bounds that provide an inequality with easily computable residual functions and explicit coefficients to bound the distance from matrices in $\mathbb{R}^{n\times r}$ to the sign-constrained Stiefel manifold. Moreover, we show

that the error bounds cannot be improved except for the multiplicative constants under some mild conditions, which explains why two square-root terms are necessary for the bounds when $1 < r < n$ and why the ℓ_1 norm can be used in the bounds when $r = n$ or $r = 1$ for the sign constraints and orthogonality, respectively. The error bounds are applied to derive exact penalty methods for minimizing a Lipschitz continuous function with orthogonality and sign constraints. To this end, we show the improvement of adding nonnegativity to the first column of the variable for the sparse principal component analysis problem. In addition, the performance of penalizing one or both constraints to the objective function is compared through testing problems.

In the second part, we propose a proximal iteratively reweighted ℓ_2 algorithm to solve the non-Lipschitz penalized problem. Under the assumption that the objective function in the original problem is continuous, our algorithm has subsequence convergence property, a sufficient decrease in each iteration, and the distance between two adjacent iteration points is square summable, any accumulation point is a stationary point. Extensive numerical experiments including Projection to $\mathbb{S}_+^{n,r}$ and Quadratic Assignment Problem show the effectiveness of our algorithm.

This thesis contains research results of the following paper which is accepted during the period of my Ph.D. study.

- X. Chen, Y. He and Z. Zhang, *Tight Error Bounds for the Sign-Constrained Stiefel Manifold*, SIAM Journal on Optimization, 35(1):302-329, 2025.

Acknowledgements

With the completion of this thesis, my doctoral career is coming to an end. Looking back on this long and challenging journey, the complex mood surged into my heart and the words blocked my throat.

First of all, I give my deepest gratitude to my chief supervisor Prof. Xiaojun Chen. With her profound academic attainments, rigorous academic attitude and selfless dedication, Prof. Chen has become the beacon on my academic path. It is impossible to finish this thesis without her penetrating advice, motherly patient guidance, continuous encouragement and generous support. She organizes weekly seminars to let our students understand each other's research content and create a harmonious atmosphere.

I would also like to thank my co-supervisor Prof. Zaikun Zhang. It is more accurate to call him my big brother than to say he is my mentor. When I encounter academic problems, he will take me step by step to find the solutions instead of telling me the answers directly. He often discusses academic issues with me and takes me to lunch. In addition, he has given me great help in my life, often taking me to dinner and giving me medicine when I am ill. His research experience has had a profound impact on my study.

I am very grateful to my master's supervisor Prof. Zhengyu Wang (Nanjing University) for his suggestions on some issues in my research.

I own multiple thanks to my academic brothers and sisters: Prof. Congpei An,

Prof. Chao Zhang, Prof. Wei Bian, Prof. Hailin Sun, Prof. Yanfang Zhang, Prof. Yang Zhou, Prof. Bo Wen, Prof. Jie Jiang, Prof. Lei Yang, Dr. Hong Wang, Dr. Chao Li, Dr. Fang He, Dr. Jianfeng Luo, Dr. Shisen Liu, Dr. Wei Liu, Dr. Xiaozhou Wang, Dr. Xiaoxia Liu, Dr. Fan Wu, Dr. Fang Fei, Dr. You Zhao, Dr. Yong Zhao, Dr. Zhihua Zhao, Dr. Lin Chen, Dr. Kaixin Gao, Dr. Ming Huang, Dr. Xingbang Cui, Mr. Zicheng Qiu, Mr. Shijie Yu, Miss. Yue Wang, Mr. Xiao Zha, Mr. Zhouxing, Luo, Mr. Guang Wang, Mr. Wentao Ma, Miss Yixuan Zhang, Miss Xin Qu, Miss Lingzi Jin and Miss Yiyang Li. I am particularly grateful to Dr. Lei Wang and Dr. Chao Li, who not only carefully reviewed my thesis, but also made it better with their original views and valuable feedback.

Over the past few years, I have navigated through several challenges, often finding myself overwhelmed by the pressures from all corners. Miss Ning Wen burst out as a radiant beam of sunshine into my grey world. I often confided to her and she always comforted me patiently. Just as the storm clears to reveal azure skies, a joyous conclusion awaits us on the horizon of our future. The tumult of love and sorrow now belongs to the past, a fresh chapter beckons us forward.

I am very grateful to the Hong Kong Polytechnic University for the support I received during my pursuit of a doctorate. I also want to thank the support staff of the Department of Applied Mathematics for their help.

Lastly, I would like to thank my family for their unwavering support, and express my heartfelt gratitude to each one of you for your companionship, which has helped ease the solitude and challenges I have encountered on this journey. My affection for you all is profound.

Contents

List of Figures

List of Tables

List of Notations

$\mathbb{R}$	the set of real numbers
$\mathbb{R}^n$	the set of n-dimensional real vectors
$\mathbb{R}_+$	the nonnegative orthant
$\mathbb{R}^{n\times r}$	the set of all $n \times r$ real matrices
$\mathbb{R}_+^{n\times r}$	the set of all $n \times r$ nonnegative real matrices
$\mathbb{S}^{n,r}$	the set of all $n \times r$ orthonormal matrices, also called $n \times r$ Stiefel manifold
$\mathbb{R}_S^{n\times r}$	the subset of $\mathbb{R}^{n\times r}$ with column-wise nonnegative or nonpositive constraints on some columns
$\mathbb{S}_S^{n,r}$	the intersection of $\mathbb{S}^{n,r}$ and $\mathbb{R}_S^{n\times r}$
I_r	the $r \times r$ identity matrix
$\mathcal{P}, \mathcal{N}$	two disjoint subsets of $\{j : 1 \leq j \leq r\}$
S	the sign matrix
$\circ$	the Hadamard product
$\mathrm{dist}(\cdot, \cdot)$	the distance between two sets
X_-	the entry-wise nonnegative part of $-X$
$\sigma(X)$	the vector formed by the sigular values of matrix X
$\|\cdot\|_{\mathrm{F}}$	the Frobenius norm
$\|\cdot\|_{\ell_p}$	the ℓ_p norm

$\mathrm{cl}C$	the closure of set C
$\hat{\partial} f$	the regular subgradient of f
∂f	the (general) subgradient of f
$\hat{\partial}_{\mathcal{R}} f$	the Riemannian regular subdifferential of f
$\partial_{\mathcal{R}} f$	the Riemannian limiting subdifferential of f
$\mathcal{N}_{\Omega}(x)$	the normal cone of x at set Ω

Chapter 1

Introduction

1.1 Problem statement

Let n and r be two integers such that $1 \leq r \leq n$, and $\mathbb{S}^{n,r} \coloneqq \{X \in \mathbb{R}^{n\times r} : X^\mathsf{T} X = I_r\}$ be the Stiefel manifold, where I_r is the $r \times r$ identity matrix. Given two disjoint subsets $\mathcal{P}$ and $\mathcal{N}$ of $\{j : 1 \leq j \leq r\}$, denote

$$\mathbb{R}^{n\times r}_S \;\coloneqq\; \left\{X \in \mathbb{R}^{n\times r} : X_{i,j} \geq 0 \text{ for } j \in \mathcal{P} \text{ and } X_{i,j} \leq 0 \text{ for } j \in \mathcal{N},\ 1 \leq i \leq n\right\},$$

which is a subset of $\mathbb{R}^{n\times r}$ with column-wise nonnegative or nonpositive constraints on some columns.

In this thesis, we consider the *sign-constrained Stiefel manifold* defined as

$$\mathbb{S}^{n,r}_S \;\coloneqq\; \mathbb{S}^{n,r} \cap \mathbb{R}^{n\times r}_S.$$

When $\mathcal{P} = \{j : 1 \leq j \leq r\}$, $\mathbb{R}^{n\times r}_S$ reduces to the nonnegative orthant $\mathbb{R}^{n\times r}_+$, and $\mathbb{S}^{n,r}_S$ reduces to the *nonnegative Stiefel manifold* $\mathbb{S}^{n,r}_+ \coloneqq \{X \in \mathbb{S}^{n,r} : X \geq 0\}$.

If we define the sign matrix $S \in \mathbb{R}^{n\times r}$ as the matrix with

$$S_{i,j} \;=\; \begin{cases} 1, & \text{if } j \in \mathcal{P}, \\ -1, & \text{if } j \in \mathcal{N}, \\ 0, & \text{otherwise,} \end{cases} \qquad 1 \leq i \leq n, \tag{1.1.1}$$

then $\mathbb{S}^{n,r}_S$ can be formulated as

$$\mathbb{S}^{n,r}_S \;=\; \{X \in \mathbb{R}^{n\times r} : S \circ X \geq 0,\ X^\mathsf{T} X = I_r\},$$

where $\circ$ signifies the Hadamard product. We will investigate error bounds

$$\operatorname{dist}(X, \mathbb{S}_S^{n,r}) \;\le\; \nu\|(S \circ X)_-\|_{\mathrm{F}}^q \qquad \text{for} \quad X \in \mathbb{S}^{n,r}, \quad (1.1.2)$$

$$\operatorname{dist}(X, \mathbb{S}_S^{n,r}) \;\le\; \nu\|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}}^q \qquad \text{for} \quad X \in \mathbb{R}_S^{n\times r}, \quad (1.1.3)$$

$$\operatorname{dist}(X, \mathbb{S}_S^{n,r}) \;\le\; \nu(\|(S \circ X)_-\|_{\mathrm{F}}^q + \|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}}^q) \quad \text{for} \quad X \in \mathbb{R}^{n\times r}, \quad (1.1.4)$$

where ν and q are positive constants, and $Y_- \coloneqq \max\{-Y, 0\}$ stands for the entry-wise nonnegative part of $-Y$ for any matrix Y. The bounds (1.1.2)–(1.1.4) are global error bounds for $\mathbb{S}_S^{n,r}$ relative to $\mathbb{S}^{n,r}$, $\mathbb{R}_S^{n\times r}$, and $\mathbb{R}^{n\times r}$, respectively, with the first two being special cases of the last one.

According to the error bound of Luo-Pang presented in [22, Theorem 2.2], there exist $\nu > 0$ and $q > 0$ such that the inequalities in (1.1.2)–(1.1.4) hold for all X in a compact subset of $\mathbb{R}^{n\times r}$. Moreover, due to the error bound for polynomial systems given in [17, Corollary 3.8], for all X in a compact subset of $\mathbb{R}^{n\times r}$, there exists a ν such that the inequalities in (1.1.2)–(1.1.4) hold with a dimension-dependent value of q that is less than 6^{-2nr}. However, to the best of our knowledge, the explicit value of ν and the value of q that is independent of the dimension in (1.1.2)–(1.1.4) are still unknown even in the special case of $\mathbb{S}_S^{n,r} = \mathbb{S}_+^{n,r}$, and it is also unknown whether the error bounds hold in an unbounded set.

Being a fundamental concept in optimization, error bound plays a crucial role in both theory and methods for solving systems of equations and optimization problems [22, 25]. One of its applications is to develop penalty methods for constrained optimization problems. Let $F : \mathbb{R}^{n\times r} \to \mathbb{R}$ be a continuous function. The minimization problem

$$\min\{F(X) : X \in \mathbb{S}_S^{n,r}\} \qquad (1.1.5)$$

can be found in a wide range of optimization models in data science, including nonnegative principal component analysis [18, 36], nonnegative Laplacian embedding [21], discriminative nonnegative spectral clustering [34], orthogonal nonnegative

matrix factorization [26, 35], and some K-indicators models for data clustering [3, 31].

Even in the special case of $\mathbb{S}_S^{n,r} = \mathbb{S}_+^{n,r}$, the constraints of problem (1.1.5) are challenging to handle due to their combinatorial nature (note that, for example, $\mathbb{S}_+^{n,n}$ equals the set of all permutation matrices on $\mathbb{R}^n$). To deal with these difficult constraints, the penalty problems

$$\min \left\{F(X) + \mu\|(S \circ X)_-\|_{\mathrm{F}}^q : X \in \mathbb{S}^{n,r}\right\}, \tag{1.1.6}$$

$$\min \left\{F(X) + \mu\|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}}^q : X \in \mathbb{R}_S^{n\times r}\right\}, \tag{1.1.7}$$

$$\min \left\{F(X) + \mu(\|(S \circ X)_-\|_{\mathrm{F}}^q + \|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}}^q) : X \in \mathbb{R}^{n\times r}\right\}, \tag{1.1.8}$$

have been widely used for solving (1.1.5) with $\mathbb{S}_S^{n,r} = \mathbb{S}_+^{n,r}$, where μ is the penalty parameter. See for example [1, 27, 34, 36] and the references therein.

The recent two paper [14, 27] use different error bounds to derive different exact penalty models for optimization on the nonnegative Stifel manifold $\mathbb{S}_+^{n,r}$. In [14], the nonnegative Stifel manifold is reformed as $\mathbb{S}_+^{n,r} = \mathcal{OB}_+^{n,r} \bigcap \{X \in \mathbb{R}^{n\times r} : \|XV\|_{\mathrm{F}} = 1\}$, where $\mathcal{OB}_+^{n,r}$ is the oblique manifold and V is a constant matrix satisfying $\|V\|_{\mathrm{F}} = 1$ and $\min_{i,j\in[r]} \left[VV^{\mathsf{T}}\right]_{ij} > 0$, they elaborate the error bound over $\mathbb{S}_+^{n,r}$ for $X \in \mathcal{OB}_+^{n,r}$ and the residue function is $(\|XV\|_{\mathrm{F}}^q - 1 + \epsilon)^{\frac{p}{2}}$, where $p, q > 0$ and $\epsilon \geq 0$. On contrast, [27] obtain a local Lipschitzian error bound over for those feasible points without zero rows when $n > r > 1$, specifically, for all $X \in \mathcal{B}(\bar{X}, \delta)$, where $\bar{X}$ has no zero rows and $\delta > 0$, it holds $\mathrm{dist}(X, \mathbb{S}_+^{n,r}) \leq (\kappa + 1)(\mathrm{dist}(X, \mathbb{R}_+^{n\times r}) + \mathrm{dist}(X, \mathbb{S}^{n,r}))$. One advantage of this error bound is the exponent of the penalty function $\|(X)_-\|_{\ell_1}$ can be set to 1. Moreover, when $n > r = 1$ or $n = r$, their error bound can be a global one. However, the exactness of problems (1.1.6)–(1.1.8) regarding global minimizers and local minimizers of problem (1.1.5) is not well understood.

1.2 Contribution of the thesis

The first contribution of this thesis is to establish the error bounds (1.1.2)–(1.1.4) with $\nu = 15r^{\frac{3}{4}}$ and $q = 1/2$ without any additional restriction on X. Moreover, we demonstrate that the error bounds cannot hold for $q > 1/2$ under mild conditions when $1 < r < n$ and $\mathbb{S}_S^{n,r} = \mathbb{S}_+^{n,r}$. In addition, we show that the error bounds (1.1.2)–(1.1.4) hold with $q = 1$ and $\nu = 7\sqrt{r}$ when $|\mathcal{P}| + |\mathcal{N}| = 1$, and hold with $q = 1$ and $\nu = 9n$ when $|\mathcal{P}| + |\mathcal{N}| = n$, but they cannot hold with $q > 1$. As an application of error bounds (1.1.2)–(1.1.4) with $\nu = 15r^{\frac{3}{4}}$ and $q = 1/2$, we show the exactness of the penalty problems (1.1.6) and (1.1.7) under the assumption that F is Lipschitz continuous, taking $\mathbb{S}_S^{n,r} = \mathbb{S}_+^{n,r}$ as an example. Moreover, we show the existence of Lipschitz continuous functions such that penalty problems (1.1.6) and (1.1.7) with $q > 1/2$ are not exact for global and local minimizers of the corresponding constrained problems. The values of q in error bounds (1.1.2)–(1.1.4) for some special sign matrices $S \in \mathbb{R}^{n\times r}$ defined in (1.1.1) by $\mathcal{P}$ and $\mathcal{N}$ are summarized in Table 1.1.

S	hold	fail
$\lvert\mathcal{P}\rvert = r$ or $\lvert\mathcal{N}\rvert = r$, $\quad 1 < r < n$	$q = 1/2$	$q > 1/2$
$\lvert\mathcal{P}\rvert = 1$ or $\lvert\mathcal{N}\rvert = 1$, $\quad 1 \le r \le n$	$q = 1$	$q > 1$
$\lvert\mathcal{P}\rvert + \lvert\mathcal{N}\rvert = n$, $\quad r = n$	$q = 1$	$q > 1$

Table 1.1: The cases that error bounds (1.1.2)–(1.1.4) hold or fail for some special sign matrices $S \in \mathbb{R}^{n\times r}$

Very recently, our error bounds and matrix inequalities have been used to study constant modulus optimization and optimal orthogonal channel selection [2, 19, 20], which have a wide variety of applications in signal processing, communications, and data science.

The second contribution of this thesis is to design an algorithm to solve problem (1.1.5). We penalize the sign-constraint to the objective function and propose

the proximal iteratively reweighted ℓ_2 algorithm to solve the non-Lipschitz penalized problem. Under the assumption that the objective function in the original problem is continuous, our algorithm has subsequence convergence property, a sufficient decrease in each iteration, the distance between two adjacent iteration points is square summable, and any accumulation point is a stationary point. Extensive numerical experiments including Projection to $\mathbb{S}_+^{n,r}$ and Quadratic Assignment Problem show the effectiveness of our algorithm.

1.3 Organization of the thesis

The rest of the thesis is organized as follows. In Chapter 2, we introduce some notation and preliminaries. Chapter 3 derives the error bounds (1.1.2)–(1.1.4) in the special case of $\mathbb{S}_S^{n,r} = \mathbb{S}_+^{n,r}$, then extends these bounds to the general case. Chapter 4 investigates the exactness of the penalty problems (1.1.6)–(1.1.8) using the new error bounds and comprises the proximal iteratively reweighted ℓ_2 algorithm (PIRL2) and convergence properties. Chapter 5 discusses some issues and advantages of penalty method for (1.1.5) via sparse trace minimization problem, it also includes the numerical experiments of PIRL2. We conclude the thesis in Chapter 6.

Chapter 2

Basic notations and preliminaries

2.1 Basic notations

For matrix $X \in \mathbb{R}^{n\times r}$, $X_+ \coloneqq \max\{X,\, 0\} = X + X_-$ is the projection of X onto $\mathbb{R}^{n\times r}_+$. In addition, the singular value vector of X is denoted by $\sigma(X) \in \mathbb{R}^r$, the entries of which are in the descent order. Meanwhile, $\Sigma(X) \in \mathbb{R}^{n\times r}$ is the matrix such that $X = U\Sigma(X)V^\mathsf{T}$ is the singular value decomposition of X, the diagonal of $\Sigma(X)$ being $\sigma(X)$. We use $\mathbf{1}$ to denote the vector with all entries being one, and its dimension will be clear from the context.

Unless otherwise specified, $\|\cdot\|$ stands for a general vector norm. For any constant $p \in [1, +\infty)$, we use $\|\cdot\|_p$ to represent either the ℓ_p-norm of vectors or the operator norm induced by this vector norm for matrices. In addition, we use $\|\cdot\|_{\ell_p}$ to denote the *entry-wise* ℓ_p-norm of a matrix, namely the ℓ_p-norm of the vector that contains all the entries of the matrix. Note that $\|\cdot\|_{\ell_2}$ is the Frobenius norm, which is also denoted by $\|\cdot\|_{\mathrm{F}}$. When $\mathbb{R}^{n\times r}$ is equipped with the Frobenius norm, we use $\mathcal{B}(X, \delta)$ to represent the open ball in $\mathbb{R}^{n\times r}$ centered at a point $X \in \mathbb{R}^{n\times r}$ with a radius $\delta > 0$, and $\mathrm{dist}(X, \mathcal{T})$ to denote the distance from a point $X \in \mathbb{R}^{n\times r}$ to a set $\mathcal{T} \subset \mathbb{R}^{n\times r}$. Finally, given a minimization problem, we use Argmin to denote the set of global minimizers.

2.2 Matrix inequalities

Lemma 2.1 is fundamental for the analysis of distances between matrices. This lemma is stated for unitarily invariant norms (see [12, Section 3.5] for this concept), although we are most interested in the case with the Frobenius norm.

Lemma 2.1 (Mirsky)**.** *For any matrices $X \in \mathbb{R}^{n\times r}$ and $Y \in \mathbb{R}^{n\times r}$, we have*

$$\|\Sigma(X) - \Sigma(Y)\| \;\leq\; \|X - Y\| \tag{2.2.1}$$

for any unitarily invariant norm $\|\cdot\|$ on $\mathbb{R}^{n\times r}$. When $\|\cdot\|$ is the Frobenius norm, the equality holds in (2.2.1) *if and only if there exist orthogonal matrices $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{r\times r}$ such that $X = U\Sigma(X)V^{\mathsf{T}}$ and $Y = U\Sigma(Y)V^{\mathsf{T}}$.*

The square case (i.e., $n = r$) of inequality (2.2.1) is due to Mirsky [23, Theorem 5], and the general case can be found in [13, Theorem 7.4.9.1]. A direct corollary of Lemma 2.1 is the following Hoffman-Wielandt [11] type bound for singular values, which is equivalent to the von Neumann trace inequality [30, Theorem I] (see also [16, Theorem 2.1]).

Lemma 2.2 (von Neumann)**.** *For any matrices $X \in \mathbb{R}^{n\times r}$ and $Y \in \mathbb{R}^{n\times r}$, we have*

$$\|\sigma(X) - \sigma(Y)\|_2 \;\leq\; \|X - Y\|_{\mathrm{F}},$$

and equivalently, $\mathrm{tr}(X^{\mathsf{T}}Y) \leq \sigma(X)^{\mathsf{T}}\sigma(Y)$.

The following lemma is another consequence of Lemma 2.1. For this result, recall that each matrix $X \in \mathbb{R}^{n\times r}$ has a polar decomposition in the form of $X = UP$, where U belongs to $\mathbb{S}^{n,r}$ and $P = (X^{\mathsf{T}}X)^{\frac{1}{2}}$, with U being called a unitary polar factor of X. The square case of this lemma is due to Fan and Hoffman [9, Theorem 1]. For the general case, see [10, Theorem 8.4], which details a proof based on Lemma 2.1.

Lemma 2.3 (Fan-Hoffman)**.** *If $U \in \mathbb{R}^{n\times r}$ is a unitary polar factor of a matrix $X \in \mathbb{R}^{n\times r}$, then*

$$\|X - U\| \;=\; \min\{\|X - V\| : V \in \mathbb{S}^{n,r}\}$$

for any unitarily invariant norm $\|\cdot\|$ on $\mathbb{R}^{n\times r}$.

Lemma 2.4 collects a few basic facts on the distance from a matrix in $\mathbb{R}^{n\times r}$ to $\mathbb{S}^{n,r}$.

Lemma 2.4. *For any matrix $X \in \mathbb{R}^{n\times r}$, we have*

$$\mathrm{dist}(X, \mathbb{S}^{n,r}) \;=\; \|\sigma(X) - \mathbf{1}\|_2 \;\le\; \min\Big\{\|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}},\; r^{\frac{1}{4}}\|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}}^{\frac{1}{2}}\Big\}.$$

In addition, $\|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}} \le (\|X\|_2 + 1)\|\sigma(X) - \mathbf{1}\|_2$.

Proof. Let $U \in \mathbb{S}^{n,r}$ be a unitary polar factor of X. By Lemma 2.3,

$$\mathrm{dist}(X, \mathbb{S}^{n,r}) = \|X - U\|_{\mathrm{F}} = \|U^{\mathsf{T}}(X - U)\|_{\mathrm{F}} = \|(X^{\mathsf{T}}X)^{\frac{1}{2}} - I_r\|_{\mathrm{F}} = \|\sigma(X) - \mathbf{1}\|_2.$$

The entry-wise inequalities $|\sigma(X) - \mathbf{1}| \le |\sigma(X)^2 - \mathbf{1}| \le (\|X\|_2 + 1)|\sigma(X) - \mathbf{1}|$ imply

$$\|\sigma(X) - \mathbf{1}\|_2 \;\le\; \|\sigma(X)^2 - \mathbf{1}\|_2 \;\le\; (\|X\|_2 + 1)\|\sigma(X) - \mathbf{1}\|_2. \tag{2.2.2}$$

Noting that $\|\sigma(X)^2 - \mathbf{1}\|_2 = \|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}}$, we obtain from (2.2.2) that

$$\|\sigma(X) - \mathbf{1}\|_2 \le \|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}} \le (\|X\|_2 + 1)\|\sigma(X) - \mathbf{1}\|_2.$$

Finally, since $|\sigma(X) - \mathbf{1}|^2 \le |\sigma(X)^2 - \mathbf{1}|$, we have

$$\|\sigma(X) - \mathbf{1}\|_2^2 \;\le\; \|\sigma(X)^2 - \mathbf{1}\|_1 \;\le\; \sqrt{r}\|\sigma(X)^2 - \mathbf{1}\|_2 \;=\; \sqrt{r}\|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}}.$$

The proof is complete. □

By Lemmas 2.3 and 2.4, $\mathrm{dist}(X, \mathbb{S}_+^{n,r}) = \|\sigma(X) - \mathbf{1}\|_2$ if X has a nonnegative unitary polar factor. It is the case in the following lemma, where this factor is $X(X^{\mathsf{T}}X)^{-\frac{1}{2}}$.

Lemma 2.5. *For a matrix $X \in \mathbb{R}_+^{n\times r}$, if $X^\mathsf{T}X$ is nonsingular and diagonal, then*

$$\operatorname{dist}(X, \mathbb{S}_+^{n,r}) \;=\; \|\sigma(X) - \mathbf{1}\|_2.$$

Lemma 2.6 is an elementary property of $\mathbb{S}_+^{n,r}$. We omit the proof.

Lemma 2.6. *For a matrix $X \in \mathbb{S}_+^{n,r}$, each row of X has at most one nonzero entry.*

Chapter 3

Error bounds for the sign-constrained Stiefel manifold

This chapter will establish the error bounds (1.1.2)–(1.1.4). Section 3.1 to Section 3.4 discusses the special case of error bounds over $\mathbb{S}_S^{n,r} = \mathbb{S}_+^{n,r}$, where $S \circ X$ reduces to X. Section 3.2 demonstrates (1.1.2)–(1.1.4) with $q = 1$ when $r = 1$ or $r = n$, and points out that they cannot hold with $q > 1$ regardless of $r \in \{1, \ldots, n\}$. In Section 3.3, we derive the bounds (1.1.2)–(1.1.4) with $q = 1/2$ for $1 \leq r \leq n$, and Section 3.4 elaborates on the tightness of these bounds when $1 < r < n$. As an application of our results, we briefly discuss the linear regularity of $\mathbb{R}_+^{n\times r}$ and $\mathbb{S}^{n,r}$ in Section 3.5. Based on the analysis in previous sections, we derive the error bounds for $\mathbb{S}_S^{n,r}$ in Sections 3.6 and Section 3.7.

3.1 Brief discussion on error bounds

General discussions on error bounds can be found in [8, Section 6.1]. Here we focus on error bounds for $\mathbb{S}_+^{n,r}$ defined by two special functions

$$\rho_1(X) \coloneqq \|X_-\|_{\mathrm{F}}^{q_1} + \|\sigma(X) - \mathbf{1}\|_2^{q_2},$$

$$\rho_2(X) \coloneqq \|X_-\|_{\mathrm{F}}^{q_1} + \|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}}^{q_2},$$

where q_1 and q_2 are positive constants. These functions are residual functions for $\mathbb{S}_+^{n,r}$ relative to $\mathbb{R}^{n\times r}$, namely nonnegative-valued functions on $\mathbb{R}^{n\times r}$ whose zeros coincide

with the elements of $\mathbb{S}_+^{n,r}$. The residual function ρ_2 is easily computable and it reduces to the one in (1.1.4) when $q_1 = q_2 = q$.

We say that ρ_1 defines a local error bound for $\mathbb{S}_+^{n,r}$ relative to $\mathbb{R}^{n\times r}$ if there exist positive constants ϵ and ν such that

$$\operatorname{dist}(X,\, \mathbb{S}_+^{n,\,r}) \;\le\; \nu\rho_1(X) \tag{3.1.1}$$

for all $X \in \mathbb{R}^{n\times r}$ satisfying $\|X_-\|_{\mathrm{F}} + \|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}} \le \epsilon$, and we say it defines a global error bound for $\mathbb{S}_+^{n,r}$ relative to $\mathbb{R}^{n\times r}$ if (3.1.1) holds for all $X \in \mathbb{R}^{n\times r}$. Likewise, we can use ρ_1 to define error bounds for $\mathbb{S}_+^{n,r}$ relative to any set $\mathcal{S} \subset \mathbb{R}^{n\times r}$ that contains $\mathbb{S}_+^{n,r}$, for example, $\mathcal{S} = \mathbb{R}_+^{n\times r}$, in which case ρ_1 reduces to its second term. Similar things can be said about ρ_2. Theorems 3.3 and 3.6 will specify the precise range of q_1 and q_2 so that ρ_1 and ρ_2 define local or global error bounds for $\mathbb{S}_+^{n,r}$ relative to $\mathbb{R}^{n\times r}$.

3.2 Tight error bounds with $r = 1$ or $r = n$

In this section, we show that the error bounds (1.1.2)–(1.1.4) hold for $q = 1$ when $r = 1$ or $r = n$. Moreover, we explain why bounds (1.1.2)–(1.1.4) cannot hold for $q > 1$ in general.

The bound for $r = 1$ is easy to establish due to the simple fact that

$$\operatorname{dist}(x,\, \mathbb{S}_+^{n,1}) \;=\; \operatorname{dist}(x,\, \mathbb{S}^{n,1}) \;=\; \big|\|x\|_2 - 1\big| \quad \text{for all} \quad x \in \mathbb{R}_+^n. \tag{3.2.1}$$

Indeed, when $x = 0$, this is trivial; when $x \neq 0$, equality (3.2.1) is true because the projection of x onto $\mathbb{S}_+^{n,1}$ equals its projection onto $\mathbb{S}^{n,1}$, which is $x/\|x\|_2 \ge 0$.

Theorem 3.1. *For any vector $x \in \mathbb{R}^n$,*

$$\operatorname{dist}(x,\, \mathbb{S}_+^{n,1}) \;\le\; 2\|x_-\|_2 + \big|\|x\|_2 - 1\big|.$$

Proof. As observed above, $\operatorname{dist}(x_+, \mathbb{S}_+^{n,1}) = \big|\|x_+\|_2 - 1\big|$. Meanwhile,

$$\big|\|x_+\|_2 - 1\big| - \big|\|x\|_2 - 1\big| \;\le\; \big|\|x_+\|_2 - \|x\|_2\big| \;\le\; \|x_+ - x\|_2 \;=\; \|x_-\|_2.$$

Thus $\operatorname{dist}(x, \mathbb{S}_+^{n,1}) \le \|x_-\|_2 + \operatorname{dist}(x_+, \mathbb{S}_+^{n,1}) \le 2\|x_-\|_2 + \big|\|x\|_2 - 1\big|$. □

To establish the error bounds for $r = n$, we first prove Proposition 3.1, which is essentially a weakened version of the observation (3.2.1) in the current situation. Note that the matrix Y defined in the proof below is indeed the rounding matrix proposed in [14, Procedure 1].

Proposition 3.1. *For any matrix $X \in \mathbb{R}_+^{n\times n}$, if $\|\sigma(X) - \mathbf{1}\|_2 < 1/(4\sqrt{n})$, then*

$$\operatorname{dist}(X, \mathbb{S}_+^{n,n}) \;\le\; 7\sqrt{n}\|\sigma(X) - \mathbf{1}\|_2. \tag{3.2.2}$$

Proof. For each $i \in \{1, \dots, n\}$, take the smallest integer $l_i \in \{1, \dots, r\}$ so that

$$X_{i,l_i} \;=\; \max\{X_{i,j} : j = 1, \dots, r\}.$$

Consider the matrix $Y \in \mathbb{R}_+^{n,r}$ defined by

$$Y_{i,j} \;=\; \begin{cases} X_{i,l_i} & \text{if } j = l_i, \\ 0 & \text{otherwise.} \end{cases} \tag{3.2.3}$$

We will demonstrate (3.2.2) by establishing bounds for $\|X - Y\|_{\mathrm{F}}$ and $\operatorname{dist}(Y, \mathbb{S}_+^{n,n})$.

Consider $\|X - Y\|_{\mathrm{F}}$ first. Due to the fact that $\|\sigma(X) - \mathbf{1}\|_2 < 1/(4\sqrt{n})$, all the n singular values of X are at least $3/4$. Since $X \ge 0$ and $X_{i,l_i} = \max\{X_{i,j} : j = 1, \dots, n\}$, we have

$$X_{i,l_i} \;\ge\; \frac{1}{\sqrt{n}}\big(XX^{\mathsf{T}}\big)^{\frac{1}{2}}_{i,i} \;\ge\; \frac{3}{4\sqrt{n}} \quad \text{for each} \quad i \in \{1, \dots, n\}.$$

Fix an integer $j \in \{1, \dots, r\}$. For each $l \in \{1, \dots, r\}$, define

$$\mathbb{1}(j \ne l) \;=\; \mathbb{1}(l \ne j) \;=\; \begin{cases} 1 & \text{if } l \ne j, \\ 0 & \text{if } l = j. \end{cases}$$

With x^j and y^j denoting the jth columns of X and Y, respectively, we have

$$\begin{aligned}
\frac{9}{16n}\|x^j - y^j\|_2^2 &= \frac{9}{16n}\sum_{i=1}^{n} X_{i,j}^2 \mathbb{1}(j \neq l_i) \\
&\leq \sum_{i=1}^{n} X_{i,l_i}^2 X_{i,j}^2 \mathbb{1}(l_i \neq j) \\
&\leq \sum_{l=1}^{n}\sum_{i=1}^{n} X_{i,l}^2 X_{i,j}^2 \mathbb{1}(l \neq j) \\
&\leq \sum_{l=1}^{n}\left(\sum_{i=1}^{n} X_{i,l}X_{i,j}\right)^2 \mathbb{1}(l \neq j) \\
&= \sum_{l=1}^{n}\left(X^{\mathsf{T}}X - I_n\right)_{l,j}^2 \mathbb{1}(l \neq j).
\end{aligned}$$

Hence

$$\|X - Y\|_{\mathrm{F}} \;\leq\; \frac{4}{3}\sqrt{n}\|X^{\mathsf{T}}X - I_n\|_{\mathrm{F}}.$$

By Lemma 2.4 and the fact that $\|X\|_2 \leq 1 + \|\sigma(X) - \mathbf{1}\|_2 \leq 5/4$, we have further

$$\|X - Y\|_{\mathrm{F}} \;\leq\; \frac{4}{3}\sqrt{n}(\|X\|_2 + 1)\|\sigma(X) - \mathbf{1}\|_2 \;\leq\; 3\sqrt{n}\|\sigma(X) - \mathbf{1}\|_2. \tag{3.2.4}$$

Now we estimate $\mathrm{dist}(Y, \mathbb{S}_+^{n,n})$. According to inequality (3.2.4) and Lemma 2.2,

$$\|\sigma(Y) - \mathbf{1}\|_2 \;\leq\; \|X - Y\|_{\mathrm{F}} + \|\sigma(X) - \mathbf{1}\|_2 \;\leq\; 4\sqrt{n}\|\sigma(X) - \mathbf{1}\|_2. \tag{3.2.5}$$

Since $\|\sigma(X) - \mathbf{1}\|_2 < 1/(4\sqrt{n})$, we have $\|\sigma(Y) - \mathbf{1}\|_2 < 1$, which implies that $Y^{\mathsf{T}}Y$ is nonsingular. Since Y has at most one nonzero entry in each row, it is clear that $Y^{\mathsf{T}}Y$ is diagonal. Thus we can invoke Lemma 2.5 and obtain

$$\mathrm{dist}(Y, \mathbb{S}_+^{n,n}) \;=\; \|\sigma(Y) - \mathbf{1}\|_2.$$

Therefore, combining inequalities (3.2.4) and (3.2.5), we conclude that (3.2.2) is true. □

Theorem 3.2 presents global and local error bounds for $\mathbb{S}_+^{n,n}$ relative to $\mathbb{R}^{n\times n}$.

Theorem 3.2. *For any matrix $X\in\mathbb{R}^{n\times n}$, we have*

$$\operatorname{dist}(X,\,\mathbb{S}_+^{n,n}) \;\le\; 9n\,(\|X_-\|_{\mathrm{F}}+\|\sigma(X)-\mathbf{1}\|_2)\,. \tag{3.2.6}$$

Moreover, if $\|X_-\|_{\mathrm{F}}+\|\sigma(X)-\mathbf{1}\|_2<1/(4\sqrt{n})$, then

$$\operatorname{dist}(X,\,\mathbb{S}_+^{n,n}) \;\le\; 8\sqrt{n}\,(\|X_-\|_{\mathrm{F}}+\|\sigma(X)-\mathbf{1}\|_2)\,. \tag{3.2.7}$$

Proof. We first prove (3.2.7), assuming that $\|X_-\|_{\mathrm{F}}+\|\sigma(X)-\mathbf{1}\|_2<1/(4\sqrt{n})$. By Lemma 2.2, this assumption ensures $\|\sigma(X_+)-\mathbf{1}\|_2<1/(4\sqrt{n})$. Thus Proposition 3.1 renders

$$\operatorname{dist}(X_+,\,\mathbb{S}_+^{n,n}) \;\le\; 7\sqrt{n}\|\sigma(X_+)-\mathbf{1}\|_2 \;\le\; 7\sqrt{n}\,(\|X_-\|_{\mathrm{F}}+\|\sigma(X)-\mathbf{1}\|_2)\,,$$

which justifies inequality (3.2.7) since $\operatorname{dist}(X,\,\mathbb{S}_+^{n,n})\le\|X_-\|_{\mathrm{F}}+\operatorname{dist}(X_+,\,\mathbb{S}_+^{n,n})$.

Now we consider inequality (3.2.6). If $\|X_-\|_{\mathrm{F}}+\|\sigma(X)-\mathbf{1}\|_2<1/(4\sqrt{n})$, then (3.2.6) holds due to (3.2.7). When $\|X_-\|_{\mathrm{F}}+\|\sigma(X)-\mathbf{1}\|_2\ge 1/(4\sqrt{n})$, inequality (3.2.6) is justified by

$$\begin{aligned}\operatorname{dist}(X,\,\mathbb{S}_+^{n,n}) \;&\le\; \operatorname{dist}(X,\,\mathbb{S}^{n,n})+2\sqrt{n}\\ &\le \|\sigma(X)-\mathbf{1}\|_2+8n(\|X_-\|_{\mathrm{F}}+\|\sigma(X)-\mathbf{1}\|_2)\\ &\le 9n\,(\|X_-\|_{\mathrm{F}}+\|\sigma(X)-\mathbf{1}\|_2)\,,\end{aligned}$$

where the first inequality holds because the diameter of $\mathbb{S}^{n,n}$ is $2\sqrt{n}$. □

Remark 3.1. *Since $\big|\|x\|_2-1\big|\le\big|\|x\|_2^2-1\big|$ and $\|\sigma(X)-\mathbf{1}\|_2\le\|X^{\mathsf{T}}X-I_n\|_{\mathrm{F}}$, Theorems 3.1 and 3.2 imply the error bounds (1.1.2)–(1.1.4) with $q=1$ for $r\in\{1,n\}$. These bounds cannot be improved except for the multiplicative constants.*

Indeed, for any matrix $X \in \mathbb{R}^{n\times r}$ with $r \in \{1,\dots,n\}$ and $\|X\|_2 \le 1$, we have

$$\begin{aligned}
\mathrm{dist}(X,\, \mathbb{S}_+^{n,r}) \;&\ge\; \max\big\{\mathrm{dist}(X,\, \mathbb{R}_+^{n\times r}),\; \mathrm{dist}(X,\, \mathbb{S}^{n,r})\big\} \\
&\ge \frac{1}{2}\big[\mathrm{dist}(X,\, \mathbb{R}_+^{n\times r}) + \mathrm{dist}(X,\, \mathbb{S}^{n,r})\big] \\
&= \frac{1}{2}\left(\|X_-\|_{\mathrm{F}} + \|\sigma(X) - \mathbf{1}\|_2\right) \\
&\ge \frac{1}{2}\|X_-\|_{\mathrm{F}} + \frac{1}{4}\|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}},
\end{aligned} \tag{3.2.8}$$

where the last two lines apply Lemma 2.4. This also implies that the bounds (1.1.2)–(1.1.4) cannot hold for any $r \in \{1,\dots,n\}$ with $q > 1$.

Theorem 3.3 is an extension of Theorems 3.1 and 3.2. It specifies the possible exponents of $\|X_-\|_{\mathrm{F}}$ and $\|\sigma(X) - \mathbf{1}\|_2$ or $\|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}}$ in local and global error bounds for $\mathbb{S}_+^{n,r}$ relative to $\mathbb{R}^{n\times r}$ for $r \in \{1, n\}$. As we will see from (b) of this theorem and its proof, when $r = 1$ or $r = n$, the error bound (1.1.2) can hold if and only if $q \le 1$, whereas (1.1.3) and (1.1.4) can hold if and only if $1/2 \le q \le 1$.

Theorem 3.3. *Let q_1 and q_2 be positive constants. Suppose that $r = 1$ or $r = n$.*

(a) *The function $\rho_1(X) := \|X_-\|_{\mathrm{F}}^{q_1} + \|\sigma(X) - \mathbf{1}\|_2^{q_2}$ defines a local error bound for $\mathbb{S}_+^{n,r}$ relative to $\mathbb{R}^{n\times r}$ if and only if $q_1 \le 1$ and $q_2 \le 1$, and it defines a global error bound if and only if $q_1 \le q_2 = 1$.*

(b) *The function $\rho_2(X) := \|X_-\|_{\mathrm{F}}^{q_1} + \|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}}^{q_2}$ defines a local error bound for $\mathbb{S}_+^{n,r}$ relative to $\mathbb{R}^{n\times r}$ if and only if $q_1 \le 1$ and $q_2 \le 1$, and it defines a global error bound if and only if $q_1 \le 1$ and $1/2 \le q_2 \le 1$.*

Proof. We consider only the case with $r = n$. The other case is similar.

(a) Based on (3.2.7) and (3.2.8), it is easy to check that ρ_1 defines a local error bound for $\mathbb{S}_+^{n,n}$ relative to $\mathbb{R}^{n\times n}$ if and only if $q_1 \le 1$ and $q_2 \le 1$. Hence we only need to consider the global error bound.

Suppose that $q_1 \le q_2 = 1$. Let us show that

$$\operatorname{dist}(X, \mathbb{S}_+^{n,n}) \;\le\; 9n\left(\|X_-\|_{\mathrm{F}}^{q_1} + \|\sigma(X) - \mathbf{1}\|_2\right) \;=\; 9n\rho_1(X) \tag{3.2.9}$$

for $X \in \mathbb{R}^{n\times n}$. If $\|X_-\|_{\mathrm{F}} \le 1$, then (3.2.9) follows from (3.2.6). When $\|X_-\|_{\mathrm{F}} > 1$,

$$\operatorname{dist}(X, \mathbb{S}_+^{n,n}) \;\le\; \operatorname{dist}(X, \mathbb{S}^{n,n}) + 2\sqrt{n} \;\le\; \|\sigma(X) - \mathbf{1}\|_2 + 2\sqrt{n}\|X_-\|_{\mathrm{F}}^{q_1},$$

which validates (3.2.9) again. Hence ρ_1 defines a global error bound for $\mathbb{S}_+^{n,n}$ relative to $\mathbb{R}^{n\times n}$.

Now suppose that ρ_1 defines a global error bound for $\mathbb{S}_+^{n,n}$ relative to $\mathbb{R}^{n\times n}$. Then it also defines a local error bound, implying $q_1 \le 1$ and $q_2 \le 1$. Consider a sequence $\{X_k\} \subset \mathbb{R}_+^{n\times n}$ such that $X_k^{\mathsf{T}} X_k = kI_n$ for each $k \ge 1$. Then

$$\operatorname{dist}(X_k, \mathbb{S}_+^{n,n}) \;\ge\; \operatorname{dist}(X_k, \mathbb{S}^{n,n}) \;=\; \|\sigma(X_k) - \mathbf{1}\|_2 \;=\; [\rho_1(X_k)]^{\frac{1}{q_2}} \;\to\; \infty.$$

By assumption, $\operatorname{dist}(X_k, \mathbb{S}_+^{n,n}) \le \nu\rho_1(X_k)$ for each $k \ge 1$ with a constant ν. Hence we know $q_2 \ge 1$. To conclude, we have $q_1 \le q_2 = 1$. The proof for (a) is complete.

(b) Based on (3.2.7), (3.2.8), and the fact that $\|\sigma(X) - \mathbf{1}\|_2 \le \|X^{\mathsf{T}}X - I_n\|_{\mathrm{F}}$ (Lemma 2.4), it is easy to check that ρ_2 defines a local error bound for $\mathbb{S}_+^{n,n}$ relative to $\mathbb{R}^{n\times n}$ if and only if $q_1 \le 1$ and $q_2 \le 1$. Hence we consider only the global error bound.

Suppose that $q_1 \le 1$ and $1/2 \le q_2 \le 1$. We will show that

$$\operatorname{dist}(X, \mathbb{S}_+^{n,n}) \;\le\; 9n\left(\|X_-\|_{\mathrm{F}}^{q_1} + \|X^{\mathsf{T}}X - I_n\|_{\mathrm{F}}^{q_2}\right) \;=\; 9n\rho_2(X) \tag{3.2.10}$$

for $X \in \mathbb{R}^{n\times n}$. If $\|X^{\mathsf{T}}X - I_n\|_{\mathrm{F}} \le 1$, then (3.2.10) holds because of (3.2.9) and the fact that $\|\sigma(X) - \mathbf{1}\|_2 \le \|X^{\mathsf{T}}X - I_n\|_{\mathrm{F}}$. When $\|X^{\mathsf{T}}X - I_n\|_{\mathrm{F}} > 1$,

$$\begin{aligned}
\operatorname{dist}(X, \mathbb{S}_+^{n,n}) \;&\le\; \operatorname{dist}(X, \mathbb{S}^{n,n}) + 2\sqrt{n} \\
&\le n^{\frac{1}{4}}\|X^{\mathsf{T}}X - I_n\|_{\mathrm{F}}^{\frac{1}{2}} + 2\sqrt{n}\|X^{\mathsf{T}}X - I_n\|_{\mathrm{F}}^{q_2} \\
&\le (n^{\frac{1}{4}} + 2\sqrt{n})\|X^{\mathsf{T}}X - I_n\|_{\mathrm{F}}^{q_2},
\end{aligned}$$

justifying (3.2.10) again, where the second inequality applies Lemma 2.4. Hence ρ_2 defines a global error bound for $\mathbb{S}^{n,n}_+$ relative to $\mathbb{R}^{n\times n}$.

Now suppose that ρ_2 defines a global error bound for $\mathbb{S}^{n,n}_+$ relative to $\mathbb{R}^{n\times n}$. Then $q_1 \le 1$ and $q_2 \le 1$, as ρ_2 also defines a local error bound. Consider again a sequence $\{X_k\} \subset \mathbb{R}^{n\times n}_+$ such that $X_k^{\mathsf{T}} X_k = kI_n$ for each $k \ge 1$. Then

$$\operatorname{dist}(X_k, \mathbb{S}^{n,n}_+) \ge \|\sigma(X_k) - \mathbf{1}\|_2 = (\sqrt{k} - 1)\sqrt{n},$$

$$\rho_2(X_k) = \|X_k^{\mathsf{T}} X_k - I_n\|_{\mathrm{F}}^{q_2} = [(k-1)\sqrt{n}]^{q_2}.$$

By assumption, $\operatorname{dist}(X_k, \mathbb{S}^{n,n}_+) \le \nu\rho_2(X_k)$ for each $k \ge 1$ with a constant ν. Hence we have $q_2 \ge 1/2$. The proof for (b) is complete. □

3.3 Error bounds with $1 \le r \le n$

Now we shift our attention to the general case with $1 \le r \le n$. Given previous bounds for $r \in \{1, n\}$, we are particularly interested in the situation where $1 < r < n$.

We will first prove a local error bound for $\mathbb{S}^{n,r}_+$ relative to $\mathbb{R}^{n\times r}_+$ as detailed in Proposition 3.2. This bound will play a role similar to what observation (3.2.1) and Proposition 3.1 do in the cases of $r = 1$ and $r = n$, respectively. To simplify its proof, we start with the following lemma.

Lemma 3.1. *For any matrix $X \in \mathbb{R}^{n\times r}_+$, there exists a matrix $Y \in \mathbb{R}^{n\times r}_+$ such that $Y^{\mathsf{T}}Y$ is diagonal and*

$$\max\left\{\|x^j - y^j\|_2,\ \left|\|y^j\|_2 - 1\right|\right\} \ \le\ \|z^j\|_1^{\frac{1}{2}} \quad \text{for each} \quad j \in \{1, \ldots, r\}, \tag{3.3.1}$$

where x^j, y^j, and z^j denote the jth column of X, Y, and $Z = X^{\mathsf{T}}X - I_r$, respectively.

Proof. Define l_i $(1 \le i \le n)$ and Y as in the proof of Proposition 3.1. Since $Y^{\mathsf{T}}Y$ is diagonal as mentioned before, it suffices to establish (3.3.1) for this Y.

Fix an index $j \in \{1, \ldots, r\}$. Recalling that $0 \leq X_{i,j} \leq X_{i,l_i}$ for each $i \in \{1, \ldots, n\}$, we have

$$\begin{aligned} \|x^j - y^j\|_2^2 &= \sum_{i=1}^{n} X_{i,j}^2 \mathbb{1}(j \neq l_i) \\ &\leq \sum_{i=1}^{n} X_{i,l_i} X_{i,j} \mathbb{1}(l_i \neq j) \\ &\leq \sum_{l=1}^{r} \left(\sum_{i=1}^{n} X_{i,l} X_{i,j} \right) \mathbb{1}(l \neq j). \end{aligned} \tag{3.3.2}$$

Since $X^\top X$ and Z have the same off-diagonal entries, inequality (3.3.2) yields

$$\|x^j - y^j\|_2^2 \leq \sum_{l=1}^{r} |Z_{l,j}| \mathbb{1}(l \neq j) = \|z^j\|_1 - |Z_{j,j}|. \tag{3.3.3}$$

It remains to prove $\left| \|y^j\|_2 - 1 \right|^2 \leq \|z^j\|_1$. To this end, note that

$$\left| \|y^j\|_2 - 1 \right|^2 \leq \left| \|y^j\|_2^2 - 1 \right| \leq \left| \|x^j\|_2^2 - 1 \right| + \|x^j - y^j\|_2^2, \tag{3.3.4}$$

where the first inequality uses the fact that $|t-1|^2 \leq |t^2 - 1|$ for any $t \geq 0$, and the second one is because $\|x^j\|_2^2 - \|y^j\|_2^2 = \|x^j - y^j\|_2^2$ due to the special construction (3.2.3) of Y. Since $\|x^j\|_2^2 - 1 = Z_{j,j}$, we can combine (3.3.3) and (3.3.4) to obtain

$$\left| \|y^j\|_2 - 1 \right|^2 \leq \left| \|x^j\|_2^2 - 1 \right| + \left(\|z^j\|_1 - |Z_{j,j}| \right) = \|z^j\|_1.$$

The proof is complete. □

Remark 3.2. *As mentioned earlier, the matrix Y in the proof of Lemma 3.1 is the rounding matrix in [14, Procedure 1]. Inequality* (3.3.2) *is essentially the second inequality in Case II of the proof for [14, Lemma 3.1]. The columns of X are assumed to be normalized in [14], but such an assumption has no effect on this inequality.*

Proposition 3.2. *For any matrix $X \in \mathbb{R}_+^{n\times r}$, if $\|\sigma(X) - \mathbf{1}\|_2 < 1/(3\sqrt{r})$, then*

$$\operatorname{dist}(X, \mathbb{S}_+^{n,r}) \;\le\; 2\sqrt{\frac{7r}{3}}\|\sigma(X) - \mathbf{1}\|_2^{\frac{1}{2}}. \tag{3.3.5}$$

Proof. Let Y and Z be the matrices specified in Lemma 3.1. Then (3.3.1) leads to

$$\|X - Y\|_{\mathrm{F}}^2 \;=\; \sum_{j=1}^{r}\|x^j - y^j\|_2^2 \;\le\; \sum_{j=1}^{r}\|z^j\|_1 \;=\; \|Z\|_{\ell_1}. \tag{3.3.6}$$

Since $Y^{\mathsf{T}}Y$ is diagonal, the entries of $\sigma(Y)$ are $\|y^1\|_2, \ldots, \|y^r\|_2$. Thus (3.3.1) also provides

$$\|\sigma(Y) - \mathbf{1}\|_2^2 \;=\; \sum_{j=1}^{r}(\|y^j\|_2 - 1)^2 \;\le\; \sum_{j=1}^{r}\|z^j\|_1 \;=\; \|Z\|_{\ell_1}. \tag{3.3.7}$$

Comparing (3.3.5) with (3.3.6)–(3.3.7), we only need to prove that $\|\sigma(X) - \mathbf{1}\|_2 < 1/(3\sqrt{r})$ ensures

$$\|Z\|_{\ell_1} \;\le\; \frac{7r}{3}\|\sigma(X) - \mathbf{1}\|_2 \tag{3.3.8}$$

and

$$\operatorname{dist}(Y, \mathbb{S}_+^{n,r}) \;=\; \|\sigma(Y) - \mathbf{1}\|_2. \tag{3.3.9}$$

Since $\|Z\|_{\ell_1} = \sum_{i=1}^{n}\sum_{j=1}^{r}|Z_{ij}| \le r\|Z\|_{\mathrm{F}}$, inequality (3.3.8) is a direct consequence of

$$\|Z\|_{\mathrm{F}} \;=\; \|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}} \;\le\; (\|X\|_2 + 1)\|\sigma(X) - \mathbf{1}\|_2 \;\le\; \frac{7}{3}\|\sigma(X) - \mathbf{1}\|_2, \tag{3.3.10}$$

where the last inequality is because $\|X\|_2 \le \|\sigma(X) - \mathbf{1}\|_2 + 1 < 4/3$. Meanwhile, inequality (3.3.10) also leads to

$$\|z^j\|_1 \;\le\; \sqrt{r}\|Z\|_{\mathrm{F}} \;\le\; \frac{7\sqrt{r}}{3}\|\sigma(X) - \mathbf{1}\|_2 \;<\; 1 \quad \text{for each} \quad j \in \{1, \ldots, r\}.$$

Therefore, inequality (3.3.1) implies that Y does not contain any zero column. Hence the diagonal entries of $Y^{\mathsf{T}}Y$ are all positive, which ensures the nonsingularity of this matrix since it is diagonal. Thus Lemma 2.5 yields (3.3.9). The proof is complete. □

Remark 3.3. *In [27, Lemma 3.2], by choosing* $\bar{X} = \begin{bmatrix} I_r & \mathbf{0}_{r\times(n-r)} \end{bmatrix}^{\mathsf{T}}$ *and special* δ *such that* $\|\sigma(X) - \mathbf{1}\|_2 < 1/(3\sqrt{r})$ *for* $X \in \mathcal{B}(\bar{X}, \delta)$ *one can derive the same result in Proposition 3.2.*

Now we are ready to establish a local error bound for $\mathbb{S}_+^{n,r}$ relative to $\mathbb{R}^{n\times r}$.

Theorem 3.4. *For any matrix* $X \in \mathbb{R}^{n\times r}$*, if* $\|X_-\|_{\mathrm{F}} + \|\sigma(X) - \mathbf{1}\|_2 < 1/(3\sqrt{r})$*, then*

$$\mathrm{dist}(X,\, \mathbb{S}_+^{n,r}) \;\le\; 4\sqrt{r}\left(\|X_-\|_{\mathrm{F}}^{\frac{1}{2}} + \|\sigma(X) - \mathbf{1}\|_2^{\frac{1}{2}}\right). \tag{3.3.11}$$

Proof. According to Lemma 2.2,

$$\|\sigma(X_+) - \mathbf{1}\|_2 \;\le\; \|X_-\|_{\mathrm{F}} + \|\sigma(X) - \mathbf{1}\|_2.$$

Thus $\|\sigma(X_+) - \mathbf{1}\|_2 < 1/(3\sqrt{r})$ by assumption, and hence Proposition 3.2 implies

$$\mathrm{dist}(X_+,\, \mathbb{S}_+^{n,r}) \;\le 2\sqrt{\frac{7r}{3}}\left(\|X_-\|_{\mathrm{F}}^{\frac{1}{2}} + \|\sigma(X) - \mathbf{1}\|_2^{\frac{1}{2}}\right). \tag{3.3.12}$$

On the other hand, since $\|X_-\|_{\mathrm{F}} < 1/(3\sqrt{r})$, it holds that

$$\|X - X_+\|_{\mathrm{F}} \;=\; \|X_-\|_{\mathrm{F}} \;\le\; \frac{1}{\sqrt{3}r^{\frac{1}{4}}}\|X_-\|_{\mathrm{F}}^{\frac{1}{2}} \;\le\; \sqrt{\frac{r}{3}}\|X_-\|_{\mathrm{F}}^{\frac{1}{2}}. \tag{3.3.13}$$

Inequality (3.3.11) follows from (3.3.12) and (3.3.13) because $2\sqrt{7/3}+1/\sqrt{3} < 4$. □

Theorem 3.4 presents only a local error bound. Indeed, $\|X_-\|_{\mathrm{F}}^{\frac{1}{2}} + \|\sigma(X) - \mathbf{1}\|_2^{\frac{1}{2}}$ does not define a global error bound for $\mathbb{S}_+^{n,r}$ relative to $\mathbb{R}^{n\times r}$, which will be explained later by Theorem 3.6. To have a global error bound, we need to replace the term $\|\sigma(X) - \mathbf{1}\|_2$ with $\|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}}$ as in the following theorem.

Theorem 3.5. *For any matrix* $X \in \mathbb{R}^{n\times r}$*, we have*

$$\mathrm{dist}(X,\, \mathbb{S}_+^{n,r}) \;\le\; 5r^{\frac{3}{4}}\left(\|X_-\|_{\mathrm{F}}^{\frac{1}{2}} + \|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}}^{\frac{1}{2}}\right). \tag{3.3.14}$$

Moreover, if $\|X_-\|_{\mathrm{F}} + \|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}} < 1/(3\sqrt{r})$*, then*

$$\operatorname{dist}(X, \mathbb{S}_+^{n,r}) \;\le\; 4\sqrt{r}\left(\|X_-\|_{\mathrm{F}}^{\frac{1}{2}} + \|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}}^{\frac{1}{2}}\right). \tag{3.3.15}$$

Proof. Recall that $\|\sigma(X) - \mathbf{1}\|_2 \le \|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}}$ (Lemma 2.4). Thus (3.3.15) is a direct consequence of Theorem 3.4 when $\|X_-\|_{\mathrm{F}} + \|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}} < 1/(3\sqrt{r})$.

Now we prove (3.3.14). Let us assume that

$$\|X_-\|_{\mathrm{F}} + \|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}} \;\ge\; \frac{1}{3\sqrt{r}},$$

as (3.3.14) is already justified by (3.3.15) when this inequality does not hold. Under this assumption,

$$\|X_-\|_{\mathrm{F}}^{\frac{1}{2}} + \|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}}^{\frac{1}{2}} \;\ge\; \frac{1}{\sqrt{3}r^{\frac{1}{4}}}. \tag{3.3.16}$$

Noting that the diameter of $\mathbb{S}^{n,r}$ is $2\sqrt{r}$, we then have

$$\begin{aligned}
\operatorname{dist}(X, \mathbb{S}_+^{n,r}) \;&\le\; \operatorname{dist}(X, \mathbb{S}^{n,r}) + 2\sqrt{r} \\
&\le r^{\frac{1}{4}}\|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}}^{\frac{1}{2}} + 2\sqrt{3}r^{\frac{3}{4}}\left(\|X_-\|_{\mathrm{F}}^{\frac{1}{2}} + \|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}}^{\frac{1}{2}}\right) \\
&\le 5r^{\frac{3}{4}}\left(\|X_-\|_{\mathrm{F}}^{\frac{1}{2}} + \|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}}^{\frac{1}{2}}\right),
\end{aligned} \tag{3.3.17}$$

where the second inequality applies Lemma 2.4 and (3.3.16). □

Recently, Theorem 3.5 has been used in [19, 20] to establish error bounds for $\operatorname{dist}(X, \mathbb{S}_+^{n,r})$ for X in the unit ball of spectral norm, i.e., $\{X \in \mathbb{R}^{n\times r} : \|X\| \le 1\}$. See (31) in [19].

3.4 Tightness of the error bounds when $1 < r < n$

The following proposition shows that the bounds presented in Theorems 3.4 and 3.5 are tight up to multiplicative constants when $1 < r < n$, no matter whether X

belongs to $\mathbb{S}^{n,r}$, $\mathbb{R}^{n\times r}_+$, or neither of them. Consequently, the error bounds (1.1.2)–(1.1.4) cannot hold with $q > 1/2$ when $1 < r < n$.

Proposition 3.3. *Suppose that* $1 < r < n$.

(a) *There exists a sequence* $\{X_k\} \subset \mathbb{S}^{n,r} \setminus \mathbb{R}^{n\times r}_+$ *such that* $(X_k)_- \to 0$ *and*

$$\operatorname{dist}(X_k, \mathbb{S}^{n,r}_+) \;\geq\; \frac{1}{\sqrt{2}}\|(X_k)_-\|_{\mathrm{F}}^{\frac{1}{2}}. \tag{3.4.1}$$

(b) *There exists a sequence* $\{X_k\} \subset \mathbb{R}^{n\times r}_+ \setminus \mathbb{S}^{n,r}$ *such that* $X_k^{\mathsf{T}}X_k \to I_r$ *and*

$$\operatorname{dist}(X_k, \mathbb{S}^{n,r}_+) \;\geq\; \frac{1}{\sqrt{2}}\|X_k^{\mathsf{T}}X_k - I_r\|_{\mathrm{F}}^{\frac{1}{2}}. \tag{3.4.2}$$

(c) *There exists a sequence* $\{X_k\} \subset \mathbb{R}^{n\times r} \setminus (\mathbb{R}^{n\times r}_+ \cup \mathbb{S}^{n,r})$ *such that* $(X_k)_- \to 0$, $X_k^{\mathsf{T}}X_k \to I_r$, *and*

$$\operatorname{dist}(X_k, \mathbb{S}^{n,r}_+) \;\geq\; \frac{1}{\sqrt{2}+1}\left(\|(X_k)_-\|_{\mathrm{F}}^{\frac{1}{2}} + \|X_k^{\mathsf{T}}X_k - I_r\|_{\mathrm{F}}^{\frac{1}{2}}\right). \tag{3.4.3}$$

Proof. Take a sequence $\{\varepsilon_k\} \subset (0, 1/2)$ that converges to 0. For each $k \geq 1$, let $X_k \in \mathbb{R}^{n\times r}$ be a matrix such that its first 3 rows are

$$\begin{bmatrix} \varepsilon_k & \varepsilon_k & \overbrace{0 & \dots & 0}^{r-2} \\ a_k & b_k & 0 & \dots & 0 \\ c_k & d_k & 0 & \dots & 0 \end{bmatrix}$$

with a_k, b_k, c_k, d_k being specified later, its 4th to $(r+1)$th rows are the last $r-2$ rows of I_r (if $r \geq 3$), and its other rows are zero (if any). In addition, let $\bar{X}_k$ be a projection of X_k onto $\mathbb{S}^{n,r}_+$. Then the first row of $\bar{X}_k$ contains at most one nonzero entry according to Lemma 2.6. Hence

$$\operatorname{dist}(X_k, \mathbb{S}^{n,r}_+) \;=\; \|X_k - \bar{X}_k\|_{\mathrm{F}} \;\geq\; \varepsilon_k. \tag{3.4.4}$$

Moreover, it is clear that $(X_k)_- \to 0$ and $X_k^\mathsf{T} X_k \to I_r$ if

$$a_k \to 1, \quad b_k \to 0, \quad c_k \to 0, \quad \text{and} \quad d_k \to 1. \tag{3.4.5}$$

In the sequel, we will configure a_k, b_k, c_k, and d_k subject to (3.4.5) so that $\{X_k\}$ validates (a), (b), and (c) one by one.

(a) Define

$$a_k \;=\; \sqrt{1-\varepsilon_k^2}, \quad b_k = -\frac{\varepsilon_k^2}{a_k}, \quad c_k = 0, \quad \text{and} \quad d_k = \sqrt{1-\varepsilon_k^2-b_k^2}.$$

Then $X_k \in \mathbb{S}^{n,r} \setminus \mathbb{R}_+^{n\times r}$. Clearly, $\|(X_k)_-\|_{\mathrm{F}} = \varepsilon_k^2/a_k$. Hence (3.4.1) holds according to (3.4.4) and the fact that $a_k \geq \sqrt{1-\varepsilon_k^2} > 1/2$ (recall that $\varepsilon_k < 1/2$).

(b) Define $a_k = d_k = 1$ and $b_k = c_k = 0$. Then $X_k \in \mathbb{R}_+^{n\times r} \setminus \mathbb{S}^{n,r}$. By straightforward calculations,

$$\|X_k^\mathsf{T} X_k - I_r\|_{\mathrm{F}} \;=\; 2\varepsilon_k^2.$$

Thus (3.4.2) holds according to (3.4.4).

(c) Define $a_k = d_k = 1$, $b_k = -\varepsilon_k^2$, and $c_k = 0$. Then $X_k \in \mathbb{R}^{n\times r} \setminus (\mathbb{R}_+^{n\times r} \cup \mathbb{S}^{n,r})$. In addition, we can calculate that

$$\|X_k^\mathsf{T} X_k - I_r\|_{\mathrm{F}} \;=\; \sqrt{\varepsilon_k^4 + (\varepsilon_k^2+\varepsilon_k^4)^2} \;\leq\; \sqrt{\varepsilon_k^4 + \left(\varepsilon_k^2 + \frac{\varepsilon_k^2}{4}\right)^2} \;\leq\; 2\varepsilon_k^2$$

and $\|(X_k)_-\|_{\mathrm{F}} = \varepsilon_k^2$. Therefore, (3.4.3) holds according to (3.4.4). □

Theorem 3.6 extends Theorems 3.4 and 3.5, allowing $\|X_-\|_{\mathrm{F}}$ and $\|\sigma(X) - \mathbf{1}\|_2$ or $\|X^\mathsf{T} X - I_r\|_{\mathrm{F}}$ to have different exponents in the error bounds. It specifies the precise range of these exponents in local and global error bounds for $\mathbb{S}_+^{n,r}$ relative to $\mathbb{R}^{n\times r}$ when $1 < r < n$. As we will see from (b) of this theorem and its proof, when $1 < r < n$, the error bound (1.1.2) can hold if and only if $q \leq 1/2$, whereas (1.1.3) and (1.1.4) can hold if and only if $q = 1/2$.

Theorem 3.6. *Let q_1 and q_2 be positive constants. Suppose that $1 < r < n$.*

(a) *The function $\rho_1(X) := \|X_-\|_{\mathrm{F}}^{q_1} + \|\sigma(X) - \mathbf{1}\|_2^{q_2}$ defines a local error bound for $\mathbb{S}_+^{n,r}$ relative to $\mathbb{R}^{n\times r}$ if and only if $q_1 \le 1/2$ and $q_2 \le 1/2$, but it cannot define a global error bound no matter what values q_1 and q_2 take.*

(b) *The function $\rho_2(X) := \|X_-\|_{\mathrm{F}}^{q_1} + \|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}}^{q_2}$ defines a local error bound for $\mathbb{S}_+^{n,r}$ relative to $\mathbb{R}^{n\times r}$ if and only if $q_1 \le 1/2$ and $q_2 \le 1/2$, and it defines a global error bound if and only if $q_1 \le q_2 = 1/2$.*

Proof. (a) Based on (3.3.11), it is easy to check that ρ_1 defines a local error bound for $\mathbb{S}_+^{n,r}$ relative to $\mathbb{R}^{n\times r}$ if $q_1 \le 1/2$ and $q_2 \le 1/2$. Conversely, if ρ_1 defines a local error bound for $\mathbb{S}_+^{n,r}$ relative to $\mathbb{R}^{n\times r}$, then $q_1 \le 1/2$ and $q_2 \le 1/2$ according to (a) and (b) of Proposition 3.3, respectively.

Now we prove that ρ_1 cannot define a global error bound. According to what has been shown above, we assume that $q_2 \le 1/2$, as a global error bound must be a local one. Consider a sequence $\{X_k\} \subset \mathbb{R}_+^{n\times r}$ with $\|X_k\|_{\mathrm{F}} \to \infty$. Then $\rho_2(X_k) = \|\sigma(X_k) - \mathbf{1}\|_2^{q_2}$, and hence

$$\frac{\mathrm{dist}(X_k,\, \mathbb{S}_+^{n,r})}{\rho_1(X_k)} \;\ge\; \frac{\|\sigma(X_k) - \mathbf{1}\|_2}{\|\sigma(X_k) - \mathbf{1}\|_2^{q_2}} \;\to\; \infty.$$

Thus ρ_1 cannot define a global error bound for $\mathbb{S}_+^{n,r}$ relative to $\mathbb{R}^{n\times r}$.

(b) Similar to (a), we can show that ρ_2 defines a local error bound for $\mathbb{S}_+^{n,r}$ relative to $\mathbb{R}^{n\times r}$ if and only if $q_1 \le 1/2$ and $q_2 \le 1/2$. Hence we only need to consider the global error bound.

Suppose that $q_1 \le q_2 = 1/2$. Let us show that

$$\mathrm{dist}(X,\, \mathbb{S}_+^{n,r}) \;\le\; 5r^{\frac{3}{4}} \left(\|X_-\|_{\mathrm{F}}^{q_1} + \|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}}^{\frac{1}{2}}\right) \;=\; 5r^{\frac{3}{4}}\rho_2(X) \tag{3.4.6}$$

for all $X \in \mathbb{R}^{n\times r}$. If $\|X_-\|_{\mathrm{F}} \le 1$, then (3.4.6) follows from (3.3.14). When $\|X_-\|_{\mathrm{F}} > 1$,

$$\mathrm{dist}(X,\, \mathbb{S}_+^{n,r}) \;\le\; \mathrm{dist}(X,\, \mathbb{S}^{n,r}) + 2\sqrt{r} \;\le\; r^{\frac{1}{4}}\|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}}^{\frac{1}{2}} + 2\sqrt{r}\|X_-\|_{\mathrm{F}}^{q_1} \;\le\; 5r^{\frac{3}{4}}\rho_2(X),$$

where the second inequality applies Lemma 2.4. Hence ρ_2 defines a global error bound for $\mathbb{S}^{n,r}_+$ relative to $\mathbb{R}^{n\times r}$.

Now suppose that ρ_2 defines a global error bound for $\mathbb{S}^{n,r}_+$ relative to $\mathbb{R}^{n\times r}$. Then it defines a local error bound, implying $q_1 \le 1/2$ and $q_2 \le 1/2$. Similar to the proof for (b) of Theorem 3.3, by considering a sequence $\{X_k\} \subset \mathbb{R}^{n\times r}_+$ such that $X_k^{\mathsf{T}} X_k = kI_r$ for each $k \ge 1$, we can prove $q_2 \ge 1/2$. The proof is complete. □

Even though the function ρ_1 in Theorem 3.6 can only define a local error bound for $\mathbb{S}^{n,r}_+$ relative to $\mathbb{R}^{n\times r}$, global error bounds can still be established if we add a suitable power of $\|\sigma(X) - \mathbf{1}\|_2$ or $\|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}}$ to ρ_1. This will be detailed in Remark 3.4 after we prove the following proposition.

Proposition 3.4. *Let ϕ_1 and ϕ_2 be two nonnegative functions on $\mathbb{R}^{n\times r}$. If there exist positive constants γ_1, γ_2, c_1 and c_2 such that*

$$\mathrm{dist}(X, \mathbb{S}^{n,r}_+) \;\le\; \gamma_1\phi_1(X) \quad \textit{when} \quad \phi_1(X) \le c_1, \tag{3.4.7}$$

$$\mathrm{dist}(X, \mathbb{S}^{n,r}) \;\le\; \gamma_2\phi_2(X) \quad \textit{when} \quad \mathrm{dist}(X, \mathbb{S}^{n,r}) \ge c_2. \tag{3.4.8}$$

Then $\mathrm{dist}(X, \mathbb{S}^{n,r}_+) \le \max\{\gamma_1, \gamma_2, c_1^{-1}(2\sqrt{r} + c_2)\}[\phi_1(X) + \phi_2(X)]$ *for all* $X \in \mathbb{R}^{n\times r}$.

Proof. Fix an $X \in \mathbb{R}^{n\times r}$. We only consider the situation where $\phi_1(X) > c_1$, due to (3.4.7). Note that

$$\mathrm{dist}(X, \mathbb{S}^{n,r}_+) \;\le\; 2\sqrt{r} + \mathrm{dist}(X, \mathbb{S}^{n,r}). \tag{3.4.9}$$

If $\mathrm{dist}(X, \mathbb{S}^{n,r}) < c_2$, then (3.4.9) implies that

$$\mathrm{dist}(X, \mathbb{S}^{n,r}_+) \;\le\; c_1^{-1}(2\sqrt{r} + c_2)\phi_1(X).$$

If $\mathrm{dist}(X, \mathbb{S}^{n,r}) \ge c_2$, then (3.4.8) and (3.4.9) imply that

$$\mathrm{dist}(X, \mathbb{S}^{n,r}_+) \;\le\; \max\{2c_1^{-1}\sqrt{r}, \gamma_2\}[\phi_1(X) + \phi_2(X)].$$

The proof is complete. □

Remark 3.4. *Suppose that* $1 < r < n$, $0 < q_1 \leq 1/2$, *and* $0 < q_2 \leq 1/2$. *According to Theorem 3.6, Proposition 3.4, and Lemma 2.4,* $\rho_1(X) + \|\sigma(X) - \mathbf{1}\|_2^q$ *with* $q \geq 1$ *defines a global error bound for* $\mathbb{S}_+^{n,r}$ *relative to* $\mathbb{R}^{n\times r}$. *So does* $\rho_1(X) + \|X^\mathsf{T}X - I_r\|_{\mathrm{F}}^q$ *with* $q \geq 1/2$. *However, the powers in* ρ_1 *cannot be greater than* $1/2$ *even with the additional terms for the global error bounds. The same can be said about* ρ_2.

3.5 Linear regularity of $\mathbb{R}_+^{n\times r}$ and $\mathbb{S}^{n,r}$

Before ending this section, we briefly mention that our analysis enables us to characterize the linear regularity of $\mathbb{R}_+^{n\times r}$ and $\mathbb{S}^{n,r}$ for $r \in \{1, \ldots, n\}$.

A pair of sets $\mathcal{A}_1$ and $\mathcal{A}_2$ in $\mathbb{R}^{n\times r}$ with $\mathcal{A}_1 \cap \mathcal{A}_2 \neq \emptyset$ are said to be boundedly linearly regular if for any bounded set $\mathcal{T} \subset \mathbb{R}^{n\times r}$ there exists a constant γ such that

$$\operatorname{dist}(X,\, \mathcal{A}_1 \cap \mathcal{A}_2) \;\leq\; \gamma \max\{\operatorname{dist}(X,\, \mathcal{A}_1),\, \operatorname{dist}(X,\, \mathcal{A}_2)\} \tag{3.5.1}$$

for all $X \in \mathcal{T}$, and they are linearly regular if (3.5.1) holds for all $X \in \mathbb{R}^{n\times r}$. Linear regularity is a fundamental concept in optimization and is closely related to error bounds. See [24] and [7, Section 8.5] for more details. Note that we can replace the maximum in (3.5.1) with a summation without essentially changing the definition of (boundedly) linear regularity.

Proposition 3.5 clarifies whether $\mathbb{R}_+^{n\times r}$ and $\mathbb{S}^{n,r}$ are linearly regular.

Proposition 3.5. *The two sets* $\mathbb{R}_+^{n\times r}$ *and* $\mathbb{S}^{n,r}$ *are linearly regular if and only if* $r = 1$ *or* $r = n$.

Proof. Recall that $\operatorname{dist}(X, \mathbb{R}_+^{n\times r}) = \|X_-\|_{\mathrm{F}}$ and $\operatorname{dist}(X, \mathbb{S}^{n,r}) = \|\sigma(X) - \mathbf{1}\|_2$ for $X \in \mathbb{R}^{n\times r}$. The "if" part of this proposition holds because of the global error bounds in Theorems 3.1 and 3.2. The "only if" part holds because $\|X_-\|_{\mathrm{F}} + \|\sigma(X) - \mathbf{1}\|_2$ does not define a global error bound for $\mathbb{S}_+^{n,r}$ relative to $\mathbb{R}^{n\times r}$ when $1 < r < n$, as we can see from (a) of Theorem 3.6. □

Proposition 3.5 remains true if we change "linearly regular" to "boundedly linearly regular". The "if" part is weakened after this change, and the other part holds because $\|X_-\|_{\mathrm{F}} + \|\sigma(X) - \mathbf{1}\|_2$ does not define a local error bound for $\mathbb{S}_+^{n,r}$ relative to $\mathbb{R}^{n\times r}$ when $1 < r < n$ according to (a) of Theorem 3.6.

3.6 A special case of error bounds

To derive the error bounds for $\mathbb{S}_S^{n,r}$, we first consider the special case with

$$\mathcal{P} = \{1, \ldots, r_1\} \ \text{ and } \ \mathcal{N} = \emptyset,$$

where $r_1 \in \{1, \ldots, r\}$. Define $r_2 = r - r_1$ henceforth. In this case, $\mathbb{S}_S^{n,r}$ reduces to

$$\mathbb{S}_{r_1,+}^{n,r} \ := \ \left\{X = (X_1, X_2) \mid X_1 \in \mathbb{R}_+^{n\times r_1},\ X_2 \in \mathbb{R}^{n\times r_2},\ X^{\mathsf{T}}X = I_r\right\}, \tag{3.6.1}$$

with $\mathbb{S}_{r_1,+}^{n,r}$ being $\mathbb{S}_+^{n,r}$ if $r_1 = r$.

Note that the results established in Chapters 2 and 3 are still valid when r is replaced with r_1 or r_2. In the sequel, we will apply these results directly without restating this fact.

Lemma 3.2. *Suppose that $r_1 < r$. Consider matrices $Y_1 \in \mathbb{R}^{n\times r_1}$ and $Y_2 \in \mathbb{R}^{n\times r_2}$. If $Y_1^{\mathsf{T}}Y_2 = 0$, then there exists a matrix Z that is a projection of Y_2 onto $\mathbb{S}^{n,r_2}$ and satisfies $Y_1^{\mathsf{T}}Z = 0$.*

Proof. Define $k = n - \mathrm{rank}(Y_1)$. Take a matrix $V \in \mathbb{S}^{n,k}$ such that $\mathrm{range}(V)$ is the orthogonal complement of $\mathrm{range}(Y_1)$ in $\mathbb{R}^n$. Since $k \geq r - r_1 = r_2$, the matrix $V^{\mathsf{T}}Y_2 \in \mathbb{R}^{k\times r_2}$ has a polar decomposition UP with $U \in \mathbb{S}^{k,r_2}$ and $P \in \mathbb{R}^{r_2\times r_2}$, the latter being positive semidefinite. Define $Z = VU \in \mathbb{R}^{n\times r_2}$. Then

$$ZP = VUP = VV^{\mathsf{T}}Y_2 = Y_2,$$

where the last equality holds because $\mathrm{range}(Y_2) \subset \mathrm{range}(V)$ according to $Y_1^{\mathsf{T}}Y_2 = 0$, and VV^{T} is the orthogonal projection onto $\mathrm{range}(V)$. Besides, $Z^{\mathsf{T}}Z = U^{\mathsf{T}}V^{\mathsf{T}}VU =$

I_{r_2}. Thus ZP is a polar decomposition of Y_2. Hence Z is a projection of Y_2 onto $\mathbb{S}^{n,r_2}$ by Lemma 2.3. Moreover, $Y_1^{\mathsf{T}} Z = Y_1^{\mathsf{T}} VU = 0$. □

Note that $\mathbb{S}^{n,r}_{r_1,+}$ can also be formulated as

$$\mathbb{S}^{n,r}_{r_1,+} \;=\; \left\{(X_1, X_2) \mid X_1 \in \mathbb{S}^{n,r_1}_{+},\; X_2 \in \mathbb{S}^{n,r_2},\; X_1^{\mathsf{T}} X_2 = 0\right\}.$$

This formulation motivates us to develop the following lemma, which provides a global error bound for $\mathbb{S}^{n,r}_{r_1,+}$ relative to $\mathbb{R}^{n\times r}$.

Lemma 3.3. *Suppose that $r_1 < r$. For any matrix $X = (X_1, X_2)$ with $X_1 \in \mathbb{R}^{n\times r_1}$ and $X_2 \in \mathbb{R}^{n\times r_2}$, we have*

$$\mathrm{dist}(X,\, \mathbb{S}^{n,r}_{r_1,+}) \;\le\; (2\|X_2\|_2+1)\,\mathrm{dist}(X_1,\, \mathbb{S}^{n,r_1}_{+}) + \mathrm{dist}(X_2,\, \mathbb{S}^{n,r_2}) + 2\|X_1^{\mathsf{T}} X_2\|_{\mathrm{F}}. \quad (3.6.2)$$

Proof. Let Y_1 be a projection of X_1 onto $\mathbb{S}^{n,r_1}_{+}$ and $Y_2 = (I_n - Y_1 Y_1^{\mathsf{T}})X_2 \in \mathbb{R}^{n\times r_2}$. Then $Y_1^{\mathsf{T}} Y_2 = 0$. By Lemma 3.2, there exists a matrix Z that is a projection of Y_2 onto $\mathbb{S}^{n,r_2}$ with $Y_1^{\mathsf{T}} Z = 0$. Define $\bar{X} = (Y_1, Z)$, which lies in $\mathbb{S}^{n,r}_{r_1,+}$. Let us estimate $\|X - \bar{X}\|_{\mathrm{F}}$. It is clear that

$$\begin{aligned}\|X - \bar{X}\|_{\mathrm{F}} &\le \|(X_1, X_2) - (Y_1, Y_2)\|_{\mathrm{F}} + \|(Y_1, Y_2) - (Y_1, Z)\|_{\mathrm{F}} \\ &\le \|X_1 - Y_1\|_{\mathrm{F}} + \|X_2 - Y_2\|_{\mathrm{F}} + \|Y_2 - Z\|_{\mathrm{F}}.\end{aligned}$$

Since $\|Y_2 - Z\|_{\mathrm{F}} = \|\sigma(Y_2) - \mathbf{1}\|_2$ (Lemma 2.4) and $\|\sigma(X_2) - \sigma(Y_2)\|_2 \le \|X_2 - Y_2\|_{\mathrm{F}}$ (Lemma 2.2), it holds that $\|Y_2 - Z\|_{\mathrm{F}} \le \|\sigma(X_2) - \mathbf{1}\|_2 + \|X_2 - Y_2\|_{\mathrm{F}}$. Therefore,

$$\|X - \bar{X}\|_{\mathrm{F}} \;\le\; \|X_1 - Y_1\|_{\mathrm{F}} + \|\sigma(X_2) - \mathbf{1}\|_2 + 2\|X_2 - Y_2\|_{\mathrm{F}}. \quad (3.6.3)$$

Meanwhile, recalling that $Y_2 = (I_n - Y_1 Y_1^{\mathsf{T}})X_2$ and $Y_1 \in \mathbb{S}^{n,r_1}$, we have

$$\|X_2 - Y_2\|_{\mathrm{F}} = \|Y_1 Y_1^{\mathsf{T}} X_2\|_{\mathrm{F}} \;=\; \|Y_1^{\mathsf{T}} X_2\|_{\mathrm{F}} \;\le\; \|(Y_1 - X_1)^{\mathsf{T}} X_2\|_{\mathrm{F}} + \|X_1^{\mathsf{T}} X_2\|_{\mathrm{F}}. \quad (3.6.4)$$

Plugging (3.6.4) into (3.6.3) while noting $\|(Y_1 - X_1)^{\mathsf{T}} X_2\|_{\mathrm{F}} \le \|X_1 - Y_1\|_{\mathrm{F}} \|X_2\|_2$, we obtain

$$\|X - \bar{X}\|_{\mathrm{F}} \;\le\; (2\|X_2\|_2 + 1)\|X_1 - Y_1\|_{\mathrm{F}} + \|\sigma(X_2) - \mathbf{1}\|_2 + 2\|X_1^{\mathsf{T}} X_2\|_{\mathrm{F}}.$$

This implies (3.6.2), because $\|X_1 - Y_1\|_{\mathrm{F}} = \mathrm{dist}(X_1, \mathbb{S}_+^{n,r_1})$ by the definition of Y_1, and $\|\sigma(X_2) - \mathbf{1}\|_2 = \mathrm{dist}(X_2, \mathbb{S}^{n,r_2})$ by Lemma 2.4. □

In light of Lemma 3.3, we can establish error bounds for $\mathbb{S}_{r_1,+}^{n,r}$ using those for $\mathbb{S}_+^{n,r}$, as will be done in Propositions 3.6 and 3.7. To this end, it is useful to note for any matrix $X = (X_1, X_2)$ that

$$\|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}} \;\geq\; \max\left\{\|X_1^{\mathsf{T}}X_1 - I_{r_1}\|_{\mathrm{F}},\; \|X_2^{\mathsf{T}}X_2 - I_{r_2}\|_{\mathrm{F}},\; \sqrt{2}\|X_1^{\mathsf{T}}X_2\|_{\mathrm{F}}\right\}. \quad (3.6.5)$$

Proposition 3.6. *For any matrix $X \in \mathbb{R}^{n\times r}$ with x_1 being its first column, we have*

$$\mathrm{dist}(X, \mathbb{S}_{1,+}^{n,r}) \;\leq\; 7\sqrt{r}\left(\|(x_1)_-\|_2 + \|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}}\right). \quad (3.6.6)$$

Moreover, if $\|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}} < 1/3$, then

$$\mathrm{dist}(X, \mathbb{S}_{1,+}^{n,r}) \;\leq\; 7\left(\|(x_1)_-\|_2 + \|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}}\right). \quad (3.6.7)$$

Proof. If $r = 1$, then (3.6.6) and (3.6.7) hold because of Theorem 3.1. Hence we suppose that $r > 1$ in the sequel. We first assume $\|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}} < 1/3$ and establish (3.6.7). Let X_2 be the matrix containing the last $r-1$ columns of X. According to Theorem 3.1 and Lemma 2.4,

$$\mathrm{dist}(x_1, \mathbb{S}_+^{n,1}) \;\leq\; 2\|(x_1)_-\|_2 + \left|x_1^{\mathsf{T}}x_1 - 1\right|, \quad (3.6.8)$$

$$\mathrm{dist}(X_2, \mathbb{S}^{n,r-1}) \;\leq\; \|X_2^{\mathsf{T}}X_2 - I_{r-1}\|_{\mathrm{F}}. \quad (3.6.9)$$

Plugging (3.6.8) and (3.6.9) into Lemma 3.3 while noting (3.6.5), we have

$$\begin{aligned}\mathrm{dist}(X, \mathbb{S}_{1,+}^{n,r}) \;&\leq\; (2\|X_2\|_2 + 1)\cdot 2\|(x_1)_-\|_2 + \left[(2\|X_2\|_2 + 1) + 1 + \sqrt{2}\right]\|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}} \\ &\leq\; 7\left(\|(x_1)_-\|_2 + \|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}}\right),\end{aligned}$$

where the second inequality uses the fact that $\|X_2\|_2^2 \leq \|X^{\mathsf{T}}X - I_r\|_2 + 1 \leq 4/3$.

To prove (3.6.6), we now only need to focus on the case with $\|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}} \geq 1/3$. In this case,

$$\mathrm{dist}(X, \mathbb{S}_{1,+}^{n,r}) \;\leq\; \mathrm{dist}(X, \mathbb{S}^{n,r}) + 2\sqrt{r} \;\leq\; \|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}} + 6\sqrt{r}\|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}},$$

which implies (3.6.6). The proof is complete. □

Proposition 3.7. *For any matrix $X \in \mathbb{R}^{n\times r}$ with X_1 being its submatrix containing the first r_1 columns, we have*

$$\operatorname{dist}(X, \mathbb{S}^{n,r}_{r_1,+}) \;\le\; 15r^{\frac{3}{4}} \left(\|(X_1)_-\|_{\mathrm{F}}^{\frac{1}{2}} + \|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}}^{\frac{1}{2}} \right). \tag{3.6.10}$$

Moreover, if $\|(X_1)_-\|_{\mathrm{F}} + \|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}} < 1/(3\sqrt{r})$, then

$$\operatorname{dist}(X, \mathbb{S}^{n,r}_{r_1,+}) \;\le\; 15\sqrt{r} \left(\|(X_1)_-\|_{\mathrm{F}}^{\frac{1}{2}} + \|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}}^{\frac{1}{2}} \right). \tag{3.6.11}$$

Proof. If $r_1 = r$, then (3.6.10) and (3.6.11) hold because of Theorem 3.5. Hence we suppose that $r_1 < r$ in the sequel. We first assume $\|(X_1)_-\|_{\mathrm{F}} + \|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}} < 1/(3\sqrt{r})$ and establish (3.6.11). Let X_2 be the matrix containing the last $r_2 = r - r_1$ columns of X. According to (3.6.5), our assumption implies

$$\|(X_1)_-\|_{\mathrm{F}} + \|X_1^{\mathsf{T}}X_1 - I_{r_1}\|_{\mathrm{F}} \;<\; \frac{1}{3\sqrt{r_1}}, \quad \|X_2^{\mathsf{T}}X_2 - I_{r_2}\|_{\mathrm{F}} \;\le\; \frac{1}{3}.$$

Hence Theorem 3.5 and Lemma 2.4 yield

$$\operatorname{dist}(X_1, \mathbb{S}^{n,r_1}_{+}) \;\le\; 4\sqrt{r_1} \left(\|(X_1)_-\|_{\mathrm{F}}^{\frac{1}{2}} + \|X_1^{\mathsf{T}}X_1 - I_{r_1}\|_{\mathrm{F}}^{\frac{1}{2}} \right), \tag{3.6.12}$$

$$\operatorname{dist}(X_2, \mathbb{S}^{n,r_2}) \;\le\; \|X_2^{\mathsf{T}}X_2 - I_{r_2}\|_{\mathrm{F}} \;\le\; \frac{1}{\sqrt{3}}\|X_2^{\mathsf{T}}X_2 - I_{r_2}\|_{\mathrm{F}}^{\frac{1}{2}}. \tag{3.6.13}$$

In addition, inequality (3.6.5) and our assumption also provide

$$\|X_1^{\mathsf{T}}X_2\|_{\mathrm{F}} \;\le\; \frac{1}{\sqrt{2}}\|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}} \;\le\; \frac{1}{\sqrt{6}}\|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}}^{\frac{1}{2}}. \tag{3.6.14}$$

Plugging (3.6.12)–(3.6.14) into Lemma 3.3 while noting (3.6.5), we obtain

$$\begin{aligned}\operatorname{dist}(X, \mathbb{S}^{n,r}_{r_1,+}) \;&\le\; \left[4\sqrt{r_1}(2\|X_2\|_2 + 1) + \frac{1}{\sqrt{3}} + \frac{2}{\sqrt{6}}\right] \left(\|(X_1)_-\|_{\mathrm{F}}^{\frac{1}{2}} + \|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}}^{\frac{1}{2}} \right) \\ &\le 15\sqrt{r} \left(\|(X_1)_-\|_{\mathrm{F}}^{\frac{1}{2}} + \|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}}^{\frac{1}{2}} \right),\end{aligned}$$

where the second inequality uses the fact that $\|X_2\|_2^2 \le \|X^\mathsf{T}X - I_r\|_2 + 1 \le 4/3$.

Now we prove (3.6.10). By the same technique as the proof of (3.3.17), we have

$$\operatorname{dist}(X,\, \mathbb{S}^{n,r}_{r_1,+}) \;\le\; 5r^{\frac{3}{4}} \left(\|(X_1)_-\|_{\mathrm{F}}^{\frac{1}{2}} + \|X^\mathsf{T}X - I_r\|_{\mathrm{F}}^{\frac{1}{2}} \right)$$

when $\|(X_1)_-\|_{\mathrm{F}} + \|X^\mathsf{T}X - I_r\|_{\mathrm{F}} \ge 1/(3\sqrt{r})$. Combining this with (3.6.11), we conclude that (3.6.10) is valid. The proof is complete. □

3.7 The general case of error bounds

We now present the error bounds for $\mathbb{S}_S^{n,r}$, detailed in Theorems 3.7–3.9. Theorems 3.7 and 3.8 can be proved using Proposition 3.6 and Theorem 3.2, respectively. We omit the proofs because they are essentially the same as that of Theorem 3.9 below.

Theorem 3.7. *Suppose that $|\mathcal{P}| + |\mathcal{N}| = 1$. For any matrix $X \in \mathbb{R}^{n\times r}$, we have*

$$\operatorname{dist}(X,\, \mathbb{S}_S^{n,r}) \;\le\; 7\sqrt{r}\left(\|(S \circ X)_-\|_{\mathrm{F}} + \|X^\mathsf{T}X - I_r\|_{\mathrm{F}} \right).$$

Moreover, if $\|X^\mathsf{T}X - I_r\|_{\mathrm{F}} < 1/3$, then

$$\operatorname{dist}(X,\, \mathbb{S}_S^{n,r}) \;\le\; 7\left(\|(S \circ X)_-\|_{\mathrm{F}} + \|X^\mathsf{T}X - I_r\|_{\mathrm{F}} \right).$$

Theorem 3.8. *Suppose that $|\mathcal{P}| + |\mathcal{N}| = n$. For any matrix $X \in \mathbb{R}^{n\times n}$, we have*

$$\operatorname{dist}(X,\, \mathbb{S}_S^{n,n}) \;\le\; 9n\left(\|(S \circ X)_-\|_{\mathrm{F}} + \|\sigma(X) - \mathbf{1}\|_2 \right).$$

Moreover, if $\|(S \circ X)_-\|_{\mathrm{F}} + \|\sigma(X) - \mathbf{1}\|_2 < 1/(4\sqrt{n})$, then

$$\operatorname{dist}(X,\, \mathbb{S}_S^{n,n}) \;\le\; 8\sqrt{n}\left(\|(S \circ X)_-\|_{\mathrm{F}} + \|\sigma(X) - \mathbf{1}\|_2 \right).$$

Theorem 3.9. *For any matrix $X \in \mathbb{R}^{n\times r}$, we have*

$$\operatorname{dist}(X,\, \mathbb{S}_S^{n,r}) \;\le\; 15r^{\frac{3}{4}} \left(\|(S \circ X)_-\|_{\mathrm{F}}^{\frac{1}{2}} + \|X^\mathsf{T}X - I_r\|_{\mathrm{F}}^{\frac{1}{2}} \right). \tag{3.7.1}$$

Moreover, if $\|(S \circ X)_-\|_{\mathrm{F}} + \|X^\mathsf{T}X - I_r\|_{\mathrm{F}} < 1/(3\sqrt{r})$, then

$$\operatorname{dist}(X,\, \mathbb{S}_S^{n,r}) \;\le\; 15\sqrt{r} \left(\|(S \circ X)_-\|_{\mathrm{F}}^{\frac{1}{2}} + \|X^\mathsf{T}X - I_r\|_{\mathrm{F}}^{\frac{1}{2}} \right). \tag{3.7.2}$$

Proof. Let $\mathcal{Q} = \{1, \ldots, r\} \setminus (\mathcal{P} \cup \mathcal{N})$. With $M_{\mathcal{P}}$, $M_{\mathcal{N}}$, and $M_{\mathcal{Q}}$ being the submatrices of I_r containing the columns indexed by $\mathcal{P}$, $\mathcal{N}$, and $\mathcal{Q}$, respectively, we take the permutation matrix

$$\Pi \;=\; (M_{\mathcal{P}},\; M_{\mathcal{N}},\; M_{\mathcal{Q}}) \;\in\; \mathbb{R}^{r\times r}.$$

In addition, we take the diagonal matrix $D \in \mathbb{R}^{r\times r}$ with $D_{j,j} = -1$ if $j \in \mathcal{N}$ and $D_{j,j} = 1$ otherwise. Define $r_1 = |\mathcal{P}| + |\mathcal{N}|$. If $r_1 = 0$, then (3.7.1) and (3.7.2) hold because of Lemma 2.4. Hence we suppose that $r_1 \geq 1$ in the sequel.

Consider any matrix $X \in \mathbb{R}^{n\times r}$. Let $Y = XD\Pi$, and $\bar{Y}$ be the projection of Y onto $\mathbb{S}^{n,r}_{r_1,+}$ defined in (3.6.1). Set $\bar{X} = \bar{Y}\Pi^{\mathsf{T}}D$, which lies in $\mathbb{S}^{n,r}_{S}$. Then

$$\mathrm{dist}(X,\, \mathbb{S}^{n,r}_{S}) \;\leq\; \|X - \bar{X}\|_{\mathrm{F}} \;=\; \|Y\Pi^{\mathsf{T}}D - \bar{Y}\Pi^{\mathsf{T}}D\|_{\mathrm{F}} \;=\; \|Y - \bar{Y}\|_{\mathrm{F}}.$$

Invoking Proposition 3.7, we have

$$\|Y - \bar{Y}\|_{\mathrm{F}} \;\leq\; 15r^{\frac{3}{4}} \left(\|(Y_1)_-\|_{\mathrm{F}}^{\frac{1}{2}} + \|Y^{\mathsf{T}}Y - I_r\|_{\mathrm{F}}^{\frac{1}{2}} \right),$$

where Y_1 is the submatrix of Y containing the first r_1 columns. It is straightforward to verify that $\|(Y_1)_-\|_{\mathrm{F}} = \|(S \circ X)_-\|_{\mathrm{F}}$ and $\|Y^{\mathsf{T}}Y - I_r\|_{\mathrm{F}} = \|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}}$. Hence we obtain (3.7.1). The bound (3.7.2) can be established in a similar way. □

Chapter 4

Exact penalties for optimization on the nonnegative Stiefel manifold

In this chapter, as an application of the error bounds established in this thesis, we consider exact penalties for optimization problem (1.1.5). For simplicity, we will focus on the special case with

$$\mathbb{S}_S^{n,r} = \mathbb{S}_+^{n,r},$$

applying the bounds in Section 3.3. Essentially the same results can be established in the general case by exploiting the bounds in Section 3.6. The exact penalty results only require (local) Lipschitz continuity of F, and hence can be applied to nonsmooth optimization, for example, F involving a group sparse regularization term [32].

The exactness of penalty methods for problem (1.1.5) with $\mathbb{S}_S^{n,r} = \mathbb{S}_+^{n,r}$ has been studied in [14, 27]. In [14], an error bound is established for $\mathbb{S}_+^{n,r}$ relative to the set

$$\{X \in \mathbb{R}_+^{n\times r} : (X^{\mathsf{T}}X)_{j,j} = 1,\ j = 1, \ldots, r\},$$

and then the bound is used to analyze a penalty method. However, the error bound in [14] cannot be used to derive the values of ν and q in (1.1.2)–(1.1.4). In [27], the authors consider the penalty problem (1.1.6) with $q = 1$, and show this problem has the same global minimizers as problem (1.1.5) if each global optimal solution of (1.1.5) has no zero rows. Our exact penalty results only need the Lipschitz con-

tinuity of the objective function F in (1.1.5). In Section 4.3, we give a warning of using penalty method for solve optimization problem over nonnegative Stiefel manifold. The exponent of penalty function should be seriously consider, otherwise errors will occur. In Section 4.4, we design the smoothing proximal reweighted algorithm to solve the penalty problem and analysis the convergence properties in Section 4.5

4.1 Exactness for Lipschitz continuous objective functions

The error bounds (1.1.2)–(1.1.4) established in this thesis enable us to have the exactness of the penalized problem

$$\min\left\{F(X)+\mu\left(\|X_-\|_{\ell_p}^{q_1}+\|X^{\mathsf{T}}X-I_r\|_{\ell_p}^{q_2}\right):X\in\mathcal{S}\right\} \tag{4.1.1}$$

for solving (1.1.5) with $\mathbb{S}_S^{n,r}=\mathbb{S}_+^{n,r}$ only under the (local) Lipschitz continuity of function F. Here the set $\mathcal{S}\subset\mathbb{R}^{n\times r}$ is a set that contains $\mathbb{S}_+^{n,r}$, while the parameters μ, p, q_1, and q_2 are all positive. If $p=2$ and $q_1=q_2=q$, then the penalized problem (4.1.1) reduces to problems (1.1.6) and (1.1.7) when $\mathcal{S}$ equals $\mathbb{S}^{n,r}$ and $\mathbb{R}_+^{n\times r}$, respectively.

During the revision of this thesis, a very recent work [20] studied another exact penalty problem for (1.1.5) with $\mathbb{S}_S^{n,r}=\mathbb{S}_+^{n,r}$ based on an error bound for $\mathbb{S}_+^{n,r}$ relative to the set

$$\{X\in\mathbb{R}_+^{n\times r}:(X^{\mathsf{T}}X)_{j,j}\le 1,\ j=1,\dots,r\}. \tag{4.1.2}$$

Since our error bounds for $\mathbb{S}_+^{n,r}$ are established relative to $\mathbb{R}^{n\times r}$, we allow the feasible set of our penalty problem to be any set $\mathcal{S}$ containing $\mathbb{S}_+^{n,r}$, whereas the feasible set in [20] can only be the set (4.1.2). In addition, with the error bounds established in Section 3.6, our results can be readily extended to the case where $\mathbb{S}_S^{n,r}$ is a sign-constrained Stiefel manifold other than $\mathbb{S}_+^{n,r}$, which is not considered in [20].

Due to the equivalence between norms, it is indeed possible to establish the exactness of (4.1.1) when the entry-wise ℓ_p-norm is changed to other ones. We choose to use the entry-wise ℓ_p-norm in (4.1.1) because it is easy to evaluate.

Theorem 4.1 presents the exactness of problem (4.1.1) regarding global optimizers when the objective function $F : \mathcal{S} \to \mathbb{R}$ is an L-Lipschitz continuous function, namely

$$|F(X) - F(Y)| \ \leq \ L\|X - Y\|_{\mathrm{F}} \tag{4.1.3}$$

for all X and Y in $\mathcal{S}$, where $L \in (0, \infty)$ is a Lipschitz constant of F with respect to the Frobenius norm. Note that the global Lipschitz continuity of the objective function F is assumed on a set $\mathcal{S}$ containing $\mathbb{S}_+^{n,r}$. For example, if $F(X) = \mathrm{trace}(X^{\mathsf{T}}A^{\mathsf{T}}AX)$ and $\mathcal{S} = \{X \in \mathbb{R}^{n\times r} : \|X\|_{\mathrm{F}} \leq \gamma\}$ with $\gamma > \sqrt{r}$, the global Lipschitz continuity of F holds on $\mathcal{S}$ with the Lipschitz constant $L = 2\gamma\|A\|_2^2$. Indeed, our theory holds even if F is undefined out of $\mathcal{S}$. The proof of Theorem 4.1 is standard and we include it in Appendix 7 for completeness.

Theorem 4.1 (Exact penalty (4.1.1) with F being Lipschitz continuous)**.** *Suppose that $\mathcal{S} \subset \mathbb{R}^{n\times r}$ is a set containing $\mathbb{S}_+^{n,r}$, $F : \mathcal{S} \to \mathbb{R}$ is an L-Lipschitz function, and $p \geq 1$ is a constant. If $0 < q \leq 1/2$ and $\mu > 5Lr^{\frac{3}{4}} \max\Big\{1, (nr)^{\frac{p-2}{4p}}\Big\}$, then*

$$\mathrm{Argmin}\{F(X) : X \in \mathbb{S}_+^{n,r}\} = \mathrm{Argmin}\Big\{F(X) + \mu\big(\|X_-\|_{\ell_p}^q + \|X^{\mathsf{T}}X - I_r\|_{\ell_p}^{\frac{1}{2}}\big) : X \in \mathcal{S}\Big\}.$$

Theorem 4.2 presents the exactness of problem (4.1.1) regarding local minimizers when F is locally Lipschitz continuous on $\mathcal{S}$, meaning that for any $\bar{X} \in \mathcal{S}$ there exists a constant $L \in (0, \infty)$ such that (4.1.3) holds for all X and Y in a certain neighborhood of $\bar{X}$ in $\mathcal{S}$. We will refer to this L as a Lipschitz constant of F around $\bar{X}$. The proof of Theorem 4.2 is also given in Appendix 7.

Theorem 4.2 (Exact penalty (4.1.1) with F being locally Lipschitz continuous)**.** *Let $\mathcal{S} \subset \mathbb{R}^{n\times r}$ be a set containing $\mathbb{S}_+^{n,r}$, $F : \mathcal{S} \to \mathbb{R}$ be a locally Lipschitz continuous*

function, and $p \geq 1$ *be a constant. Suppose that* $0 < q_1 \leq 1/2$ *and* $0 < q_2 \leq 1/2$. *For any local minimizer* X^* *of* F *on* $\mathbb{S}_+^{n,r}$, X^* *is also a local minimizer of*

$$\min \left\{ F(X) + \mu(\|X_-\|_{\ell_p}^{q_1} + \|X^\mathsf{T} X - I_r\|_{\ell_p}^{q_2}) : X \in \mathcal{S} \right\} \tag{4.1.4}$$

for all $\mu > 4L^* \sqrt{r} \max\left\{1, (nr)^{\frac{q_1(p-2)}{2p}}, r^{\frac{q_2(p-2)}{p}}\right\}$, *where* L^* *is a Lipschitz constant of* F *around* X^*. *Conversely, if* X^* *lies in* $\mathbb{S}_+^{n,r}$ *and there exists a constant* μ *such that* X^* *is a local minimizer of* (4.1.4), *then* X^* *is also a local minimizer of* F *on* $\mathbb{S}_+^{n,r}$.

Suppose that $p \leq 2$. It is noteworthy that the thresholds for μ in Theorems 4.1 and 4.2 are independent of n (even the dependence on r is mild). This is favorable in practice, as r can be much smaller than n in applications. We also note that the second part of Theorem 4.2 requires $X^* \in \mathbb{S}_+^{n,r}$. This is indispensable without additional assumptions on the problem structure (see [7, Remark 9.1.1]).

4.2 The exponents in the penalty term

When $1 < r < n$, the requirements on the exponents of $\|X_-\|_{\mathrm{F}}$ and $\|X^\mathsf{T} X - I_r\|_{\mathrm{F}}$ in Theorems 4.1 and 4.2 cannot be relaxed. This is elaborated in Proposition 4.1, with $\mathcal{S} = \mathbb{R}^{n\times r}$ being an example. Similar results can be proved for $\mathcal{S} = \mathbb{S}^{n,r}$ and $\mathcal{S} = \mathbb{R}_+^{n\times r}$.

Proposition 4.1. *Suppose that* $1 < r < n$, $p \geq 1$, $q_1 > 0$, *and* $q_2 > 0$. *Define the function* $\rho(X) = \|X_-\|_{\ell_p}^{q_1} + \|X^\mathsf{T} X - I_r\|_{\ell_p}^{q_2}$ *for* $X \in \mathbb{R}^{n\times r}$. *There exists a Lipschitz continuous function* $F : \mathbb{R}^{n\times r} \to \mathbb{R}$ *such that the following statements hold.*

(a) $\mathrm{Argmin}\{F(X) : X \in \mathbb{S}_+^{n,r}\} = \mathbb{S}_+^{n,r}$.

(b) *If* $q_1 > 1/2$ *or* $q_2 \neq 1/2$, *then any* $X^* \in \mathbb{S}_+^{n,r}$ *is not a global minimizer of* $F + \mu\rho$ *on* $\mathbb{R}^{n\times r}$ *for any* $\mu > 0$.

(c) *If $q_1 > 1/2$ or $q_2 > 1/2$, then there exists an $X^* \in \mathbb{S}_+^{n,r}$ that is not a local minimizer of $F + \mu\rho$ on $\mathbb{R}^{n\times r}$ for any $\mu > 0$.*

Proof. Define

$$F(X) \;=\; -\operatorname{dist}(X, \mathbb{S}_+^{n,r}) \quad \text{for} \quad X \in \mathbb{R}^{n\times r}.$$

Then F is Lipschitz continuous on $\mathbb{R}^{n\times r}$. We will justify (a)–(c) one by one.

(a) This holds because F takes a constant value 0 on $\mathbb{S}_+^{n,r}$.

(b) Assume for contradiction that there exists an $X^* \in \mathbb{S}_+^{n,r}$ such that X^* is a global minimizer of $F + \mu^*\rho$ on $\mathbb{R}^{n\times r}$ for a certain $\mu^* > 0$. Then

$$F(X) + \mu^*\rho(X) \;\geq\; F(X^*) + \mu^*\rho(X^*) \;=\; 0 \quad \text{for all} \quad X \in \mathbb{R}^{n\times r}.$$

By the definition of F, we then have $\operatorname{dist}(X, \mathbb{S}_+^{n,r}) \leq \mu^*\rho(X)$ for all $X \in \mathbb{R}^{n\times r}$. Hence ρ defines a global error bound for $\mathbb{S}_+^{n,r}$ relative to $\mathbb{R}^{n\times r}$, contradicting (b) of Theorem 3.6 (note that $\|\cdot\|_{\ell_p}$ and $\|\cdot\|_{\mathrm{F}}$ are equivalent norms).

(c) According to (b) of Theorem 3.6, the function ρ does not define a local error bound for $\mathbb{S}_+^{n,r}$ relative to $\mathbb{R}^{n\times r}$. Thus there is a sequence $\{X_k\} \subset \mathbb{R}^{n\times r}$ such that

$$\|(X_k)_-\|_{\mathrm{F}} + \|X_k^{\mathsf{T}}X_k - I_r\|_{\mathrm{F}} \;\leq\; k^{-1}, \tag{4.2.1}$$

$$\operatorname{dist}(X_k, \mathbb{S}_+^{n,r}) \;>\; k\rho(X_k) \tag{4.2.2}$$

for each $k \geq 1$. According to (4.2.1), $\|X_k^{\mathsf{T}}X_k\|_{\mathrm{F}} \leq \sqrt{r} + k^{-1}$. Thus $\{X_k\}$ has a subsequence $\{X_{k_\ell}\}$ that converges to a certain point X^*. Using (4.2.1) again, we have $\|X_-^*\|_{\mathrm{F}} + \|(X^*)^{\mathsf{T}}X^* - I_r\|_{\mathrm{F}} = 0$, and hence $X^* \in \mathbb{S}_+^{n,r}$. It remains to show that X^* is not a local minimizer of $F + \mu\rho$ for any $\mu > 0$. Assume for contradiction that X^* is such a local minimizer for a certain $\mu^* > 0$. Then for all sufficiently large ℓ,

$$F(X_{k_\ell}) + \mu^*\rho(X_{k_\ell}) \;\geq\; F(X^*) + \mu^*\rho(X^*) \;=\; 0.$$

By the definition of F, we then have $\operatorname{dist}(X_{k_\ell}, \mathbb{S}_+^{n,r}) \leq \mu^*\rho(X_{k_\ell})$, contradicting (4.2.2). The proof is complete. □

When $r = 1$ or $r = n$, since the exponents of $\|X_-\|_{\rm F}$ and $\|X^{\mathsf{T}}X - I_r\|_{\rm F}$ in the error bounds can be increased from 1/2 to 1, their exponents in the penalty term of (4.1.1) can be taken from a larger range while keeping the exactness of (4.1.1). This is briefly summarized in Remark 4.1.

Remark 4.1. *Suppose that $r = 1$ or $r = n$. If F is Lipschitz continuous on $\mathcal{S}$, then we can establish a result similar to Theorem 4.1 for $0 < q_1 \le 1$ and $1/2 \le q_2 \le 1$ based on the error bound* (3.2.10). *When F is only locally Lipschitz continuous, similar to Theorem 4.2, the exactness of problem* (4.1.1) *regarding local minimizers can be established if $0 < q_1 \le 1$ and $0 < q_2 \le 1$. Proposition 4.1 can also be adapted to the case of $r = 1$ or $r = n$. It is also worth noting that $\mathbb{S}_+^{n,n}$ is precisely the set of $n \times n$ permutation matrices, and hence $\min\{F(X) : X \in \mathbb{S}_+^{n,n}\}$ represents optimization problems over permutation matrices.*

4.3 Warning of penalty methods for optimization over nonnegative Stifel manifold

When $\mathbb{S}_S^{n,r} = \mathbb{S}_+^{n,r}$, problem (1.1.5) reduces to the nonnegative orthogonal constrained optimization problem

$$\min_{X\in\mathbb{S}_+^{n,r}} F(X). \tag{4.3.1}$$

Many papers use penalty methods for problem (4.3.1) with penalty functions $\|\cdot\|_{\rm F}^2$, $\|\cdot\|_{\rm F}$ or $\|\cdot\|_{\ell_1}$ of X_- or $X^{\mathsf{T}}X - I_r$, e.g., [1, 21, 34, 36]. However, there is not a satisfactory answer in existing literature whether the penalty problem using $\|\cdot\|_{\rm F}^2$, $\|.\|_{\rm F}$ or $\|\cdot\|_{\ell_1}$ is an exact penalty regarding local and global minimizers of problem (4.3.1) for a Lipschitz continuous objective function.

In 2024, the authors of [27] proved that the penalty problem

$$\min_{X\in\mathbb{S}^{n,r}} F(X) + \mu\|X_-\|_{\ell_1} \tag{4.3.2}$$

is a global exact penalty for problem (4.3.1) under the assumption that any global minimizer has no zero rows. Moreover, they aimed to show that such strong assumption cannot be removed by Example 3.9 in [27], which is as follows

$$\min_{X\in\mathbb{S}_+^{3,2}} f(X) := -2X_{1,1} - 2X_{2,2} - X_{3,1} - X_{3,2}. \tag{4.3.3}$$

The authors of [27] claimed $X^* = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}^\mathsf{T}$ is a global minimizer of (4.3.3), but is not a solution of the penalty problem

$$\min_{X\in\mathbb{S}^{3,2}} f(X) + \mu\|X_-\|_{\ell_1}$$

for any $\mu > 0$. However, X^* is not a global minimizer of (4.3.3), since $f(X^*) = -4 > -\sqrt{5} - 2 = f(\hat{X})$, where $\hat{X} = \begin{bmatrix} 2/\sqrt{5} & 0 & 1/\sqrt{5} \\ 0 & 1 & 0 \end{bmatrix}^\mathsf{T}$. Thus the claim with this example in [27] is wrong.

The column vectors of an orthogonal matrix are not only satisfy unity but also orthogonal to each other, which is also one of the difficulties in the optimization over the Stiefel manifold. Coupling-constrained optimization problems are often challenging. It seems the authors in [27] want to increase the weight of some elements in the orthogonal matrix in the objective function to satisfy the unity and orthogonality with a simple matrix which submatrix is an identity matrix, but such an operation fails to decouple all the relationships. To be specific, there is a conflict between the linearity of the objective function and the unity, orthogonality, and non negativity of the column vectors.

In this thesis, we give a warning for the penalty problem (4.3.2) in the case where the objective function is only Lipschitz continuous. From Proposition 4.1, we know that there is a Lipschitz continuous function F such that any global (local) minimizer

of (4.3.1) is not a global (local) minimizer of (4.3.2) for any $\mu > 0$. On the other hand, from Theorem 4.1 and Theorem 4.2, we know that

$$\min_{X \in \mathbb{S}^{n,r}} F(X) + \mu \|X_-\|_{\ell_1}^q$$

is an exact penalty problem for (4.3.1) regarding global and local minimizers for $\mu > 5Lr^{\frac{3}{4}}$ and $q \in (0, 1/2]$, where L is a Lipschitz constant of F. Our results provide theoretical warning and guarantee for penalty methods of nonnegative orthogonal constrained optimization problem (4.3.1).

4.4 The smoothing proximal reweighted algorithm

Let $f : \mathbb{R}^{n\times r} \to \mathbb{R}$ be a convex and L-Lipschitz continuous function and $\mathbb{S}_S^{n,r} = \mathbb{S}_+^{n,r}$. Consider the following nonnegative orthogonal constrained optimization problem,

$$\min_{X \in \mathbb{S}_+^{n,r}} f(X). \tag{4.4.1}$$

We can construct the following penalized optimization problem,

$$\begin{aligned} \min \quad & f(X) + \lambda \left(\sum_{i=1}^{n} \sum_{j=1}^{r} \max(-X_{ij}, 0) \right)^{\frac{1}{2}} \\ \text{s.t.} \quad & X^\top X = I_r, \end{aligned} \tag{4.4.2}$$

where $\lambda > 5Lr^{\frac{3}{4}}$ is a penalty parameter. Setting $\mathcal{S} = \mathbb{S}^{n,r}$ in (5.1.1), by the argument in Chapter 5, it is clear that (4.4.1) and (4.4.2) share the same global minimizers.

Since the inequality $(t+s)^{\frac{1}{2}} \leq t^{\frac{1}{2}} + s^{\frac{1}{2}}$ holds for any two nonnegative numbers t and s, we have

$$\left(\sum_{i=1}^{n} \sum_{j=1}^{r} \max(-X_{ij}, 0) \right)^{\frac{1}{2}} \leq \sum_{i=1}^{n} \sum_{j=1}^{r} \max(-X_{ij}, 0)^{\frac{1}{2}}.$$

This observation motivates us to consider the following optimization problem,

$$\begin{aligned} \min \quad & F(X) \coloneqq f(X) + \lambda P(X) \\ \text{s.t.} \quad & X^{\mathsf{T}} X = I_r, \end{aligned} \tag{4.4.3}$$

where

$$P(X) := \sum_{i=1}^{n} \sum_{j=1}^{r} \max(-X_{ij}, 0)^{\frac{1}{2}}.$$

Then it can be readily verified that the global minimizers of problems (4.4.1) and (4.4.3) coincide with each other. However, it is highly challenging to solve problem (4.4.3) since its objective function fails to be locally Lipschitz continuous. To address this issue, we endeavor to solve the approximation problem of (4.4.3) as follows,

$$\begin{aligned} \min \quad & F_\varepsilon(X) \ \coloneqq \ f(X) + \lambda P_\varepsilon(X) \\ \text{s.t.} \quad & X^{\mathsf{T}} X = I_r, \end{aligned} \tag{4.4.4}$$

where

$$P_\varepsilon(X) \ \coloneqq \ \sum_{i=1}^{n} \sum_{j=1}^{r} (\max(-X_{ij}, 0) + \varepsilon)^{\frac{1}{2}},$$

and $\varepsilon > 0$ is a small constant.

The Lagrangian function of problem (4.4.4) is

$$\mathcal{L}(X, \Lambda) = f(X) + \lambda P_\varepsilon(X) - \frac{1}{2} \langle \Lambda, X^{\mathsf{T}} X - I_r \rangle.$$

For any local minimizer $\tilde{X}$, the corresponding KKT system is

$$0 \in \partial \mathcal{L}(\tilde{X}, \Lambda), \quad \tilde{X}^{\mathsf{T}} \tilde{X} = I_r.$$

Here, Λ is the associated Lagrangian multiplier, which is symmetric due to the symmetry of $X^{\mathsf{T}} X$. Then there exists $\tilde{V} \in \partial f(\tilde{X})$ and $\tilde{P} \in \partial P_\varepsilon(\tilde{X})$ such that

$$\tilde{V} + \lambda \tilde{P} - \tilde{X} \Lambda = 0.$$

By solving the above equation, we can obtain that $\Lambda = \tilde{X}^{\mathsf{T}}(\tilde{V} + \lambda\tilde{P})$. Together with the symmetric condition, we derive the following first-order necessary condition of (4.4.4),

$$\begin{cases} (I_n - \tilde{X}\tilde{X}^{\mathsf{T}})(\tilde{V} + \lambda\tilde{P}) \;=\; 0, \\ \tilde{X}^{\mathsf{T}}(\tilde{V} + \lambda\tilde{P}) \;=\; (\tilde{V} + \lambda\tilde{P})^{\mathsf{T}}\tilde{X}, \\ \tilde{X}^{\mathsf{T}}\tilde{X} \;=\; I_r. \end{cases} \tag{4.4.5}$$

Definition 4.1. *A point $\tilde{X}$ is called a limiting stationary point of problem* (4.4.4) *if there exists $\tilde{V} \in \partial f(\tilde{X})$ and $\tilde{P} \in \partial P_\varepsilon(\tilde{X})$ such that the conditions in* (4.4.5) *are satisfied.*

In this chapter, we follow the idea of [6] to apply the iteratively reweighted ℓ_2 minimization algorithm (IRL2) to solve problem (4.4.4). For convenience, we define $Y_{ij} = \max(-X_{ij}, 0) + \varepsilon$. Let

$$Q_\varepsilon(Y) \;\coloneqq\; \sum_{i=1}^{n}\sum_{j=1}^{r} Y_{ij}^{\frac{1}{2}}.$$

Then it holds that

$$P_\varepsilon(X) = Q_\varepsilon(Y).$$

In our algorithm, at the current iterate X^k, we construct an approximation of $F_\varepsilon(X)$ as follows,

$$F_\varepsilon^k(X) \;\coloneqq\; f(X) + \lambda P_\varepsilon^k(X),$$

where

$$P_\varepsilon^k(X) \;\coloneqq\; \sum_{i=1}^{n}\sum_{j=1}^{r} W_{ij}^k(\max(-X_{ij}, 0) + \varepsilon)^2,$$

and

$$W_{ij}^k \coloneqq \frac{1}{4}(\max(-X_{ij}^k, 0) + \varepsilon)^{-\frac{3}{2}}$$

is the (i,j)-th entry of the weight matrix W^k. Let

$$Q_\varepsilon^k(Y) \coloneqq \sum_{i=1}^{n}\sum_{j=1}^{r} W_{ij}^k Y_{ij}^2.$$

Then we have

$$P_\varepsilon^k(X) = Q_\varepsilon^k(Y).$$

Algorithm 1 outlines the complete procedure of our approach for solving problem (4.4.4), which is named *proximal iteratively reweighted ℓ_2 method* and abbreviated to PIRL2. In each iteration, we solve the following proximal reweighted problem

$$\begin{aligned} \min \quad & F_\varepsilon^k(X) + \frac{\gamma_k}{2}\|X - X^k\|_{\mathrm{F}}^2. \\ \text{s.t.} \quad & X^\mathsf{T} X = I_r \end{aligned} \tag{4.4.6}$$

to update the next iterate X^{k+1}. The first-order stationary condition of a point $\tilde{X}$ for problem (4.4.6) can be stated as follows,

$$\begin{cases} (I_n - \tilde{X}\tilde{X}^\mathsf{T})(\tilde{V} + \lambda\tilde{P} + \gamma_k(\tilde{X} - X^k)) \;=\; 0, \\ \tilde{X}^\mathsf{T}(\tilde{V} + \lambda\tilde{P} + \gamma_k(\tilde{X} - X^k)) \;=\; (\tilde{V} + \lambda\tilde{P} + \gamma_k(\tilde{X} - X^k))^\mathsf{T}\tilde{X}, \\ \tilde{X}^\mathsf{T}\tilde{X} \;=\; I_r, \end{cases} \tag{4.4.7}$$

where $\tilde{V} \in \partial f(\tilde{X})$ and $\tilde{P} \in \partial P_\varepsilon^k(\tilde{X})$.

In the remaining part of this chapter, we will prove that any accumulation point of the iterate sequence generated by Algorithm 1 is a stationary point of problem (4.4.4).

Algorithm 1: Proximal Iteratively Reweighted ℓ_2 Method (PIRL2) for (4.4.4).

Input: $\gamma_0 > 0$, $\bar{\gamma} > 0$, $\tau > 1$.

Initialization: Generate the initial point $X^0 \in \mathbb{S}^{n,r}$ randomly.

for $k = 0, 1, \cdots$ **do**

Update X^{k+1} by solving the following subproblem,

$$X^{k+1} \in \underset{X^{\mathsf{T}}X = I_r}{\arg\min} F_\varepsilon^k(X) + \frac{\gamma_k}{2}\|X - X^k\|_{\mathrm{F}}^2. \tag{4.4.8}$$

Set $\gamma_{k+1} = \min(\tau\gamma_k, \bar{\gamma})$

end

/Post-procedure

In the post-procedure of PIRL2, we adopt the same method in [14, Algorithm 4.1] to improve the quality of the solution.

4.5 Sufficient decrease and subsequence convergence

Lemma 4.1. *Let $\{X^k\}$ be the sequence generated by PIRL2. Then we have*

$$F_\varepsilon(X^k) - F_\varepsilon(X^{k+1}) \geq \frac{\gamma_k}{2}\|X^{k+1} - X^k\|_{\mathrm{F}}^2. \tag{4.5.1}$$

Proof. By the global optimality of X^{k+1} in the subproblem (4.4.8), we have

$$F_\varepsilon^k(X^{k+1}) + \frac{\gamma_k}{2}\|X^{k+1} - X^k\|_{\mathrm{F}}^2 \ \leq \ F_\varepsilon^k(X^k),$$

which implies that

$$\begin{aligned} F_\varepsilon(X^{k+1}) - F_\varepsilon(X^k) \ \leq & \lambda P_\varepsilon(X^{k+1}) - \lambda P_\varepsilon(X^k) \\ & + \lambda P_\varepsilon^k(X^k) - \lambda P_\varepsilon^k(X^{k+1}) \\ & - \frac{\gamma_k}{2}\|X^{k+1} - X^k\|_{\mathrm{F}}^2. \end{aligned} \tag{4.5.2}$$

Consider the univariate function $r(x) = x^{\frac{1}{2}}$. Since $\nabla^2 r(x) = -\frac{1}{4}x^{-\frac{3}{2}} < -\frac{1}{4}(1+\varepsilon)^{-\frac{3}{2}} < -2^{-3}$, the function r is -2^{-3}- strongly concave over $[\varepsilon, 1+\varepsilon]$. Similarly, the univariate

function $r_{ij}^k(x) = W_{ij}^k x^2$ is 2^{-2}-strongly convex since $\nabla^2 r^k(x) = 2W_{ij}^k = \frac{1}{2}(Y_{ij}^k)^{-\frac{3}{2}} > 2^{-2}$ due to $Y_{ij} \in [\varepsilon, 1+\varepsilon]$. By virtue of the relationships $Q_\varepsilon(Y) = \sum_{i=1}^n \sum_{j=1}^r r(Y_{ij})$ and $Q_\varepsilon^k(Y) = \sum_{i=1}^n \sum_{j=1}^r r_{ij}^k(Y_{ij})$, we can obtain that

$$\begin{aligned}
F_\varepsilon(X^{k+1}) - F_\varepsilon(X^k) \leq\ & \lambda(P_\varepsilon(X^{k+1}) - P_\varepsilon(X^k) + P_\varepsilon^k(X^k) - P_\varepsilon^k(X^{k+1})) \\
& - \frac{\gamma_k}{2}\|X^{k+1} - X^k\|_{\mathrm{F}}^2 \\
\leq\ & \lambda(Q_\varepsilon(Y^{k+1}) - Q_\varepsilon(Y^k) + Q_\varepsilon^k(Y^k) - Q_\varepsilon^k(Y^{k+1})) \\
& - \frac{\gamma_k}{2}\|X^{k+1} - X^k\|_{\mathrm{F}}^2 \\
\leq\ & \lambda(\langle \nabla Q_\varepsilon(Y^k), Y^{k+1} - Y^k\rangle - 2^{-4}\|Y^{k+1} - Y^k\|_{\mathrm{F}}^2 \\
& - \langle \nabla Q_\varepsilon^k(Y^k), Y^{k+1} - Y^k\rangle - 2^{-3}\|Y^{k+1} - Y^k\|_{\mathrm{F}}^2) \\
& - \frac{\gamma_k}{2}\|X^{k+1} - X^k\|_{\mathrm{F}}^2 \\
\leq\ & -2^{-3}\lambda\|Y^{k+1} - Y^k\|_{\mathrm{F}}^2 - \frac{\gamma_k}{2}\|X^{k+1} - X^k\|_{\mathrm{F}}^2 \\
\leq\ & -\frac{\gamma_k}{2}\|X^{k+1} - X^k\|_{\mathrm{F}}^2.
\end{aligned} \tag{4.5.3}$$

Here, we use the strong concavity of Q_ε and strong convexity of Q_ε^k to derive the third inequality. □

Lemma 4.2. *Let $\{X^k\}$ be the sequence generated by PIRL2. Then we have*

$$\sum_{k=1}^{\infty} \|X^{k+1} - X^k\|_{\mathrm{F}}^2 < \infty. \tag{4.5.4}$$

Proof. From Lemma 4.1, we have

$$\begin{aligned}\sum_{i=1}^{k}\|X^{k+1}-X^{k}\|_{\mathrm{F}}^{2} &\leq \sum_{i=1}^{k}\frac{2}{\gamma_i}(F_{\varepsilon}(X^{i})-F_{\varepsilon}(X^{i+1}))\\ &\leq 2\sum_{i=1}^{k}F_{\varepsilon}(X^{i})-F_{\varepsilon}(X^{i+1})\\ &\leq 2\sum_{i=1}^{k}F_{\varepsilon}(X^{i})-F_{\varepsilon}(X^{i+1})\\ &= 2(F_{\varepsilon}(X^{1})-F_{\varepsilon}(X^{k+1})).\end{aligned} \tag{4.5.5}$$

Notice that F_ε is bounded below. We finish the proof by combining the last equality of (4.5.5) and letting $k \to \infty$. □

Theorem 4.3. *Let $\{X^k\}$ be the sequence generated by PIRL2. Then any accumulation point of $\{X^k\}$ is a limiting stationary point of problem* (4.4.4).

Proof. To begin with, the sequence $\{X^k\}$ is bounded due to the compactness of the Stiefel manifold $\mathbb{S}^{n,r}$. Hence, there exists a convergent subsequence $\{X^{n_k}\}$ of $\{X^k\}$. We assume that it converges to $\bar{X} \in \mathbb{S}^{n,r}$. Since X^{n_k} is a first-order stationary point of (4.4.6), there exists $v^{n_k}(X^{n_k}) \in \partial f(X^{n_k})$ and $p^{n_k-1}(X^{n_k}) \in \partial P_\varepsilon^{n_k-1}(X^{n_k})$ satisfying the necessary condition (4.4.7), where

$$(p^{n_k-1}(X^{n_k}))_{ij} \;=\; 2W_{ij}^{n_k-1}(\max(-X_{ij}^{n_k},0)+\varepsilon)C_{ij}^{n_k} \tag{4.5.6}$$

with some $C_{ij}^{n_k} \in \partial \max(-X_{ij}^{n_k},0)$.

Next, it follows from Lemma 4.2 that $\lim_{k\to\infty}\|X^{n_k}-X^{n_k-1}\|_{\mathrm{F}} = 0$. Since X^{n_k}

converges to $\bar{X}$, there exists $\bar{C}_{ij} \in \partial \max(-\bar{X}_{ij}, 0)$ such that

$$\begin{aligned}\lim_{k\to\infty}(p^{n_k-1}(X^{n_k}))_{ij} &= \lim_{k\to\infty}\frac{1}{2}(\max(-X_{ij}^{n_k-1},0)+\varepsilon)^{-\frac{3}{2}}(\max(-X_{ij}^{n_k},0)+\varepsilon)C_{ij}^{n_k}\\ &= \lim_{k\to\infty}\frac{1}{2}(\max(-X_{ij}^{n_k},0)+\varepsilon)^{-\frac{3}{2}}(\max(-X_{ij}^{n_k},0)+\varepsilon)C_{ij}^{n_k}\\ &= \frac{1}{2}(\max(-\bar{X}_{ij},0)+\varepsilon)^{-\frac{1}{2}}\bar{C}_{ij},\end{aligned} \tag{4.5.7}$$

where the last equality holds due to the upper hemicontinuity of subdifferentials [29, 28]. Similarly, we can prove that there exists $v(\bar{X}) \in \partial f(\bar{X})$ such that

$$\lim_{k\to\infty} v^{n_k}(X^{n_k}) = v(\bar{X}).$$

Let $(p(\bar{X}))_{ij} = \frac{1}{2}(\max(-\bar{X}_{ij},0)+\varepsilon)^{-\frac{1}{2}}\bar{C}_{ij}$. Then it is clear that $p(\bar{X}) \in \partial P_\varepsilon(\bar{X})$. Combining with the fact that X^{n_k} is a first-order stationary point of (4.4.6), we have

$$\begin{cases} (I_n - X^{n_k}(X^{n_k})^{\mathsf{T}})(v^{n_k}(X^{n_k}) + \lambda p^{n_k-1}(X^{n_k}) + \gamma_{n_k-1}(X^{n_k} - X^{n_k-1})) = 0,\\ (X^{n_k})^{\mathsf{T}}(v^{n_k}(X^{n_k}) + \lambda p^{n_k-1}(X^{n_k}) + \gamma_{n_k-1}(X^{n_k} - X^{n_k-1}))\\ \quad = (v^{n_k}(X^{n_k}) + \lambda p^{n_k-1}(X^{n_k}) + \gamma_{n_k-1}(X^{n_k} - X^{n_k-1}))^{\mathsf{T}}X^{n_k},\\ (\bar{X})^{\mathsf{T}}\bar{X} - I_r = 0. \end{cases}$$

Taking $k \to \infty$ in the above relationships, we can obtain that

$$\begin{cases} (I_n - \bar{X}\bar{X}^{\mathsf{T}})(v(\bar{X}) + \lambda p(\bar{X})) = 0,\\ (\bar{X})^{\mathsf{T}}(v(\bar{X}) + \lambda p(\bar{X})) = (v(\bar{X}) + \lambda p(\bar{X}))^{\mathsf{T}}\bar{X},\\ (\bar{X})^{\mathsf{T}}\bar{X} - I_r = 0. \end{cases}$$

Therefore, $\bar{X}$ is a limiting stationary point of problem (4.4.4). □

Remark 4.2. *In this thesis, we use the penalty method to solve the optimization problem with the sign and orthogonality constraints. Firstly, we establish the error bound over the sign-constrained Stiefel manifold to derive the exact penalty model. Then we use the reweighted method to solve the exact penalty model* (4.4.3) *in Sections 4.4 and 4.5. The exact penalty model of the problem* (1.1.5) *can be solved in*

the same way by replacing $P(X) \coloneqq \sum_{i=1}^{n}\sum_{j=1}^{r}\max(-X_{ij},0)^{\frac{1}{2}}$ *with* $P(X) \coloneqq \sum_{i=1}^{n}\sum_{j=1}^{r}\max(-S_{ij}X_{ij},0)^{\frac{1}{2}}$. *Moreover the above convergence properties can be applied to the sign-constrained Stiefel manifold optimization problem* (1.1.5) *similarly.*

Chapter 5

Numerical experiments

In this chapter, we first show the advantages of posing sign-constraint via the sparse trace minimization problem in Section 5.1. In the following two sections, we test the performance of PIRL2 based on two applications, including projection to nonnegative Stiefel manifold and quadratic assignment problem.

5.1 Sparse trace minimization with sign-constraint

Note that Remark 4.1 can be extended to the case $|\mathcal{P}| + |\mathcal{N}| = 1$ or $|\mathcal{P}| + |\mathcal{N}| = n$. In particular, the penalty problem

$$\min_{X \in \mathbb{S}^{n,r}} F(X) + \mu \|(S \circ X)_-\|_{\ell_1}$$

is an exact penalty problem of (1.1.5) with $S_{i,1} = 1$ and $S_{i,j} = 0$, for $j \neq 1, i = 1, \ldots, n$.

Consider the following sparse trace maximization problem [4]

$$\min_{X \in \mathbb{S}^{n,r}} -\mathrm{tr}(X^\mathsf{T} A^\mathsf{T} A X) + \lambda \|X\|_{\ell_1}, \tag{5.1.1}$$

where $A \in \mathbb{R}^{m \times n}$ is a given matrix. If $A^\mathsf{T} A$ is a positive or an irreducible nonnegative matrix, then by the Perron-Frobenius theorem, the largest eigenvalue of $A^\mathsf{T} A$ is positive and the corresponding eigenvector is positive. Hence, for a dense nonnegative

data matrix A, it is interesting to consider

$$\min_{X\in\mathbb{S}^{n,r}_S} -\mathrm{tr}(X^\mathsf{T}A^\mathsf{T}AX) + \lambda\|T\circ X\|_{\ell_1}, \tag{5.1.2}$$

with $S_{i,1}=1$, $S_{i,j}=0$, $T_{i,1}=0, T_{i,j}=1$, for $j\neq 1, i=1,\ldots,n$. Since the objective function of (5.1.2) is Lipschitz continuous with Lipschitz constant $L=2\|A\|_2^2+r\lambda\sqrt{n}$ over $\mathbb{S}^{n,r}$, our results show that

$$\min_{X\in\mathbb{S}^{n,r}} -\mathrm{tr}(X^\mathsf{T}A^\mathsf{T}AX) + \lambda\|T\circ X\|_{\ell_1} + \mu\|(S\circ X)_-\|_{\ell_1} \tag{5.1.3}$$

is an exact penalty problem of (5.1.2) with $\mu > 5Lr^{\frac{3}{4}}$.

In [4], Chen et. al proposed a ManPG (Manifold Proximal Gradient) algorithm to solve the following nonsmooth optimization problem

$$\min_{X\in\mathbb{S}^{n,r}} F(X) := f(X) + h(X),$$

where f is smooth, ∇f is Lipschitz continuous and h is nonsmooth, convex and Lipschitz continuous. The objective functions in problem (5.1.1) and problem (5.1.3) satisfy these conditions. Numerical results in [4] show that ManPG outperforms some existing algorithms for solving problem (5.1.1). We compare the two models (5.1.1) and (5.1.3) for sparse trace maximization problem by using the code of [4] downloaded from https://github.com/chenshixiang/ManPG, with the same initial points that are randomly generated by the code. Other algorithms for solving nonsmooth matrix optimization over $\mathbb{S}^{n,r}$ can be found in [38] and its references. Moreover, we can also replace the orthogonal constraint by adding a penalty term $\|X^\mathsf{T}X - I_r\|_{\ell_1}$ to (5.1.3).

5.1.1 Synthetic simulations

For given m, n, we randomly generated 20 nonnegative matrices and then normalized the columns by Matlab functions as follows

$$A = \mathrm{rand}(m,n), \quad A = \mathrm{normc}(A).$$

For each randomly generated matrix A, we use ManPG to find an approximate solution $\hat{X}$ of (5.1.1) and (5.1.3), respectively. The reconstructed matrix and its relative reconstruction error (RRE) and percentage of explained variance (PEV) [37] by using $\hat{X}$ are defined by

$$\hat{A} = A\hat{X}(\hat{X}^\mathsf{T}\hat{X})^{-1}\hat{X}^\mathsf{T}, \quad \text{RRE} = \frac{\|A - \hat{A}\|_\text{F}}{\|A\|_\text{F}}, \quad \text{PEV} = \frac{\text{tr}(\hat{A}^\mathsf{T}\hat{A})}{\text{tr}(A^\mathsf{T}A)}(\times 100\%). \quad (5.1.4)$$

In Table 5.1 and Table 5.2, we report the average values of RRE and PEV of $\hat{A}$ by using the randomly generated 20 nonnegative matrices A for each m and n to compare models (5.1.1) and (5.1.3) with $r = 10$. All computed solutions $\hat{X}$ for calculating RRE and PEV in Table 5.1 and Table 5.2, satisfy

$$\|\hat{X}^\mathsf{T}\hat{X} - I_r\|_\text{F} \le 10^{-14} \quad \text{and} \quad \|\hat{X}^\mathsf{T}\hat{X} - I_r\|_\text{F} + \|(S \circ \hat{X})_-\|_{\ell_1} \le 10^{-14},$$

for model (5.1.1) and model (5.1.3), respectively.

$m = 40$, $n = 30$							
λ, μ	$0.6, 150$	$0.6, 170$	$0.6, 190$	$0.6, 200$	$1, 100$	$1,\ 110$	$1,\ 130$
model (5.1.1)	0.4029	0.4029	0.4029	0.4029	0.4046	0.4046	0.4046
model (5.1.3)	0.3999	0.3992	0.3988	0.3953	0.4029	0.4008	0.3981
$\lambda = 0.6$, $\mu = 100$							
m, n	$50, 25$	$50, 50$	$80, 25$	$80, 40$	$80, 80$	$200,\ 25$	$200,\ 50$
model (5.1.1)	0.3811	0.4427	0.3846	0.4315	0.4652	0.3860	0.4464
model (5.1.3)	0.3806	0.4409	0.3843	0.4284	0.4636	0.3847	0.4451

Table 5.1: Comparison on RRE with different (m, n, λ, μ) by randomly generated A

5.1.2 Numerical results using Yale face dataset

The Yale Face dataset contains 165 GIF format gray scale images of 15 individuals with 11 images for each subject, and one for each different facial expression or configuration. From http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html, we download the 165×1024 facial image matrix F_{ace}. The $(15 \times (i-1) + t)$-th row of F_{ace} is the t-th image of the i-th person, with $i = 1, \ldots, 15$ and $t = 1, \ldots, 11$.

$m = 40,\ n = 30$							
λ, μ	$0.6, 150$	$0.6, 170$	$0.6, 190$	$0.6, 200$	$1, 100$	$1,\ 110$	$1,\ 130$
model (5.1.1)	0.8376	0.8376	0.8376	0.8376	0.8363	0.8363	0.8363
model (5.1.3)	0.8400	0.8406	0.8410	0.8410	0.8376	0.8391	0.8404
$\lambda = 0.6,\ \mu = 100$							
m, n	$50, 25$	$50, 50$	$80, 25$	$80, 40$	$80, 80$	$200, 25$	$200, 50$
model (5.1.1)	0.8547	0.8040	0.8520	0.8138	0.7836	0.8510	0.8007
model (5.1.3)	0.8551	0.8055	0.8523	0.8164	0.7850	0.8520	0.8019

Table 5.2: Comparison on PEV with different (m, n, λ, μ) by randomly generated A

Each row of F_{ace} defines a 32×32 nonnegative matrix. We use the all rows of F_{ace}, which include 11 images of all people, to get 165 32×32 nonnegative matrices and then use Matlab function normc to normalize each of these matrices.

For each 32×32 matrix A, we use ManPG to find an approximate solution $\hat{X}$ of (5.1.1) and (5.1.3), respectively. We compute the reconstructed matrix $\hat{A}$ and its RRE and PEV by using computed $\hat{X}$ as (5.1.4).

From Figure 5.1 to Figure 5.6, we can see that in almost every case, the reconstructed matrix $\hat{A}$ by model (5.1.3) has lower values RRE and higher values PEV than that computed by model (5.1.1) without sacrificing the sparsity of solutions. In our numerical experiments, we only restricted the power of the penalty term to be one, but did not restrict the penalty parameter $\mu > 5Lr^{\frac{3}{4}}$.

5.2 Projection to $\mathbb{S}_+^{n,r}$

Consider the following problem that finds the projection of a given matrix $C \in R^{n\times r}$ onto $\mathbb{S}_+^{n,r}$

$$\begin{aligned} \min \quad & \frac{1}{2}\|X - C\|_{\mathrm{F}}^2 \\ \text{s.t.} \quad & X^{\mathsf{T}}X = I_r \\ & X \geq 0. \end{aligned} \tag{5.2.1}$$

Define gap $\coloneqq$ $|\|\bar{X} - C\|_{\mathrm{F}}/\|X^* - C\|_{\mathrm{F}} - 1|$, where $\bar{X}$ is the solution found by

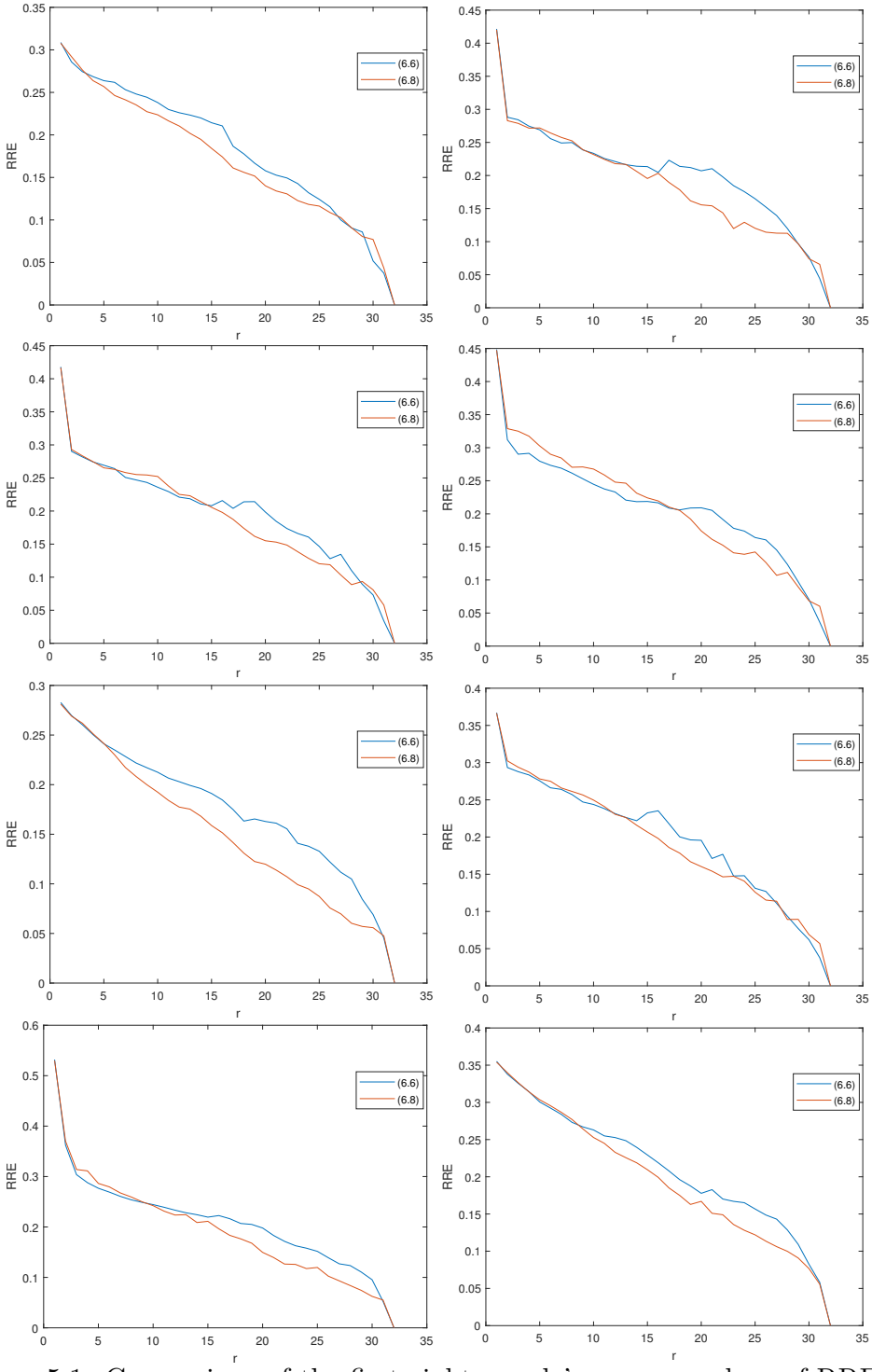

Figure 5.1: Comparison of the first eight people's average values of RRE by models (5.1.1) with $\lambda = 1$ and (5.1.3) with $\lambda = 1, \mu = 6$, respectively, for $r = 1, \ldots, 32$.

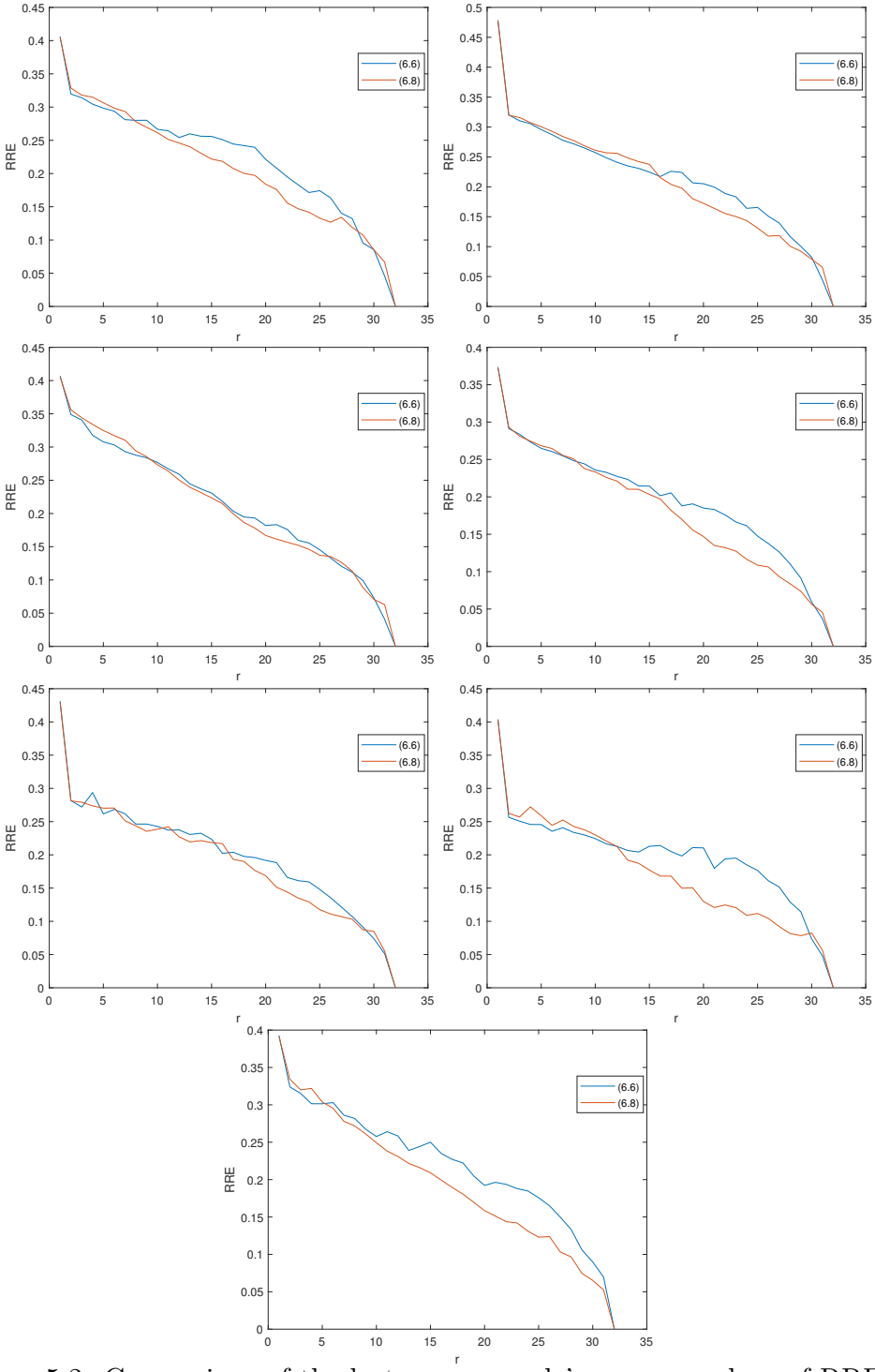

Figure 5.2: Comparison of the last seven people's average values of RRE by models (5.1.1) with $\lambda = 1$ and (5.1.3) with $\lambda = 1, \mu = 6$, respectively, for $r = 1, \ldots, 32$.

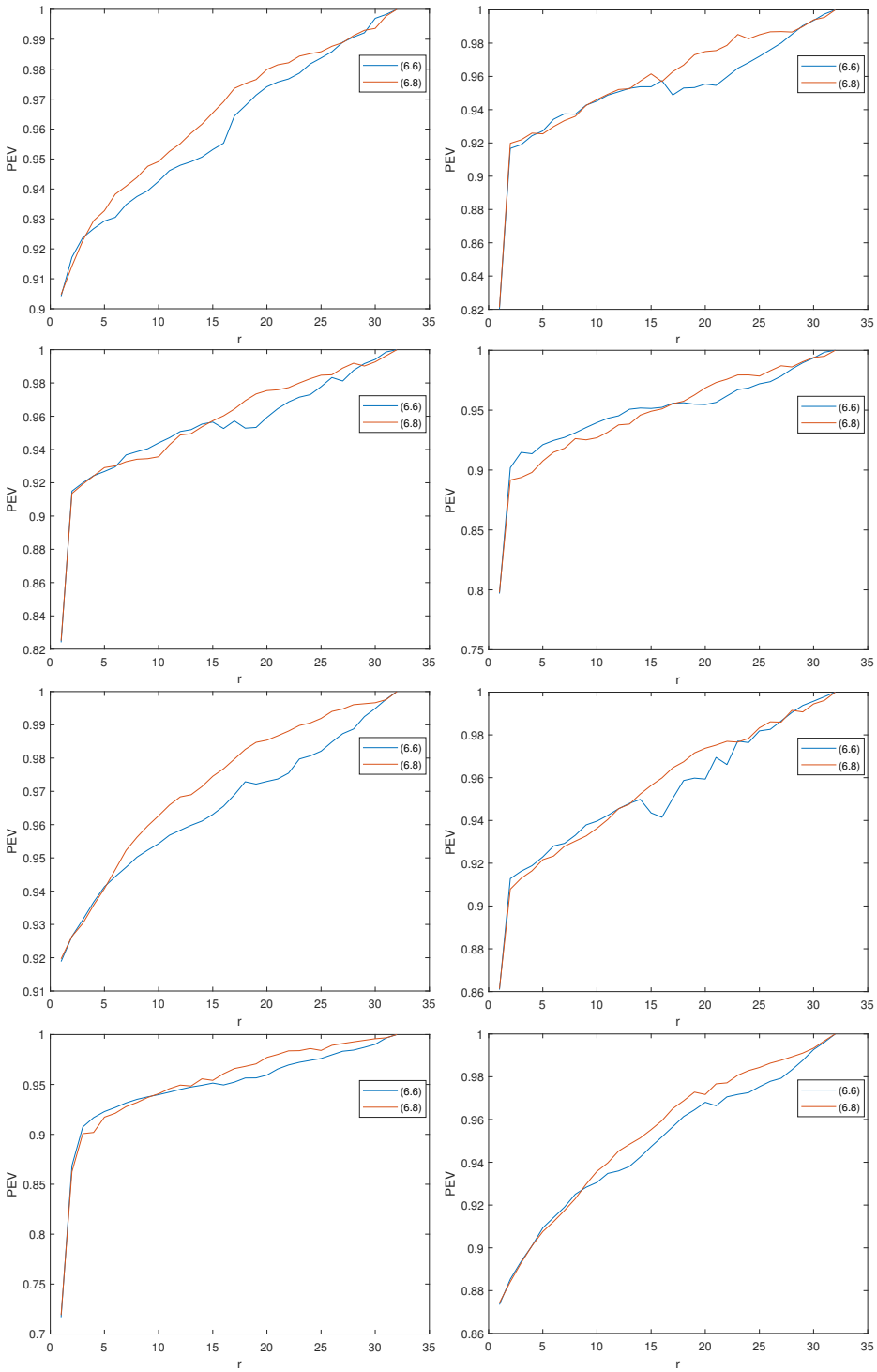

Figure 5.3: Comparison of the first eight people's average values of PEV by models (5.1.1) with $\lambda = 1$ and (5.1.3) with $\lambda = 1, \mu = 6$, respectively, for $r = 1, \ldots, 32$.

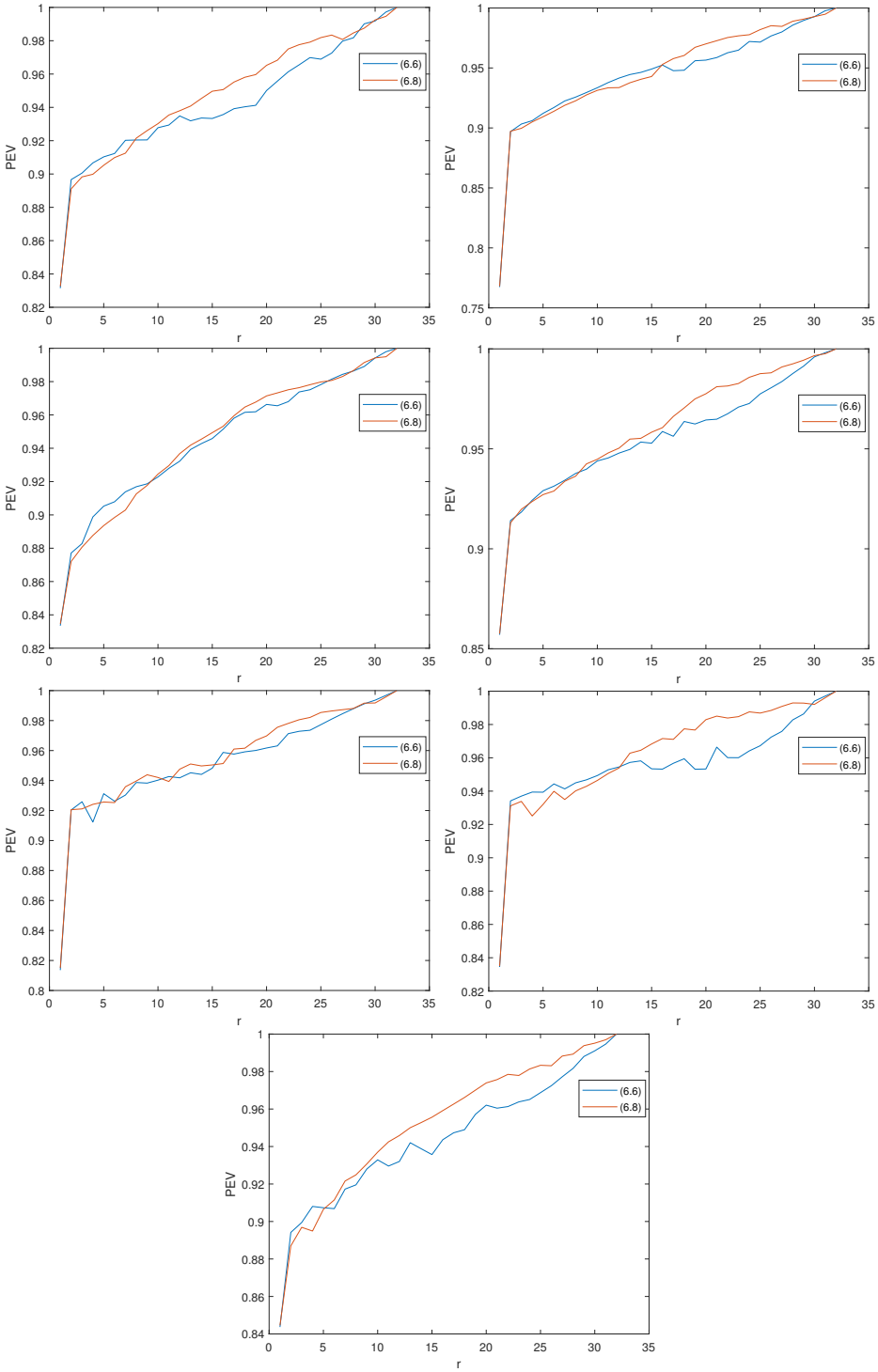

Figure 5.4: Comparison of the last seven people's average values of PEV by models (5.1.1) with $\lambda = 1$ and (5.1.3) with $\lambda = 1, \mu = 6$, respectively, for $r = 1, \ldots, 32$.

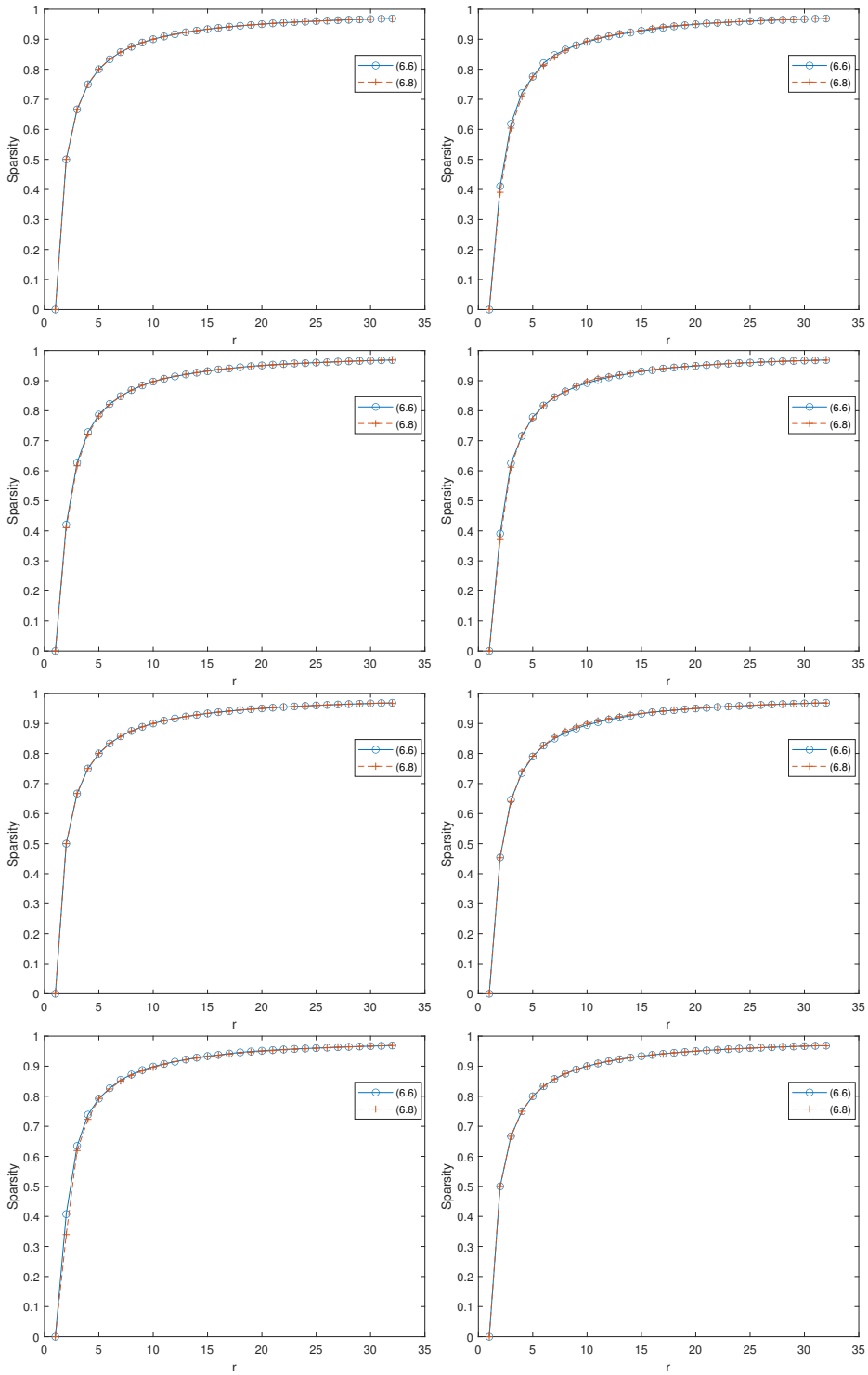

Figure 5.5: Comparison of the first eight people's average values of sparsity by models (5.1.1) with $\lambda = 1$ and (5.1.3) with $\lambda = 1, \mu = 6$, respectively, for $r = 1, \ldots, 32$.

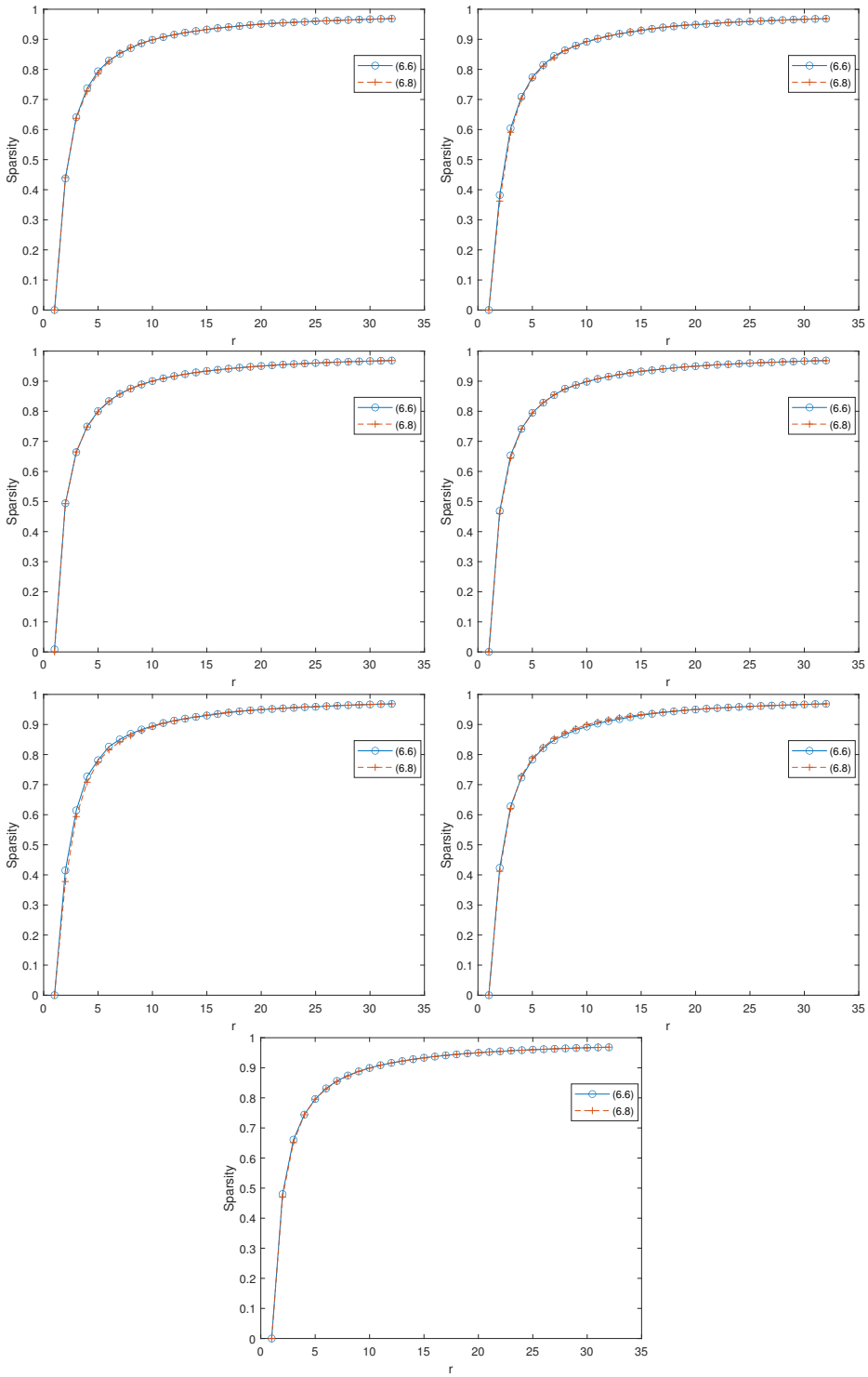

Figure 5.6: Comparison of the last seven people's average values of sparsity by models (5.1.1) with $\lambda = 1$ and (5.1.3) with $\lambda = 1, \mu = 6$, respectively, for $r = 1, \ldots, 32$.

PIRL2 and X^* is the optimal point of the test problem, the initial point X^0 is same as EP4Orth+ [14].

We use PenCF [33] to solve the subproblem (4.4.8) and the code of PenCF can be downloaded from this website https://portrait.gitee.com/stmopt/stop.

When the algorithm is completed, we adopt the same post-procedure in [14, Algorithm 4.1] to improve the quality of the solutions. The parameter ξ in the following tables controls the noise level when generating matrix C, the larger the ξ, the harder it is to find the projection of C, one can refer [14, Section 6.1] for more detail.

We run 60 times for PIRL2 choose alternating Barzilai–Borwein stepsize, and set the tolerance for KKT violation to 10^{-8}, the maximum number of iterations to 800 in the PenCF's option. We run 50 times for each case and take the average relative gap as the final result.

We use Pencf to solve

$$\begin{aligned} \min \quad & \frac{1}{2}\|X - C\|_{\mathrm{F}}^2 + \lambda \|\max(-X, 0)\|_{\mathrm{F}}^2 \\ \text{s.t.} \quad & X^{\mathsf{T}} X = I_r, \end{aligned} \tag{5.2.2}$$

and use PencfQuad instead of Pencf in Table to avoid ambiguity. Under the same framework of Algorithm 1, we change

$$P_\varepsilon^k(X) \ := \ \sum_{i=1}^{n}\sum_{j=1}^{r} W_{ij}^k(\max(-X_{ij}, 0) + \varepsilon), \tag{5.2.3}$$

where $W_{ij}^k = \frac{1}{2}(\max(-X_{ij}^k, 0) + \varepsilon)^{-\frac{1}{2}}$, the modified algorithm is called PIRL1.

Table 5.3 to Table 5.8 show PIRL2 can improve the results than directly solving quadratic penalty model (5.2.2) by Pencf and ℓ_2 weight performs better than ℓ_1 weight.

For problems with size $n = 2000$, $r = 10$ to $r = 400$, we compare the relative

Table 5.3: The comparison of relative gap among PencfQuad, PIRL2 and PIRL2 for problem size $n = 2000$, $r = 10$

	PencfQuad	PIRL1	PIRL2
$\xi = 0.5$	0	2.43e-02	0
$\xi = 0.7$	0	5.49e-05	0
$\xi = 0.9$	1.63e-05	1.62e-03	**0**
$\xi = 0.95$	1.46-04	1.98e-02	**0**
$\xi = 0.98$	9.88e-05	2.18e-02	**1.36e-05**
$\xi = 1$	1.53e-04	2.31e-02	**8.60e-05**

Table 5.4: The comparison of relative gap among PencfQuad, PIRL1 and PIRL2 for problem size $n = 2000$, $r = 50$

	PencfQuad	PIRL1	PIRL2
$\xi = 0.5$	0	6.12e-03	0
$\xi = 0.7$	3.66e-05	1.36e-02	**0**
$\xi = 0.9$	8.69e-04	2.09e-02	**0**
$\xi = 0.95$	8.42e-04	2.18e-02	**1.4e-04**
$\xi = 0.98$	8.38e-04	2.24e-02	**2.4e-04**
$\xi = 1$	6.93e-04	2.27e-02	**2.3e-04**

gap computed by PIRL2 and EP4Orth+ from Table 5.9 to Table 5.14. PIRL2 performs better than EP4Orth+ when $r \leq 100$ while EP4Orth+ performs better as the problem size increases. Generally, PIRL2 exhibits superior performance in relatively small-scale cases, whereas EP4Orth+ demonstrates better performance in relatively large-scale cases.

In Table 5.15, PIRL2 gives the lowest relative gap among the three algorithms

Table 5.5: The comparison of relative gap among PencfQuad, PIRL1 and PIRL2 for problem size $n = 2000$, $r = 100$

	PencfQuad	PIRL1	PIRL2
$\xi = 0.5$	0	1.6e-02	0
$\xi = 0.7$	5.07e-04	2.3e-02	**0**
$\xi = 0.9$	1.12e-03	2.0e-02	**0**
$\xi = 0.95$	9.50e-04	1.9e-02	**2.7e-04**
$\xi = 0.98$	7.77e-04	2.0e-02	**2.6e-04**
$\xi = 1$	5.73e-04	2.0e-02	**2.1e-04**

Table 5.6: The comparison of relative gap among PencfQuad, PIRL1 and PIRL2 for problem size $n = 2000$, $r = 200$

	PencfQuad	PIRL1	PIRL2
$\xi = 0.5$	0	2.0e-03	0
$\xi = 0.7$	4.14e-04	1.0e-02	**0**
$\xi = 0.9$	6.91e-04	1.4e-02	**0**
$\xi = 0.95$	5.31e-04	1.4e-02	**3.5e-04**
$\xi = 0.98$	4.17e-04	1.5e-02	**3.1e-04**
$\xi = 1$	3.05e-04	1.5e-02	**2.3e-04**

Table 5.7: The comparison of relative gap among PencfQuad, PIRL1 and PIRL2 for problem size $n = 2000$, $r = 300$

	PencfQuad	PIRL1	PIRL2
$\xi = 0.5$	0	1.4e-02	0
$\xi = 0.7$	6.22e-04	1.4e-02	**0**
$\xi = 0.9$	6.85e-04	1.4e-02	**0**
$\xi = 0.95$	5.58e-04	4.3e-02	**4.2e-04**
$\xi = 0.98$	4.28e-04	4.2e-02	**3.4e-04**
$\xi = 1$	3.62e-04	4.1e-02	**2.7e-04**

but consumes more time than SEPPG and EP4Orth+.

5.3 Quadratic assignment problem

The quadratic assignment problem (QAP) was first proposed by Koopmans and Beckmann (1957) [15]. Consider a_{ij} the flow from i-th facility to j-th facility, and b_{ij} the distant between location k and location l, the mathematical expression of QAP

Table 5.8: The comparison of relative gap among PencfQuad, PIRL1 and PIRL2 for problem size $n = 2000$, $r = 400$

	PencfQuad	PIRL1	PIRL2
$\xi = 0.5$	1.35e-05	1.6e-02	**0**
$\xi = 0.7$	6.73e-04	1.4e-02	**0**
$\xi = 0.9$	7.13e-04	1.3e-02	**0**
$\xi = 0.95$	5.80e-04	3.8e-02	**4.7e-04**
$\xi = 0.98$	5.05e-04	3.7e-02	**3.8e-04**
$\xi = 1$	4.11e-04	3.6e-02	**3.1-04**

Table 5.9: The comparison of relative gap between PIRL2 and EP4Orth+ for problem size $n = 2000$, $r = 10$

	PIRL2	EP4Orth+
	gap	gap
$\xi = 0.5$	0	0
$\xi = 0.7$	0	0
$\xi = 0.9$	0	0
$\xi = 0.95$	**0**	7.2e-05
$\xi = 0.98$	**1.36e-05**	8.9e-4
$\xi = 1$	**8.60e-05**	1.2e-3

Table 5.10: The comparison of relative gap between PIRL2 and EP4Orth+ for problem size $n = 2000$, $r = 50$

	PIRL2	EP4Orth+
	gap	gap
$\xi = 0.5$	0	0
$\xi = 0.7$	0	0
$\xi = 0.9$	0	0
$\xi = 0.95$	**1.4e-04**	2.1e-04
$\xi = 0.98$	**2.4e-04**	5.0e-4
$\xi = 1$	**2.3e-04**	2.6e-3

Table 5.11: The comparison of relative gap between PIRL2 and EP4Orth+ for problem size $n = 2000$, $r = 100$

	PIRL2	EP4Orth+
	gap	gap
$\xi = 0.5$	0	0
$\xi = 0.7$	0	0
$\xi = 0.9$	0	0
$\xi = 0.95$	2.72e-04	**6.6e-07**
$\xi = 0.98$	**2.67e-04**	8.0e-4
$\xi = 1$	**2.16e-04**	2.6e-3

Table 5.12: The comparison of relative gap between PIRL2 and EP4Orth+ for problem size $n = 2000$, $r = 200$

	PIRL2	EP4Orth+
	gap	gap
$\xi = 0.5$	0	0
$\xi = 0.7$	0	0
$\xi = 0.9$	0	0
$\xi = 0.95$	3.59e-04	**0**
$\xi = 0.98$	**3.19e-04**	4.5e-4
$\xi = 1$	**2.36e-04**	1.9e-3

Table 5.13: The comparison of relative gap between PIRL2 and EP4Orth+ for problem size $n = 2000$, $r = 300$

	PIRL2	EP4Orth+
	gap	gap
$\xi = 0.5$	0	0
$\xi = 0.7$	0	0
$\xi = 0.9$	0	0
$\xi = 0.95$	4.21e-04	**0**
$\xi = 0.98$	3.45e-04	**2.5e-4**
$\xi = 1$	**2.79e-04**	1.8e-3

Table 5.14: The comparison of relative gap between PIRL2 and EP4Orth+ for problem size $n = 2000$, $r = 400$

	PIRL2	EP4Orth+
	gap	gap
$\xi = 0.5$	0	0
$\xi = 0.7$	0	0
$\xi = 0.9$	0	0
$\xi = 0.95$	4.73e-04	**0**
$\xi = 0.98$	3.85e-04	**1.7e-4**
$\xi = 1$	**3.11e-04**	1.6e-3

Table 5.15: Comparison of relative gap, infeasibility and computation time among PIRL2, SEPPG and EP4Orth+ for the projection problems onto $\mathbb{S}_+^{n,r}$

(ξ, n, r)	PIRL2			SEPPG			EP4Orh+		
	Relgap	Infeas	Time	Relgap	Infeas	Time	Relgap	Infeas	Time
(0.5,500,5)	**0**	0	3.6	3.80e-09	4.10e-08	0.1	0	0	0
(0.8,500,5)	**0**	0	4.9	4.10e-06	6.30e-08	0.2	0	0	0.1
(1,500,5)	**0**	0	4.6	2.60e-05	5.10e-08	0.6	2.80e-04	0	0.1
(0.5,1000,10)	**0**	0	13.7	3.50e-09	5.60e-08	0.6	0	0	0.3
(0.8,1000,10)	**0**	0	13.3	3.20e-06	5.30e-08	3.4	0	0	0.5
(1,1000,10)	**3.70e-05**	0	14	1.10e-03	0	10.1	4.10e-04	0	0.5

is

$$\begin{aligned} \min \quad & \sum_{i,j=1}^{n} \sum_{k,l=1}^{n} a_{ij} b_{kl} x_{ik} x_{jl} \\ \text{s.t.} \quad & \sum_{i=1}^{n} x_{ij} = 1 \quad 1 \le j \le n, \\ & \sum_{j=1}^{n} x_{ij} = 1 \quad 1 \le i \le n, \\ & x_{ij} \in \{0, 1\} \quad 1 \le i, j \le n. \end{aligned} \tag{5.3.1}$$

Let $A = (a_{ij})$, $B = (b_{ij})$ and $X = (x_{ij})$, by the definition of a_{ij} and b_{kl}, we know A, B are nonnegative. Notice X is a permutation matrix in QAP, together with the trace operator, the QAP can be written in the following matrix form

$$\begin{aligned} \min \quad & \langle A, XBX^{\mathsf{T}} \rangle \\ \text{s.t.} \quad & X^{\mathsf{T}} X = I_n, \\ & X \in \mathbb{R}_+^{n \times n}. \end{aligned} \tag{5.3.2}$$

Due to the combinatorial property, QAP is an NP-hard problem and a hard problem when $n > 15$. Moreover, we adopt the same reformulation of (5.3.2) as (5.4) in [27],

i.e.,

$$\begin{aligned} \min \quad & \langle A, (X \circ X) B (X \circ X)^{\mathsf{T}} \rangle \\ \text{s.t.} \quad & X^{\mathsf{T}} X = I_n, \\ & X \in \mathbb{R}_+^{n \times n}. \end{aligned} \tag{5.3.3}$$

We use the 133 instants in QAPLIB to test PIRL2, and compare its performance with SEPPG [27] and EP4Orth+ [14]. Since the matrix A in "**esc16f**" is zero matrix, the optimal solutions or bounds in "**tai10a**" and "**tai10b**" are not provided, we exclude the three cases in the numerical tests. The subprobem (4.4.8) in PIRL2 is still solved by Pencf. One can find the data in QAPLIB via the link https://coral.ise.lehigh.edu/data-sets/qaplib/qaplib-problem-instances-and-solutions/. Define the violation of nonnegativity

$$\textbf{Ninf} \coloneqq \| \max(-X, 0) \|_{l_1}, \tag{5.3.4}$$

and the relative gap

$$\textbf{relgap} \coloneqq \left[\frac{f(\bar{X}) - \text{Best}}{\text{Best}} \times 100 \right] \%, \tag{5.3.5}$$

where $\bar{X}$ is the solution obtained by PIRL2 and Best is the optimal objective value or the lower bounds provided by QAPLIB. The relative gap will be used for measuring the efficiency of different algorithms. SEPPG contains two versions of algorithms SEPPG+ and SEPPG0, we use the better result between them as the result of SEPPG. The initial points of PIRL2 are generated randomly by the **MATLAB** functions,

$$\tilde{X}^0 = \text{randn}(n, r), \quad X^0 = \text{orth}(\tilde{X}^0), \tag{5.3.6}$$

in the same way as the initial points are generated in Section 5.3 in [27]. The maximum number of iterations in Pencf is set to 500, while the number of outer loop iterations in PIRL2 is set to 20.

Since the initial points are generated randomly, we run 100 times PIRL2 for each case and report the minimum and median relative gap of 133 instances in Table 5.16 and Table 5.17. When the relative gap equals 0, it means that our algorithm finds the optimal solution or better lower bound. Due to the exact penalty property of our model and the post-procedure, the solutions found by PIRL2 always lie in $\mathbb{S}_+^{n,n}$, therefore, the feasibility of the solutions is not reported in Table 5.16 and Table 5.17.

In Table 5.18, we compare relative gap, violation of nonnegativity and computation time among PIRL2, SEPPG and EP4Orth+ for 21 QAPLIB cases with $n \geq 80$. PIRL2 obtains lower relative gap on "**sko**", "**tai80a**" , "**tai256c**"and"**tho150**" than SEPPG and EP4Orth+, but consumes more time than SEPPG. For the left cases, SEPPG achieves the lowest relative gap.

From Table 5.19 and Table 5.20, the performance of SEPPG is the best among the three algorithms, followed by PIRL2.

Table 5.16: Relative gap calculated by PIRL2 algorithm from **bur26a** to **nug15**

Name	Min_gap	Median_gap	Name	Min_gap	Median_gap
bur26a	0.37	1.18	esc32c	0	5.919
bur26b	0.3427	1.2281	esc32d	1	13.9487
bur26c	0.1759	1.2076	esc32e	0	0
bur26d	0.003	1.283	esc32g	0	33.3333
bur26e	0.2065	1.1579	esc32h	0.6018	6.1817
bur26f	0.0894	1.2574	esc64a	0	4.3898
bur26g	0.2908	1.6345	esc128	0	21.875
bur26h	0.3223	1.8002	had12	0.6053	6.4165
chr12a	21.8802	73.361	had14	0.0185	5.4104
chr12b	9.1768	99.4798	had16	1.9355	5.8065
chr12c	18.806	86.5005	had18	1.4184	5.2072
chr15a	23.848	133.7692	had20	1.7914	5.1719
chr15b	44.9812	148.6859	kra30a	0	0
chr15c	27.8199	141.8245	kra30b	0	0
chr18a	60.6596	171.8418	kra32	6.4374	13.3897
chr18b	14.2112	43.1551	lipa20a	1.6527	4.2357
chr20a	44.3431	118.1113	lipa20b	0	22.3741
chr20b	37.859	111.9669	lipa30a	0.6643	2.6764
chr20c	47.6453	204.8084	lipa30b	12.3717	21.3824
chr22a	15.757	43.2099	lipa40a	0.798	2.0971
chr22b	17.7591	46.1737	lipa40b	7.5861	22.7184
chr25a	58.6407	136.2094	lipa50a	0.5698	1.853
els16	0	0	lipa50b	12.5726	22.351
esc16a	0	17.6471	lipa60a	1.6341	1.9162
esc16b	0	0.6849	lipa60b	20.4867	23.8722
esc16c	0	6.5202	lipa70a	0.471	1.1192
esc16d	0	0.9984	lipa70b	22.11	22.8882
esc16e	0.026	21.4286	lipa80a	1.1785	1.4157
esc16g	0	16.376	lipa80b	21.0963	23.5712
esc16h	0	0	lipa90a	0.9223	1.1933
esc16i	0	0	lipa90b	0	24.642
esc16j	0	9.9287	nug12	0.126	9.9958
esc32a	23.0769	42.259	nug14	2.9586	14.6943
esc32b	23.8095	45.2381	nug15	4.1739	15.3913

Table 5.17: Relative gap calculated by PIRL2 algorithm from **nug16a** to **wil100**

Name	Min_gap	Median_gap	Name	Min_gap	Median_gap
nug16a	1.2865	11.9278	tai15a	3.2848	10.6269
nug16b	1.1272	12.1965	tai15b	0.6879	1.7339
nug17	2.2943	10.9122	tai17b	2.6978	7.216
nug18	3.1612	12.5389	tai20a	5.7844	12.3436
nug20	5.1362	13.3337	tai20b	3.4	8.9205
nug21	4.5119	16.0788	tai25a	5.0486	9.0112
nug22	4.505	13.7653	tai25b	3.2056	16.9442
nug24	6.3073	14.7015	tai30a	4.6263	7.7533
nug25	5.9497	12.8324	tai30b	1.5413	17.3936
nug27	3.5921	9.7822	tai35a	3.9346	9.4566
nug28	5.6523	11.9628	tai35b	2.1468	12.1028
nug30	5.454	9.03	tai40a	0	8.9766
rou12	5.1272	14.3613	tai40b	3.6644	11.8569
rou15	2.6402	16.3776	tai50a	6.6278	8.6116
rou20	6.3585	11.908	tai50b	2.1821	9.1281
scr12	8.1312	32.9322	tai60a	0	0
scr15	5.6332	23.1776	tai60b	1.3827	9.0733
scr20	4.277	24.0246	tai64c	0.5412	4.0924
sko42	3.1363	5.8548	tai80a	0	0
sko49	2.5531	5.0757	tai80b	3.6573	7.5264
sko56	3.941	6.3672	tai100a	4.2627	5.9205
sko64	0	0	tai100b	2.7349	6.6633
sko72	0	0	tai150b	1.9384	4.2301
sko81	0	0	tai256c	0.6928	1.218
sko90	0	0	ste36a	8.6041	21.8551
sko100a	0	0	ste36b	9.0242	28.2887
sko100b	0	0	ste36c	0	0
sko100c	0	0	tho30	0	0
sko100d	0	0	tho40	3.7279	8.0257
sko100e	0	0	tho150	0	0
sko100f	0	0	wil50	1.2537	2.6868
tai12a	1.7161	12.656	wil100	1.1603	1.9671
tai12b	0	13.5227			

Table 5.18: Comparison of the relative gap, violation of nonnegativity and computation time among PIRL2, SEPPG and EP4Orth+ for 21 QAPLIB cases with $n \geq 80$

Name	PIRL2			SEPPG			EP4Orth+	
	Median_gap	Time	Ninf	Median_gap	Time	Ninf	Median_gap	Time
esc128	21.875	27.2	0	**4.156**	73.9	3.40e-06	194	56
lipa80a	1.415	10.2	0	**0.778**	17.7	1.60e-06	2.028	0.7
lipa80b	23.571	10.4	0	**14.891**	1.6	0	27.207	22.9
lipa90a	1.193	12.2	0	**0.703**	18.6	6.10e-07	1.82	1
lipa90b	24.642	12.3	0	**13.541**	1.8	0	27.7	23.8
sko81	**0**	16.1	0	1.618	5.1	0	13.24	33
sko90	**0**	18.9	0	1.637	5.7	0	12.87	38.7
sko100a	**0**	24.1	0	1.426	7.2	0	12.773	51.7
sko100b	**0**	24.1	0	1.451	7.1	0	12.29	51
sko100c	**0**	22.3	0	1.622	7.5	0	13.234	51.6
sko100d	**0**	25.1	0	1.471	7	0	12.54	51.7
sko100e	**0**	30.7	0	1.66	7.2	0	13.478	52.5
sko100f	**0**	21.8	0	1.464	7.3	0	12.25	51.7
tai80a	**0**	11.8	0	3.02	1.3	0	9.481	22.3
tai80b	7.526	12.7	0	**4.58**	48.4	0	33.349	34.7
tai100a	5.920	16.1	0	**2.737**	2.1	0	8.72	29.9
tai100b	6.663	20.3	0	**3.908**	76.9	0	39.591	58.5
tai150b	4.230	36.9	0	**2.991**	35	0	22.58	154.6
tai256c	1.218	280.2	0	**1.216**	32.7	0	203.637	317.5
tho150	**0**	32.4	0	1.885	12.2	0	15.948	147.2
wil100	1.967	22.8	0	**0.686**	8.3	0	8.162	52.6

Table 5.19: Level of the minimum relative gaps on the 133 QAPLIB instances

Min_gap($\leq$ %)	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	2	3	4
PIRL2	35	39	41	43	46	47	49	54	55	55	56	71	79	90
SEPPG	**46**	**55**	**62**	**64**	**70**	**75**	**79**	**89**	**93**	**97**	**98**	**112**	**119**	**121**
EP4Orth+	6	6	6	7	7	9	9	11	14	14	14	18	22	25

Table 5.20: Level of the median relative gaps on the 133 QAPLIB instances

Median_gap($\leq$ %)	0.3	0.5	0.7	1	3	5	7	10	15	20	25	30	40
PIRL2	**21**	**21**	**21**	24	42	46	59	74	94	101	114	115	117
SEPPG	4	9	17	**25**	**58**	**86**	**96**	**102**	**114**	**117**	**121**	**122**	**122**
EP4Orth+	0	0	0	0	10	15	17	23	44	53	73	84	91

Chapter 6

Conclusions and future work

6.1 Conclusions

In this thesis, we solve the optimization problem over the sign-constrained Stiefel manifold by establishing the exact penalty model and designing an efficient algorithm.

In the first part, we present the error bounds (1.1.2)–(1.1.4) with explicit values of ν and q in Theorems 3.1, 3.2 and 3.5 for $\mathbb{S}_s^{n,r} = \mathbb{S}_+^{n,r}$. Furthermore, we show that these error bounds cannot hold with $q > 1/2$ when $1 < r < n$ in Proposition 3.3, and point out that they cannot hold with $q > 1$ for any $r \in \{1, \ldots, n\}$ in Remark 3.1. In Chapter 4 we present the error bounds (1.1.2)–(1.1.4) with explicit values of ν and q in Theorems 3.7–3.9 for the sign-constrained Stiefel manifold. The exponent q in the error bounds is $1/2$ for any $r \in \{1, \ldots, n\}$ and can take the value 1 for $|\mathcal{P}| + |\mathcal{N}| \in \{1, n\}$. The new error bounds help us to establish the exactness of penalty problems (1.1.6)–(1.1.8) for problem (1.1.5). Compared with existing results on error bounds for the set $\mathbb{S}_+^{n,r}$ and penalty methods for minimization with nonnegative orthogonality constraints, our results have explicit values of the error bound parameters and penalty parameters, and do not need any condition other than the (local) Lipschitz continuity of the objective function for the exact penalty. Moreover, exponents in our error bounds are independent of the dimension of the

underlying space.

In the second part, we propose a proximal iteratively reweighted ℓ_2 algorithm to solve the nondifferentiable non-Lipschitz penalized problem. Under the assumption that the objective function in the original problem is continuous, our algorithm has subsequence convergence property, a sufficient decrease in each iteration, and the distance between two adjacent iteration points is square summable, any accumulation point is a stationary point. In numerical parts, we use projection to $\mathbb{S}_+^{n,r}$ and QAP two problems to test the PIRL2 algorithm and compare its performance with SEPPG and EP4Orth+. Extensive numerical examples show the effectiveness of the PIRL2 algorithm.

6.2 Future work

In this thesis, the sign constraint is posed on each column of the matrix in the Stiefel manifold, we wish to extend the sign constraint to each component of the matrix, but the significance of the component sign constraint remains to be explored. The fraction exponent in the error bounds leads to the non-Lipschitz of the penalty term, we ever changed the exponent of the penalty term to 1 in some numerical experiments and found that it had little effect on the results, so we guess there may exist some conditions to improve the exponents in these error bounds. In the framework of PIRL2, the accuracy and speed of solving subproblems greatly affect the effectiveness of the entire algorithm, moreover, in this thesis, we obtain the first-order stationary point of the subproblem with the help of the Pencf algorithm, so we hope to design a better algorithm to solve this subproblem in the future.

Chapter 7

Appendix. Proofs of Theorems 4.1 and 4.2

We first present the following lemma on a simple inequality between the entry-wise ℓ_p-norm and the Frobenius norm. Its proof is elementary and hence omitted.

Lemma 7.1. *For any $X \in \mathbb{R}^{n\times r}$ and any $p \geq 1$,*

$$\|X\|_{\mathrm{F}} \;\leq\; \max\left\{1,\; (nr)^{\frac{p-2}{2p}}\right\} \|X\|_{\ell_p}.$$

The proofs of Theorems 4.1 and 4.2 are as follows.

Proof of Theorem 4.1. Define $h(X) = \|X_-\|_{\ell_p}^q + \|X^{\mathsf{T}}X - I_r\|_{\ell_p}^{\frac{1}{2}}$ for $X \in \mathcal{S}$, and set $\nu = 5r^{\frac{3}{4}} \max\left\{1,\, (nr)^{\frac{p-2}{4p}}\right\}$. By (3.4.6) and Lemma 7.1, we have

$$\mathrm{dist}(X,\, \mathbb{S}_+^{n,r}) \;\leq\; 5r^{\frac{3}{4}} \left(\|X_-\|_{\mathrm{F}}^q + \|X^{\mathsf{T}}X - I_r\|_{\mathrm{F}}^{\frac{1}{2}}\right) \;\leq\; \nu h(X) \quad \text{for} \quad X \in \mathcal{S}.$$

For any $X \in \mathcal{S}$, setting $\bar{X}$ to a projection of X onto $\mathbb{S}_+^{n,r}$, and combining the L-Lipschitz continuity of F with the above error bound, we have

$$F(\bar{X}) \;\leq\; F(X) + L\,\mathrm{dist}(X,\, \mathbb{S}_+^{n,r}) \;\leq\; F(X) + \mu h(X).$$

This implies that

$$\inf\{F(X) : X \in \mathbb{S}_+^{n,r}\} \;\leq\; \inf\{F(X) + \mu h(X) : X \in \mathcal{S}\}.$$

Meanwhile, $\inf\{F(X) : X \in \mathbb{S}_+^{n,r}\} \ \geq \ \inf\{F(X) + \mu h(X) : X \in \mathcal{S}\}$ as h is zero on $\mathbb{S}_+^{n,r} \subset \mathcal{S}$. Thus

$$\inf\{F(X) : X \in \mathbb{S}_+^{n,r}\} \ = \ \inf\{F(X) + \mu h(X) : X \in \mathcal{S}\}. \tag{7.0.1}$$

For any $X^* \in \mathrm{Argmin}\{F(X) : X \in \mathbb{S}_+^{n,r}\}$, we have $h(X^*) = 0$ and

$$F(X^*) + \mu h(X^*) \ = \ F(X^*) \ = \ \inf\{F(X) : X \in \mathbb{S}_+^{n,r}\},$$

which together with (7.0.1) ensures $X^* \in \mathrm{Argmin}\{F(X) + \mu h(X) : X \in \mathcal{S}\}$.

Now take any $X^* \in \mathrm{Argmin}\{F(X) + \mu h(X) : X \in \mathcal{S}\}$, and let $\bar{X}^*$ be a projection of X^* onto $\mathbb{S}_+^{n,r}$. Then we have

$$F(X^*) + \mu h(X^*) \ \leq \ F(\bar{X}^*) + \mu h(\bar{X}^*) \ = \ F(\bar{X}^*) \ \leq \ F(X^*) + \nu L h(X^*).$$

This leads to $h(X^*) = 0$, as $\mu > \nu L$ and $h(X^*) \geq 0$. Hence X^* lies in $\mathbb{S}_+^{n,r}$, and

$$F(X^*) \ = \ F(X^*) + \mu h(X^*) \ = \ \inf\{F(X) + \mu h(X) : x \in \mathcal{S}\},$$

which implies that $X^* \in \mathrm{Argmin}\{F(X) : X \in \mathbb{S}_+^{n,r}\}$ with the help of (7.0.1). We complete the proof. □

Proof of Theorem 4.2. Define $h(X) = \|X_-\|_{\ell_p}^{q_1} + \|X^{\mathsf{T}}X - I_r\|_{\ell_p}^{q_2}$ for $X \in \mathcal{S}$, and set $\nu = 4\sqrt{r}\max\left\{1, (nr)^{\frac{q_1(p-2)}{2p}}, r^{\frac{q_2(p-2)}{p}}\right\}$. For any $X \in \mathbb{S}_+^{n,r}$ and any $Y \in \mathcal{S}$ such that $\|Y - X\|_{\mathrm{F}} < 1/(6\sqrt{r})$, we have

$$\|Y_-\|_{\mathrm{F}} + \|\sigma(Y) - \mathbf{1}\|_2 \ \leq \ \|Y - X\|_{\mathrm{F}} + \|\sigma(Y) - \sigma(X)\|_2 \ \leq \ 2\|Y - X\|_{\mathrm{F}} \ < \ \frac{1}{3\sqrt{r}},$$

where the first inequality is because $X_- = 0$ and $\sigma(X) - \mathbf{1} = 0$, while the second invokes Lemma 2.2. Hence (3.3.15) and Lemma 7.1 yield

$$\begin{aligned}
&\mathrm{dist}(Y, \mathbb{S}_+^{n,r}) \\
&\leq 4\sqrt{r}\left(\|Y_-\|_{\mathrm{F}}^{q_1} + \|Y^{\mathsf{T}}Y - I_r\|_{\mathrm{F}}^{q_2}\right) \\
&\leq 4\sqrt{r}\left(\max\left\{1, (nr)^{\frac{q_1(p-2)}{2p}}\right\}\|Y_-\|_{\ell_p}^{q_1} + \max\left\{1, (r^2)^{\frac{q_2(p-2)}{2p}}\right\}\|Y^{\mathsf{T}}Y - I_r\|_{\ell_p}^{q_2}\right) \\
&\leq \nu h(Y).
\end{aligned}$$

Given a local minimizer X^* of F on $\mathbb{S}_+^{n,r}$, there exists a $\delta \in (0, 1/(3\sqrt{r}))$ such that X^* is a global minimizer of F on $\mathbb{S}_+^{n,r} \cap \mathcal{B}(X^*, \delta)$ and F is L^*-Lipschitz continuous in the same set.

It suffices to demonstrate that X^* is a global minimizer of $F+\mu h$ on $\mathcal{S} \cap \mathcal{B}(X^*, \delta/2)$ for all $\mu > \nu L^*$. Take any point $Y \in \mathcal{S} \cap \mathcal{B}(X^*, \delta/2)$, let $\bar{Y}$ be a projection of Y onto $\mathbb{S}_+^{n,r}$, and note that $\bar{Y}$ lies in $\mathcal{B}(X^*, \delta)$, which is because

$$\|\bar{Y} - X^*\|_{\mathrm{F}} \;\leq\; \|\bar{Y} - Y\|_{\mathrm{F}} + \|Y - X^*\|_{\mathrm{F}} \;\leq\; \|X^* - Y\|_{\mathrm{F}} + \|Y - X^*\|_{\mathrm{F}} \;<\; \delta.$$

Then, using the fact that $h(X^*) = 0$, we have

$$F(X^*) + \mu h(X^*) = F(X^*) \leq F(\bar{Y}) \leq F(Y) + L^* \operatorname{dist}(Y, \mathbb{S}_+^{n,r}) \leq F(Y) + \mu h(Y),$$

which is what we desire.

If X^* is a local minimizer of $F + \mu h$ on $\mathcal{S}$, and X^* happens to lie in $\mathbb{S}_+^{n,r}$, then

$$F(X^*) \;=\; F(X^*) + \mu h(X^*) \;\leq\; F(Y) + \mu h(Y) \;=\; F(Y)$$

for any Y that is close to X^* and located in $\mathbb{S}_+^{n,r} \subset \mathcal{S}$. Hence X^* is also a local minimizer of F on $\mathbb{S}_+^{n,r}$. We complete the proof. □

Bibliography

[1] M. Ahookhosh, L. T. K. Hien, N. Gillis, and P. Patrinos. Multi-block Bregman proximal alternating linearized minimization and its application to orthogonal nonnegative matrix factorization. *Computational Optimization and Applications*, 79:681–715, 2021.

[2] J. Alegría, J. Thunberg, and O. Edfors. Channel orthogonalization with reconfigurable surfaces: General models, theoretical limits, and effective configuration. *arXiv:2403.15165*, 2024.

[3] F. Chen, Y. Yang, L. Xu, T. Zhang, and Y. Zhang. Big-data clustering: K-means or K-indicators? *arXiv:1906.00938*, 2019.

[4] S. Chen, S. Ma, A. M.-C. So, and T. Zhang. Proximal gradient method for nonsmooth optimization over the Stiefel manifold. *SIAM Journal on Optimization*, 30:210–239, 2020.

[5] X. Chen, Y. He, and Z. Zhang. Tight error bounds for the sign-constrained stiefel manifold. *SIAM Journal on Optimization*, 35(1):302–329, 2025.

[6] X. Chen and W. Zhou. Convergence of the reweighted ℓ_1 minimization algorithm for $\ell_2 - \ell_p$ minimization. *Computational Optimization and Applications*, 59:47–61, 2014.

[7] Y. Cui and J.-S. Pang. *Modern Nonconvex Nondifferentiable Optimization*, volume 29 of *MOS-SIAM Series on Optimization*. SIAM, Philadelphia, 2021.

[8] F. Facchinei and J.-S. Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems — Volume I*. Springer, New York, 2003.

[9] K. Fan and A. J. Hoffman. Some metric inequalities in the space of matrices. *Proceedings of the American Mathematical Society*, 6:111–116, 1955.

[10] N. J. Higham. *Functions of Matrices: Theory and Computation*. SIAM, Philadelphia, 2008.

[11] A. J. Hoffman and H. W. Wielandt. The variation of the spectrum of a normal matrix. *Duke Mathematical Journal*, 20:37–39, 1953.

[12] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, Cambridge, 2008.

[13] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, second edition, 2012.

[14] B. Jiang, X. Meng, Z. Wen, and X. Chen. An exact penalty approach for optimization with nonnegative orthogonality constraints. *Mathematical Programming*, 198:855–897, 2023.

[15] T. C. Koopmans and M. Beckmann. Assignment problems and the location of economic activities. *Optimization Methods and Software*, pages 53–76, 1957.

[16] A. S. Lewis. The convex analysis of unitarily invariant matrix functions. *Journal of Convex Analysis*, 2:173–183, 1995.

[17] G. Li, B. S. Mordukhovich, and T. S. Phạm. New fractional error bounds for polynomial systems with applications to Hölderian stability in optimization and spectral theory of tensors. *Mathematical Programming*, 153:333–362, 2015.

[18] C. Liu and N. Boumal. Simple algorithms for optimization on Riemannian manifolds with constraints. *Applied Mathematics & Optimization*, 82:949–981, 2020.

[19] J. Liu, Y. Liu, W.-K. Ma, M. Shao, and A. M.-C. So. Extreme point pursuit–part I: A framework for constant modulus optimization. *IEEE Transactions on Signal Processing*, 72:4541–4556, 2024.

[20] J. Liu, Y. Liu, W.-K. Ma, M. Shao, and A. M.-C. So. Extreme point pursuit–part II: Further error bound analysis and applications. *IEEE Transactions on Signal Processing*, 72:4557–4572, 2024.

[21] D. Luo, C. Ding, H. Huang, and T. Li. Non-negative Laplacian embedding. In W. Wang, H. Kargupta, S. Ranka, P. S. Yu, and X. Wu, editors, *ICDM '09: Proceedings of the 2009 Ninth IEEE International Conference on Data Mining*, pages 337–346. IEEE, 2009.

[22] Z.-Q. Luo and J.-S. Pang. Error bounds for analytic systems and their applications. *Mathematical Programming*, 67:1–28, 1994.

[23] L. Mirsky. Symmetric gauge functions and unitarily invariant norms. *The Quarterly Journal of Mathematics*, 11:50–59, 1960.

[24] K. F. Ng and W. H. Yang. Regularities and their relations to error bounds. *Mathematical Programming*, 99:521–538, 2004.

[25] J.-S. Pang. Error bounds in mathematical programming. *Mathematical Programming*, 79:299–332, 1997.

[26] F. Pompili, N. Gillis, P.-A. Absil, and F. Glineur. Two algorithms for orthogonal nonnegative matrix factorization with application to clustering. *Neurocomputing*, 141:15–25, 2014.

[27] Y. Qian, S. Pan, and L. Xiao. Error bound and exact penalty method for optimization problems with nonnegative orthogonal constraint. *IMA Journal of Numerical Analysis*, 44:120–156, 2024.

[28] R. Rockafellar and R.-B. Wets. *Variational Analysis*. Springer Science & Business Media, 2009.

[29] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970.

[30] J. von Neumann. Some matrix inequalities and metrization of matric-space. *Tomsk University Reviews*, 1:286–300, 1937.

[31] S. Wang, T.-H. Chang, Y. Cui, and J.-S. Pang. Clustering by orthogonal NMF model and non-convex penalty optimization. *IEEE Transactions on Signal Processing*, 69:5273–5288, 2021.

[32] N. Xiao, X. Liu, and Y. Yuan. Exact penalty function for $\ell_{2,1}$ norm minimization over the Stiefel manifold. *SIAM Journal on Optimization*, 31:3097–3126, 2021.

[33] N. Xiao, X. Liu, and Y. Yuan. A class of smooth exact penalty function methods for optimization problems with orthogonality constraints. *Optimization Methods and Software*, 37:1205–1241, 2022.

[34] Y. Yang, Y. Yang, H. T. Shen, Y. Zhang, X. Du, and X. Zhou. Discriminative nonnegative spectral clustering with out-of-sample extension. *IEEE Transactions on Knowledge and Data Engineering*, 25:1760–1771, 2013.

[35] Z. Yang and E. Oja. Linear and nonlinear projective nonnegative matrix factorization. *IEEE Transactions on Neural Networks*, 21:734–749, 2010.

[36] R. Zass and A. Shashua. Nonnegative sparse PCA. In B. Schölkopf, J. Platt, and T. Hofmann, editors, *Advances in Neural Information Processing Systems 19 (NIPS 2006)*, pages 1561–1568. IEEE, 2007.

[37] Q. Zhao, D. Meng, Z. Xu, and C. Gao. A block coordinate descent approach for sparse principal component analysis. *Neurocompting*, 153:180–190, 2015.

[38] Y. Zhou, C. Bao, C. Ding, and J. Zhu. A semismooth Newton based augmented Lagrangian method for nonsmooth optimization on matrix manifolds. *Mathematical Programming*, 201:1–61, 2023.