MATRIX COMPLETION MODELS WITH FIXED BASIS COEFFICIENTS AND RANK REGULARIZED PROBLEMS WITH HARD CONSTRAINTS

MIAO WEIMIN

(M.Sc., UCAS; B.Sc., PKU)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF MATHEMATICS

NATIONAL UNIVERSITY OF SINGAPORE

2013

This thesis is dedicated to

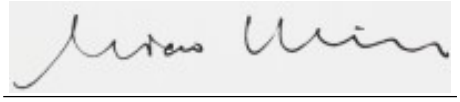
my parents

DECLARATION

I hereby declare that the thesis is my original work and it has been written by me in its entirety.

I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.



Miao Weimin

January 2013

Acknowledgements

I am deeply grateful to Professor Sun Defeng at National University of Singapore for his supervision and guidance over the past five years, who constantly oriented me with promptness and kept offering insightful advice on my research work. His depth of knowledge and wealth of ideas have enriched my mind and broadened my horizons.

I have been privileged to work with Professor Pan Shaohua at South China University of Technology throughout the thesis during her visit at National University of Singapore — her kindness in agreeing to our collaboration and continually making immense contribution in significantly improving our work have spurred a great deal of inspirations.

I am greatly indebted to Professor Yin Hongxia at Minnesota State University, without whom I would not have been in this PhD program. My grateful thanks also go to Professor Liu Yongjin at Shenyang Aerospace University for many fruitful discussions with him on my research topics.

I would like to convey my gratitude to Professor Toh Kim Chuan and Professor Zhao Gongyun at National University of Singapore and Professor Yin Wotao at

Rice University for their valuable comments on my thesis.

I would like to offer special thanks to Dr. Jiang Kaifeng for his generosity in supplying me with impressive understanding and support in coding. I would also like to thank Dr. Ding Chao and Mr. Wu Bin for their helpful suggestions and useful questions on my thesis.

Heartfelt appreciation goes to my dearest friends Zhao Xinyuan, Gu Weijia, Gao Yan, Shi Dongjian, Gong Zheng, Bao Chenglong and Chen Caihua for sharing joy and fun with me in and out mathematics, preserving the years of my PhD study an unforgettable memory of mine.

Lastly, I am tremendously thankful for my parents' care and support all these years; their love and faith in me has nurtured a promising environment that I could always follow my heart and pursue my dreams.

Miao Weimin

(First submission) January 2013

(Final submission) May 2013

Contents

Summary

The problems with embedded low-rank structures arise in diverse areas such as engineering, statistics, quantum information, finance and graph theory. The nuclear norm technique has been widely-used in the literature but its efficiency is not universal. This thesis is devoted to dealing with the low-rank structure via techniques beyond the nuclear norm for achieving better performance.

In the first part, we address low-rank matrix completion problems with fixed basis coefficients, which include the low-rank correlation matrix completion in various fields such as the financial market and the low-rank density matrix completion from the quantum state tomography. For this class of problems, with a reasonable initial estimator, we propose a rank-corrected procedure to generate an estimator of high accuracy and low rank. For this new estimator, we establish a non-asymptotic recovery error bound and analyze the impact of adding the rank-correction term on improving the recoverability. We also provide necessary and sufficient conditions for rank consistency in the sense of Bach [7], in which the concept of constraint nondegeneracy in matrix optimization plays an important role. These obtained results, together with numerical experiments, indicate the superiority of our proposed

rank-correction step over the nuclear norm penalization.

In the second part, we propose an adaptive semi-nuclear norm regularization approach to address rank regularized problems with hard constraints. This approach is designed via solving a nonconvex but continuous approximation problem iteratively. The quality of solutions to approximation problems is also evaluated. Our proposed adaptive semi-nuclear norm regularization approach overcomes the difficulty of extending the iterative reweighted l_1 minimization from the vector case to the matrix case. Numerical experiments show that the iterative scheme of our proposed approach has advantages of achieving both the low-rank-structure-preserving ability and the computational efficiency.

List of Figures

List of Tables

Notation

- Let $\mathbb{R}^n_+$ denote the cone of all nonnegative real n-vectors and let $\mathbb{R}^n_{++}$ denote the cone of all positive real n-vectors.

- Let $\mathbb{R}^{n_1\times n_2}$ and $\mathbb{C}^{n_1\times n_2}$ denote the space of all $n_1 \times n_2$ real and complex matrices, respectively. Let $\mathbb{M}^{n_1\times n_2}$ represent $\mathbb{R}^{n_1\times n_2}$ for the real case and $\mathbb{C}^{n_1\times n_2}$ for the complex case.

- Let $\mathcal{S}^n(\mathcal{S}^n_+, \mathcal{S}^n_{++})$ denote the set of all $n\times n$ real symmetric (positive semidefinite, positive definite) matrices and $\mathcal{H}^n(\mathcal{H}^n_+, \mathcal{H}^n_{++})$ denote the set of all $n \times n$ Hermitian (positive semidefinite, positive definite) matrices. Let $\mathbb{S}^n$ ($\mathbb{S}^n_+$, $\mathbb{S}^n_{++}$) represent $\mathcal{S}^n$ ($\mathcal{S}^n_+$, $\mathcal{S}^n_{++}$) for the real case and $\mathcal{H}^n$ ($\mathcal{H}^n_+$, $\mathcal{H}^n_{++}$) for the complex case.

- Let $\mathbb{V}^{n_1\times n_2}$ represent $\mathbb{R}^{n_1\times n_2}$, $\mathbb{C}^{n_1\times n_2}$, $\mathcal{S}^n$ or $\mathcal{H}^n$. We define $n := \min(n_1, n_2)$ for the previous two cases and stipulate $n_1 = n_2 = n$ for the latter two cases. Let $\mathbb{V}^{n_1\times n_2}$ be endowed with the trace inner product $\langle\cdot,\cdot\rangle$ and its induced norm $\|\cdot\|_F$, i.e., $\langle X, Y\rangle := \mathrm{Re}\big(\mathrm{Tr}(X^{\mathbb{T}}Y)\big)$ for $X, Y \in \mathbb{V}^{n_1\times n_2}$, where "Tr" stands for the trace of a matrix and "Re" means the real part of a complex

number.

- For the real case, $\mathbb{O}^{n\times k}$ denotes the set of all $n \times k$ real matrices with orthonormal columns, and for the complex case, $\mathbb{O}^{n\times k}$ denotes the set of all $n \times k$ complex matrices with orthonormal columns. When $k = n$, we write $\mathbb{O}^{n\times k}$ as $\mathbb{O}^n$ for short.

- Let $\mathbb{Q}_k$ be the set of all permutation matrices that have exactly one entry 1 in each row and column and 0 elsewhere. Let $\mathbb{Q}_k^{\pm}$ be the set of all signed permutation matrices that have exactly one entry 1 or -1 in each row and column and 0 elsewhere.

- The notation $^{\mathbb{T}}$ denotes the transpose for the real case and the conjugate transpose for the complex case. The notation * means the adjoint of operator.

- For any index set π, let $|\pi|$ denote the cardinality of π, i.e., the number of elements in π. For any $x \in \mathbb{R}^n$, let x_π denote the vector in $\mathbb{R}^{|\pi|}$ containing the components of x indexed by π, let $|x|$ denote the vector in $\mathbb{R}^n_+$ whose i-th component is $|x_i|$, and let x_+ denote the vector in $\mathbb{R}^n_+$ whose i-th component is $\max(0, x_i)$.

- For any given vector x, let $\mathrm{Diag}(x)$ denote the rectangular diagonal matrix of suitable size with the i-th diagonal entry being x_i.

- For any $x \in \mathbb{R}^n$, let $\|x\|_0$, $\|x\|_1$, $\|x\|_2$ and $\|x\|_\infty$ denote the l_0 norm (cardinality), the l_1 norm, the Euclidean norm and the maximum norm, respectively. For any $X \in \mathbb{V}^{n_1\times n_2}$, let $\|X\|$ and $\|X\|_*$ denote the spectral norm and the nuclear norm, respectively.

- Let I_n denote the $n \times n$ identity matrix. Let e denote the vector of suitable length whose entries are all ones. Let e_i denote the vector of suitable length whose i-th entries is one and the others are zeros.

- The notations $\overset{a.s.}{\to}$, $\overset{p}{\to}$ and $\overset{d}{\to}$ mean almost sure convergence, convergence in probability and convergence in distribution, respectively. We write $x_m = O_p(1)$ if x_m is bounded in probability.

- Let $\mathrm{sgn}(\cdot)$ denote the sign function defined over $\mathbb{R}$, i.e.,

$$\mathrm{sgn}(t) := \begin{cases} 1 & \text{if } t > 0, \\ 0 & \text{if } t = 0, \\ -1 & \text{if } t < 0. \end{cases}$$

  Let $\mathbb{1}(\cdot)$ denote the indicator function defined over $\mathbb{R}_+$, i.e.,

$$\mathbb{1}(t) := \begin{cases} 0 & \text{if } t = 0, \\ 1 & \text{if } t > 0. \end{cases}$$

  Let $\mathrm{id}(\cdot)$ denote the identity function defined over $\mathbb{R}_+$, i.e, $\mathrm{id}(t) := t$, $t \geq 0$.

- For any set K, let δ_K denote the characteristic function of K, i.e.,

$$\delta_K(x) := \begin{cases} 0 & \text{if } x \in K, \\ \infty & \text{if } x \notin K. \end{cases}$$

  This function is also called the indicator function of K. To avoid confusion with $\mathbb{1}(\cdot)$, we adopt the former name.

Chapter 1

Introduction

The problems with embedded low-rank structures arise in diverse areas such as engineering, statistics, quantum information, finance and graph theory. An important class of them is the low-rank matrix completion, which is of considerable interest recently in many applications, from machine learning to quantum state tomography. This problem refers to recovering an unknown low-rank or approximately low-rank matrix from a small number of noiseless or noisy observations of its entries, or more general, basis coefficients. In some cases, the unknown matrix to be recovered may possess a certain structure, for example, a correlation matrix from the financial market or a density matrix from the quantum state tomography. Besides, some reliable prior information on entries (or basis coefficients) may also be known, for example, the correlation coefficient between two pegged exchange rates can be fixed to be one in a correlation matrix of exchange rates. Existing algorithms such as OptSpace [82], SVP [123], ADMiRA [103], GROUSE [9] and LMaFit [178] have difficulty in dealing with such matrix completions problems with fixed entries (or basis coefficients), unless those additional requirements of the unknown matrix are ignored, which is of course an unwilling choice. An available choice, as far as we can see, is the nuclear norm technique. The nuclear

norm, i.e., the sum of all the singular values, is the convex envelope of the rank function over the unit ball of the spectral norm [49]. It has been shown that the nuclear norm technique is efficient for encouraging a low-rank solution in many cases including matrix completion in the literature. However, for structured matrix completion problems with fixed basis coefficients considered in this thesis, the efficiency of the nuclear norm could be highly weakened — may not able to lead to a desired low-rank solution with a small estimation error. How to address such matrix completion models constitutes our primary interest.

Another important problem is the rank regularized problem, which refers to minimizing the tradeoff between a loss function and the rank function over a convex set of matrices. The rank function can be used to measure the simplicity of a model in many applications, with its specific meaning may varying in different problems to be such as order, complexity or dimensionality. Many application problems in collaborative filtering [161, 162], system identification [50, 52], dimensionality reduction [108, 187], video inpainting [35] and graph theory [41, 1], to name but a few, can be cast into rank regularized problems, including also the matrix completion problem described above as a special case. Rank regularized problems are NP-hard in general due to the discontinuity and non-convexity of the rank function. The nuclear norm technique — replacing the rank function with the nuclear norm for a convex relaxation problem, is widely-used for finding a low-rank solution. For example, as a special case, the rank minimization problem — minimizing the rank of a matrix over a convex set, can be expressed as

$$\begin{array}{ll} \min & \operatorname{rank}(X) \\ \text{s.t.} & X \in K, \end{array} \tag{1.1}$$

where X is the decision variable and K is a convex subset of $\mathbb{M}^{n_1 \times n_2}$. Its convex relaxation using the nuclear norm is termed the nuclear norm minimization, taking

the form

$$\begin{aligned} \min \quad & \|X\|_* \\ \text{s.t.} \quad & X \in K. \end{aligned} \tag{1.2}$$

The equivalence of the rank minimization (1.1) and the nuclear norm minimization (1.2) has been establish under certain conditions. Nevertheless, there is still a big gap between them. Several iterative algorithms have been proposed in the literature to step forward to close the gap. However, when hard constraints are involved, how to efficiently address such low-rank optimization problems is still a challenge.

In view of above, in this thesis, we focus on dealing with the low-rank structure beyond the nuclear norm technique for matrix completion models with fixed basis coefficients and rank regularized problems with hard constraints. Partial results in this thesis come from the author's recent papers [127] and [128].

1.1 Literature review

The nuclear norm technique has been observed to provide a low-rank solution in practice for a long time, e.g., see [125, 124, 49]. The quality of the solution produced by using this technique is of particular interest in the literature. Among which, most works focus on the low-rank matrix recovery problem, which refers to recovering an unknown low-rank matrix from a number of its linear measurements. The nuclear norm minimization (1.2) is an important and efficient approach. The first remarkable theoretical characterization for the minimum rank solution via the nuclear norm minimization with linear equality constraints was given by Recht, Fazel and Parrilo [150]. It was shown that the success of recovering a low-rank matrix of rank at most r from its partial noiseless linear measurements via the nuclear norm minimization is guaranteed under a certain restricted isometric property

(RIP) condition for the linear map, which can be satisfied for several random ensembles with high probability as long as the number of linear measurements is larger than $O(r(n_1 + n_2)\log(n_1 n_2))$. The success of recovery indeed implies that with constraints defined by the linear map, the rank minimization (1.1) and the nuclear norm minimization (1.2) are equivalent in terms of their solutions. RIP is a powerful technique that can be used in the analysis of low-rank matrix recovery problems, not only for the nuclear norm minimization but also for other algorithms for recovery such as SVP [123] and ARMiDA [103]. However, RIP has its drawback — it is not invariant under measurement amplification. More precisely, given a linear map $\mathcal{A} : \mathbb{M}^{n_1 \times n_2} \to \mathbb{R}^m$, the linear system $\mathcal{A}(X) = b$ is equivalent to $\mathcal{B}(\mathcal{A}(X)) = \mathcal{B}(b)$ for any non-singular linear map $\mathcal{B} : \mathbb{R}^m \to \mathbb{R}^m$. But the RIP constant of the linear map $\mathcal{A}$ and $\mathcal{B} \circ \mathcal{A}$ could be dramatically different.

In the later work, different from the RIP-based analysis, necessary and sufficient null space conditions were provided by Recht, Xu and Hassibi [151] and Oymak and Hassibi [139] for exact low-rank matrix recovery, leading to an explicit relationship between the rank and the number of noiseless linear measurement for the success of recovery for Gaussian random ensembles. Meanwhile, Dvijotham and Fazel [37] presented another analysis of recovery based on the spherical section property (SSP) of the null space of the linear map. Very recently, Kong, Tuncel and Xiu [92] introduced the concepts of G-numbers of a linear map and derived necessary and sufficient conditions for recovery based on them. The obtained condition was shown to be equivalent to the null space condition in [139] and can be considered as a dual characterization.

All the results mentioned above focus on the noiseless case. In a more realistic setting, the available measurements are corrupted by a small amount of noise. Candès and Plan [20] derived recovery error bounds based on the RIP for two nuclear-norm-minimization based algorithms (the matrix Dantzig selector and the

matrix Lasso) and showed that linear measurements of order $O(r(n_1 + n_2))$ are enough for recovery provided they are sufficiently random. Other works for the RIP-type error bounds can be found in [102, 131, 93, 18]. Negahban and Wainwright [134] analyzed the nuclear norm penalized least squares estimator based on the restricted strong convexity (RSC) of a loss function introduced in [133] and established non-asymptotic error bounds on the Frobenius norm that are applicable to both exactly and approximately low-rank matrices. From a different perspective, Bach [7] derived necessary and sufficient conditions on the rank consistency of nuclear norm penalized least squares estimator and provided an adaptive version of it for free rank consistency. Almost all the results about the low-rank matrix recovery using the nuclear norm technique are somewhat extended from that about the sparse vector recovery via the l_1 norm. In view of this, Oymak et al. [140] provided a general approach for extending some sufficient conditions for recovery from vector cases to matrix cases.

The nuclear norm technique deserves its popularity not only because of its theoretical favor but also its computational efficiency. Fast algorithms for solving the nuclear norm minimization or its regularized versions, to name a few, include the singular value thresholding (SVT) algorithm [17], fixed point continuation with approximate SVD (FPCA) algorithm [114], the (inexact) accelerated proximal gradient (APG) algorithm [168, 79], the linearized alternating direction (LADM) method [185], the proximal point algorithm (PPA) [109] and the partial PPA [81, 80]. The efficiency of all these algorithms owes to the full use of the so-called singular value soft-thresholding operator, which is the proximal point mapping of the nuclear norm.

The low-rank matrix completion problem currently dominates the applications of the low-rank matrix recovery. For this problem, the linear measurements are specialized to be a small number of observations of entries, or more generally, basis

coefficients of the unknown matrix. In spite of being a special case of low-rank matrix recovery, unfortunately, the matrix completion problem does not have the Gaussian measurement ensemble and does not obey the RIP. Therefore, the theoretical results for low-rank matrix recovery mentioned above are not applicable. Instead of the RIP, Candès and Recht [21] introduced the concept of incoherence property and proved that most low-rank matrices can be exactly recovered from a surprisingly small number of noiseless observations of randomly sampled entries via the nuclear norm minimization. The bound of the number of sampled entries was later improved to be near-optimal of by Candés and Tao [22] through a counting argument. It was shown that, under suitable conditions, the number of entries required for recovery under the uniform sampling via the nuclear norm minimization is at most the degree of freedom by a poly-logarithmic factor in the dimension of matrix. Such a bound was also obtained by Keshavan et al. [82] for their proposed OptSpace algorithm, which is based on spectral methods followed by local manifold optimization. Later, Gross [70] sharpened the bound by employing a novel technique from quantum information theory developed in [71], in which noiseless observations were extended from entries to coefficients relative to any basis. This technique was also adapted by Recht [149], leading to a short and intelligible analysis. Besides the above results for the noiseless case, matrix completion with noise was first addressed by Candés and Plan [19]. More recently, nuclear norm penalized estimators for matrix completion with noise have been well studied by Koltchinskii, Lounici and Tsybakov [91], Negahban and Wainwright [135], Klopp [86] and Koltchinskii [89] under different settings. Several non-asymptotic order-optimal (up to logarithmic factors) error bounds in Frobenius norm have been correspondingly established. Besides the nuclear norm, other estimators for matrix completion with different penalties have also been considered in terms of recoverability in the literature, including the Schatten-p quasi-norm penalty by

Rohde and Tsybakov [156], the rank penalty by Klopp [85], the von Neumann entropy penalty by Koltchinskii [90], the max-norm (or $\gamma_{2:l_1 \to l_\infty}$ norm) by Srebro and colleagues [162, 163, 57] and the spectrum elastic net by Sun and Zhang [164].

However, the efficiency of the nuclear norm for finding a low-rank solution is not universal. The efficiency may be challenged in some circumstances. For example, the conditions characterized by Bach [7] for rank consistency of the nuclear norm penalized least squares estimator may not be satisfied, especially when certain constraints are involved. In particular for matrix completion problems, general sampling schemes may highly weaken the efficiency of the nuclear norm. Salakhutdinov and Srebro [158] showed that the nuclear norm minimization may fail for matrix completion when certain rows and/or columns are sampled with high probability. The failure is in the sense that the number of observations required for recovery are much more than the setting of most matrix completion problems, at least the degree of freedom by a polynomial factor in the dimension rather than a poly-logarithmic factor. Negahban and Wainwright [135] also pointed out the impact of such heavy sampling schemes on the recovery error bound. As a remedy for this, a weighted nuclear norm (trace norm), based on row- and column-marginals of the sampling distribution, was suggested in [135, 158, 56] for achieving better performance if the prior information on sampling distribution is available.

In order to go beyond the limitation of the nuclear norm, several iterative algorithms have also been proposed for solving rank regularized problems (rank minimization problems) in the literature. Fazel, Hindi and Boyd in [51] (see also [49]) proposed the reweighted trace minimization for minimizing the rank of a positive semidefinite matrix, which falls into the class of majorization methods. The log-det function, which is concave over the positive semidefinite cone, is typically used to be the surrogate of the rank function, leading to a linear majorization in each iteration. Later, an attempt to extend this approach to the the reweighted

nuclear norm minimization for the rectangular case was conducted Mohan and Fazel [132]. Iterative reweighted least squares algorithms were also independently proposed by Mohan and Fazel [130] and Fornasier, Rauhut and Ward [54], which enjoy improved performance beyond the nuclear norm and may allow for efficient implementations. Besides, Lu and Zhang [113] proposed penalty decomposition methods for both rank regularized problems and rank constrained problems which make use of the closed-form solutions of some special minimization involving the rank function.

1.2 Contributions

In the first part of this thesis, we address low-rank matrix completion models with fixed basis coefficients. In our setting, given a basis of the matrix space, a few basis coefficients of the unknown matrix are assumed to be fixed due to a certain structure or some prior information, and the rest are allowed to be observed with noises under general sampling schemes. Certainly, one can apply the nuclear norm penalized technique to recover the unknown matrix. The challenge is that, this may not yield a desired low-rank solution with a small estimation error.

Our consideration is strongly motivated by correlation and density matrix completion problems. When the true matrix possesses a symmetric/Hermitian positive semidefinite structure, the impact of general sampling schemes on the recoverability of the nuclear norm technique is more remarkable. In this situation, the nuclear norm reduces to the trace and thus only depends on diagonal entries rather than all entries as the rank function does. As a result, if diagonal entries are heavily sampled, the rank-promoting ability of the nuclear norm, as well as the recoverability, will be highly weakened. This phenomenon is fully reflected in the widely-used correlation matrix completion problem, for which the nuclear norm

becomes a constant and severely loses its effectiveness for matrix recovery. Another example of particular interest in quantum state tomography is to recover a density matrix of a quantum system from Pauli measurements (e.g., see [71, 53, 175]). A density matrix is a Hermitian positive semidefinite matrix of trace one. Obviously, if the constraints of positive semidefiniteness and trace one are simultaneously imposed on the nuclear norm minimization, the nuclear norm completely fails in promoting a low-rank solution. Thus, one of the two constraints has to be abandoned in the nuclear norm minimization and then be restored in the post-processing stage. In fact, this idea has been much explored in [71, 53] and the numerical results there indicated its relative efficiency though it is at best sub-optimal.

In order to optimally address the difficulties in low-rank matrix completion with fixed basis coefficient, especially in correlation and density matrix completion problems, we propose a low-rank matrix completion model with fixed basis coefficients. A rank-correction step is introduced to address this critical issue provided that a reasonable initial estimator is available. A satisfactory choice of the initial estimator is the nuclear norm penalized estimator or one of its analogies. The rank-correction step solves a convex "nuclear norm − rank-correction term + proximal term" regularized least squares problem with fixed basis coefficients (and the possible positive semidefinite constraint). The rank-correction term is a linear term constructed from the initial estimator, and the proximal term is a quadratic term added to ensure the boundedness of the solution to the convex problem. The resulting convex matrix optimization problem can be solved by the efficient algorithms recently developed in [79, 81, 80] even for large-scale cases.

The idea of using a two-stage or even multi-stage procedure is not brand new for dealing with sparse recovery in the statistical and machine learning literature. The l_1 norm penalized least squares method, also known as the Lasso [167], is very

attractive and popular for variable selection in statistics, thanks to the invention of the fast and efficient LARS algorithm [39]. On the other hand, the l_1 norm penalty has long been known by statisticians to yield biased estimators and cannot attain the estimation optimality [43, 47]. The issue of bias could be mitigated or overcome by nonconvex penalization methods. Commonly-used nonconvex penalties include the l_q norm penalty $(0 < q < 1)$ by Frank and Friedman [104], the smoothly clipped absolute deviation (SCAD) penalty by Fan [42], and the minimax concave penalty (MCP) by Zhang [190]. A multi-stage procedure naturally occurs if the nonconvex problem obtained is solved by an iterative algorithm [195]. In particular, once a good initial estimator is used, a two-stage estimator is enough to achieve the desired asymptotic efficiency, e.g., the adaptive Lasso proposed by Zou [194] and the relaxed Lasso proposed by Meinshausen [121]. There are also a number of important papers in this line on variable selection, including [104, 122, 191, 77, 193, 44], to name only a few. For a broad overview, the interested readers are referred to the recent survey papers [45, 46]. It is natural to extend the ideas from the vector case to the matrix case. Recently, Bach [7] made an important step in extending the adaptive Lasso of Zou [194] to the matrix case for seeking rank consistency under general sampling schemes. However, it is not clear how to apply Bach's idea to our matrix completion model with fixed basis coefficients since the required rate of convergence of the initial estimator for achieving asymptotic properties is no longer valid as far as we can see. More critically, there are numerical difficulties in efficiently solving the resulting optimization problems. Such difficulties also occur when the reweighted nuclear norm proposed by Mohan and Fazel [132] is applied to the rectangular matrix completion problems.

The rank-correction step to be proposed in this thesis is for the purpose to overcome the above difficulties. This approach is inspired by the majorized penalty

method recently proposed by Gao and Sun [62] for solving structured matrix optimization problems with a low-rank constraint. For our proposed rank-correction step, we provide a non-asymptotic recovery error bound in the Frobenius norm, following a similar argument adopted by Klopp in [86]. The obtained error bound indicates that adding the rank-correction term could help to substantially improve the recoverability. As the estimator is expected to be of low-rank, we also study the asymptotic property — rank consistency in the sense of Bach [7], under the setting that the matrix size is assumed to be fixed. This setting may not be ideal for analyzing asymptotic properties for matrix completion, but it does allow us to take the crucial first step to gain insights into the limitation of the nuclear norm penalization. In particular, the concept of constraint nondegeneracy for conic optimization problem plays a key role in our analysis. Interestingly, our results of recovery error bound and rank consistency consistently suggest a criterion for constructing a suitable rank-correction function. In particular, for the correlation and density matrix completion problems, we prove that the rank consistency automatically holds for a broad selection of rank-correction functions. For most cases, a single rank-correction step is enough for significantly reducing the recovery error. But if the initial estimator is not good enough, e.g., the nuclear norm penalized least squares estimator when the sample ratio is very low, the rank-correction step may also be iteratively used for several times for achieving better performance. Finally, we remark that our results can also be used to provide a theoretical foundation for the majorized penalty method of Gao and Sun [62] and Gao [61] for structured low-rank matrix optimization problems.

In the second part of this thesis, we address rank regularized problems with hard constraints. Although the nuclear norm technique is still a choice for such problems, its rank-promoting ability could be much more limited, since the problems of consideration is more general than low-rank matrix recovery problems and

could hardly have any property for guaranteeing the efficiency of its convex relaxation. To go a further step beyond the nuclear norm, inspired by the efficiency of the rank-correction step for matrix completions problems (with fixed basis coefficients), we propose an adaptive semi-nuclear norm regularization approach for rank regularized problems (with hard constraints). This approach aims to solve an approximation problem instead, whose regularization term is a nonconvex but continuous surrogate of the rank function that can be written as the nuclear norm of the Löwner's (singular value) operator associated with a concave increasing function over $\mathbb{R}_+$. The relationship between a rank regularized problem and its approximations is examined by using the epi-convergence. In particular for affine rank minimization problems, we establish a necessary and sufficient null space condition for ensuring the minimum-rank solution to the approximation problem. This result further indicates that the considered nonconvex candidate surrogate function possesses better rank-promoting ability (recoverability) than the nuclear norm. Compared with the nuclear norm regularization, the convexity for computational convenience is sacrificed in change of the improvement of rank-promoting ability.

Being an application of the majorized proximal gradient method proposed for general nonconvex optimization problems, the adaptive semi-nuclear norm regularization approach solves a sequence of convex optimization problems regularized by a semi-nuclear norm in each iteration. Under mild conditions, we show that any limit point of the sequence generated by this approach is a stationary point of the corresponding approximation problem. Thanks to the semi-nuclear norm, each subproblem can be efficiently solved by recently developed methodologies with high accuracy, allowing for the use of the singular value soft-thresholding operator. Still thanks to the semi-nuclear norm, each iteration of this approach produces a low-rank solution. This property is crucial since when hard constraints are involved,

each subproblem could be computational costly so that the fewer iterations the better. Our proposed adaptive semi-nuclear norm regularization approach overcomes the difficulty of extending the iterative reweighted l_1 minimization from the vector case to the matrix case. In particular for rank minimization problems, we specified our approach to be the adaptive semi-nuclear norm minimization. For the positive semidefinite case, this specified algorithm recovers the reweighted trace minimization proposed by Fazel, Hindi and Boyd in [51], expect for an additional proximal term. Therefore, the idea of using adaptive semi-nuclear norms can be essentially regarded as an extension of the reweighted trace minimization from the positive semidefinite case to the rectangular case. In spite of this, even for the positive semidefinite rank minimization, our approach is still distinguished for its computational efficiency due to the existence of the proximal term. Compared with other existing iterative algorithms for rank regularized problems (rank minimization problems), the iterative scheme of our proposed approach has advantages of both the low-rank-structure-preserving ability and the computational efficiency, both of which are especially crucial and favorable for rank regularized problems with hard constraints.

1.3 Outline of the thesis

This thesis is organized as follows: In Chapter 2, we provide some preliminaries that will be used in the subsequent discussions. Besides introducing some concepts and properties of the majorization, the spectral operator and the epi-convergence (in distribution), we derive the Clarke generalized gradients of the w-weighted norm, provide f-version inequalities of singular values and propose the majorized proximal gradient method for solving general nonconvex optimization problems. In Chapter 3, we introduce the observation model of matrix completion with fixed

basis coefficients and the formulation of the rank-correction step. We establish a non-asymptotic recovery error bound and discuss the impact of the rank-correction term on recovery. We also provide necessary and sufficient conditions for rank consistency. The construction of the rank-correction function is discussed based on the obtained results. Numerical results are reported to validate the efficiency of our proposed rank-corrected procedure. In Chapter 4, we propose an adaptive semi-nuclear norm regularization approach for rank regularized problems with hard constraints. We discuss the approximation quality of the problem solved in this approach and establish the convergence of this approach. Several families of candidate surrogate functions available for this approach are provided with a further discussion. We also compare this approach with some existing iterative algorithms for rank regularized problem (rank minimization problem). Numerical experiments are provided to support the superiority of our approach. We conclude this thesis and discuss the further work in Chapter 5.

Chapter 2

Preliminaries

In this chapter, we introduce some basic properties which are essential to our discussions in the subsequent chapters.

2.1 Majorization

This concept of majorization — a partial ordering on vectors, was introduced by Hardy, Littlewood and Pólya [74].

Definition 2.1. *For any $x, y \in \mathbb{R}^n$, we say that x is weakly majorized by y, denoted by $x \prec_w y$, if*

$$\sum_{i=1}^{k} x_{[i]} \leq \sum_{i=1}^{k} y_{[i]}, \quad k = 1, \ldots, n, \tag{2.1}$$

where $(x_{[1]}, \ldots, x_{[n]})^T$ is the vector by rearranging the components x_i, $i = 1, \ldots, n$, in the nonincreasing order, i.e., $x_{[1]} \geq \cdots \geq x_{[n]}$. Moreover, we say that x is majorized by y, denoted by $x \prec y$, if (2.1) holds with equality for $k = n$.

Given any $x \in \mathbb{R}^n$, let $\Delta(x)$ denote the convex hull of the set of vectors obtained from x by all possible permutation of its components, i.e.,

$$\Delta(x) := \operatorname{conv}\big(\{z \in \mathbb{R}^n \mid z = Qx, Q \in \mathbb{Q}_n\}\big);$$

and let $\Gamma(x)$ denote the convex hull of the set of vectors obtained from x by all possible signed permutation of its components, i.e.,

$$\Gamma(x) := \operatorname{conv}\big(\{z \in \mathbb{R}^n \mid z = Qx, Q \in \mathbb{Q}_n^{\pm}\}\big);$$

Rado [147] first characterized the convex hull of permutations of a given vector as follows:

Theorem 2.1 (Rado [147]). *Let $x, y \in \mathbb{R}^n$. Then $x \prec y$ if and only if $x \in \Delta(y)$.*

As observed by Horn [76], this result can also be obtained by combining earlier results of Hardy, Littlewood and Pólya [74] and Birkhoff [14]. Later, Markus [117, Theorem 1.2] characterized the convex hull of signed permutations of a given vector.

Theorem 2.2 (Markus [117]). *Let $x, y \in \mathbb{R}^n$. Then $x \prec_w y$ if and only if $x \in \Gamma(y)$.*

The following result taken from [118, Proposition 3.C.1] will also be useful in the sequel.

Proposition 2.3. *Let $I \subseteq \mathbb{R}$ be an interval and $g : I \to \mathbb{R}$ be a convex function. Define $\phi(x) := \sum_{i=1}^n g(x_i)$. Then $x \prec y$ on I^n implies $\phi(x) \leq \phi(y)$.*

2.2 The spectral operator

For any real or complex matrix $X \in \mathbb{M}^{n_1 \times n_2}$, the singular value decomposition (SVD) of X is a factorization of the form

$$X = U\operatorname{Diag}\big(\sigma(X)\big)V^{\mathbb{T}},$$

where $\sigma(X) = \big(\sigma_1(X), \ldots, \sigma_n(X)\big)^{\mathbb{T}}$ denotes the vector of singular values of X arranged in the nonincreasing order, and $U \in \mathbb{O}^{n_1}, V \in \mathbb{O}^{n_2}$ are orthogonal matrices corresponding to the left and right singular vectors respectively. We define

$$\mathbb{O}^{n_1,n_2}(X) := \big\{(U, V) \in \mathbb{O}^{n_1} \times \mathbb{O}^{n_2} \mid X = U\operatorname{Diag}\big(\sigma(X)\big)V^{\mathbb{T}}\big\}.$$

In particular for any real symmetric matrix or complex Hermitian matrix $X \in \mathbb{S}^n$, an eigenvalue decomposition of X takes the form

$$X = P\mathrm{Diag}\big(\lambda(X)\big)P^{\mathbb{T}},$$

where $\lambda(X) = \big(\lambda_1(X), \ldots, \lambda_n(X)\big)^{\mathbb{T}}$ denotes the vector of eigenvalues of X arranged in the nonincreasing order and $P \in \mathbb{O}^n$ is an orthogonal matrix corresponding to eigenvectors. We define

$$\mathbb{O}^n(X) := \big\{P \in \mathbb{O}^n \mid X = P\mathrm{Diag}\big(\lambda(X)\big)P^{\mathbb{T}}\big\}.$$

Now, we introduce the concept of spectral operator associated with a symmetric vector-valued function.

Definition 2.2. *A function $f : \mathbb{R}^n \to \mathbb{R}^n$ is said to be symmetric if*

$$f(x) = Q^{\mathbb{T}} f(Qx) \quad \forall\, Q \in \mathbb{Q}_n^{\pm} \text{ and } \forall\, x \in \mathbb{R}^n,$$

Definition 2.3. *The spectral operator $F : \mathbb{M}^{n_1\times n_2} \to \mathbb{M}^{n_1\times n_2}$ associated with the function f is defined by*

$$F(X) := U\mathrm{Diag}\big(f(\sigma(X))\big)V^{\mathbb{T}}, \tag{2.2}$$

where $(U, V) \in \mathbb{O}^{n_1,n_2}(X)$ and $X \in \mathbb{M}^{n_1\times n_2}$.

Notice that the symmetry of f implies that

$$\big(f(x)\big)_i = 0 \quad \text{if } x_i = 0.$$

This guarantees the well-definedness of the spectral operator F ([33, Theorem 3.1]). Moreover, the continuous differentiability of f implies the continuous differentiability of F. When $X \in \mathbb{S}^n$, we have an equivalent representation of (2.2) as

$$F(X) = P\mathrm{Diag}\big(f(|\lambda(X)|)\big)\big(P\mathrm{Diag}(s(X))\big)^{\mathbb{T}},$$

where $P \in \mathbb{O}^n(X)$, and $s(X) \in \mathbb{R}^n$ with its i-th component taking the value

$$s_i(X) = \begin{cases} -1, & \text{if } \lambda_i(X) < 0 \\ 1, & \text{if otherwise.} \end{cases}$$

In particular for the positive semidefinite case, both U and V in (2.2) reduce to P. For more details on spectral operators, the reader may refer to the PhD thesis of Ding [33].

Based on the well-definedness of the spectral operator, the well-known Löwner's (eigenvalue) operators [112] can be extended from symmetric matrices to nonsymmetric matrices defined as follows.

Definition 2.4. *Let $f : \mathbb{R}_+ \to \mathbb{R}$ be a function such that $f(0) = 0$. The Löwner's (singular value) operator $F : \mathbb{M}^{n_1 \times n_2} \to \mathbb{M}^{n_1 \times n_2}$ associated with f is defined by*

$$F(X) := U\mathrm{Diag}\big(f(\sigma_1(X)), \ldots, f(\sigma_n(X))\big)V^{\mathbb{T}}, \tag{2.3}$$

where $(U, V) \in \mathbb{O}^{n_1,n_2}(X)$ and $X \in \mathbb{M}^{n_1 \times n_2}$.

It is not hard to check that the Löwner's operator $F(X)$ defined by (2.3) can be regarded as a special spectral operator associated with the symmetric function $g : \mathbb{R}^n \to \mathbb{R}^n$ defined by

$$g(x) = \big(\mathrm{sgn}(x_1)f(|x_1|), \cdots, \mathrm{sgn}(x_n)f(|x_n|)\big)^{\mathbb{T}}, \quad x \in \mathbb{R}^n.$$

This operator has also been discussed in the Master thesis of Yang [186].

For preparation of discussions in the sequel, we introduce the so-called singular values soft- and hard-thresholding operators, which in fact fall in the class of special spectral operators (or Löwner's operators). For any matrix $Z \in \mathbb{M}^{n_1 \times n_2}$ and any real number $\tau > 0$, the singular values soft-thresholding operator $\mathcal{P}_\tau^{\mathrm{soft}} : \mathbb{M}^{n_1 \times n_2} \to \mathbb{M}^{n_1 \times n_2}$ is defined by

$$\mathcal{P}_\tau^{\mathrm{soft}}(Z) := U\mathrm{Diag}\big((\sigma(Z) - \tau)_+\big)V^{\mathbb{T}}, \quad (U, V) \in \mathbb{O}^{n_1,n_2}(Z), \tag{2.4}$$

and the singular value hard-thresholding operator is defined by

$$\mathcal{P}_\tau^{\mathrm{hard}}(Z) := U\mathrm{Diag}\big(\widetilde{\sigma}(Z)\big)V^{\mathbb{T}}, \quad (U,V) \in \mathbb{O}^{n_1,n_2}(Z), \tag{2.5}$$

where

$$\widetilde{\sigma}_i(Z) = \begin{cases} \sigma_i(Z) & \text{if } \sigma_i(Z) > \tau, \\ 0 & \text{if } \sigma_i(Z) \le \tau, \end{cases} \qquad i = 1, \ldots, n.$$

The name "soft" and "hard" come from the two different ways in dealing with the singular values of a matrix. It is well-known that in fact, the singular value soft-thresholding operator is the proximal point mapping of the nuclear norm, i.e.,

$$\mathcal{P}_\tau^{\mathrm{soft}}(Z) = \underset{X \in \mathbb{M}^{n_1 \times n_2}}{\arg\min} \Big\{ \|X\|_* + \frac{1}{2\tau}\|X - Z\|^2 \Big\},$$

e.g., see [17, 115]; while the singular value hard-thresholding operator is the (non-convex) proximal point mapping of the rank function, i.e.,

$$\mathcal{P}_\tau^{\mathrm{hard}}(Z) \in \underset{X \in \mathbb{M}^{n_1 \times n_2}}{\arg\min} \Big\{ \mathrm{rank}(X) + \frac{1}{\tau^2}\|X - Z\|^2 \Big\}.$$

2.3 Clarke's generalized gradients

Let $\phi : \mathbb{R}^n \to [-\infty, \infty]$ be any absolutely symmetric function, i.e., $\phi(Qx) = \phi(x)$ for any $x \in \mathbb{R}^n$ and any signed permutation matrix $Q \in \mathbb{Q}_n^{\pm}$. Define a singular value function $\Phi : \mathbb{M}^{n_1 \times n_2} \to [-\infty, \infty]$ as

$$\Phi(X) := \phi\big(\sigma(X)\big).$$

For any $X \in \mathbb{M}^{n_1 \times n_2}$, it is easy to see that ϕ is Lipschitz near $\sigma(X)$ if and only if Φ is Lipschitz near X. In this case, from [105, Theorem 3.7], the Clarke generalized gradient of Φ at X, denoted by $\partial\Phi(X)$, can be characterized as

$$\partial\Phi(X) = \big\{ U\mathrm{Diag}(d)V^{\mathbb{T}} \mid d \in \partial\phi\big(\sigma(X)\big), (U,V) \in \mathbb{O}^{n_1,n_2}(X) \big\}. \tag{2.6}$$

It is not hard to see from (2.6) that Φ is differential at X if and only if ϕ is differential at $\sigma(X)$. Based on this remarkable result, we next characterize Clarke's generalized gradients for two classes of singular valued functions.

Theorem 2.4. *Given an extended real-valued function f on $\mathbb{R}$, define $\phi : \mathbb{R}^n \to [-\infty, \infty]$ and $\Phi : \mathbb{M}^{n_1\times n_2} \to [-\infty, \infty]$ as*

$$\phi(x) := \sum_{i=1}^{n} f(|x_i|) \quad \text{and} \quad \Phi(X) := \phi\big(\sigma(X)\big) = \sum_{i=1}^{n} f\big(\sigma_i(X)\big).$$

For any $X \in \mathbb{M}^{n_1\times n_2}$, if f is Lipschitz continuous near all $\sigma_1(X), \ldots, \sigma_n(X)$, then the Clarke's generalized gradient $\partial\Phi(X)$ can be characterized as (2.6) with

$$\partial\phi\big(\sigma(X)\big) = \left\{ d \in \mathbb{R}^n \;\middle|\; \begin{array}{l} d_i \in \partial f\big((\sigma_i(X)\big) \text{ if } \sigma_i(X) > 0 \\ d_i \in \big[-|f'_+(0)|, |f'_+(0)|\big] \text{ if } \sigma_i(X) = 0 \end{array} \right\}, \tag{2.7}$$

where f'_- and f'_+ denote the left derivative and the right derivatives of f respectively.

Proof. It is known from [27, Theorem 2.5.1] that $\partial\phi\big(\sigma(X)\big)$ is the convex hull of $\partial_B\phi\big(\sigma(X)\big)$ taking the form

$$\partial_B\phi\big(\sigma(X)\big) := \Big\{ \lim_{k\to\infty} \nabla\phi(x^k) \;\big|\; x^k \to \sigma(X), \phi \text{ is differentiable at } x^k \Big\},$$

where $\nabla\phi(x^k)$ denotes the gradient of ϕ at x^k. By direct calculation, we have

$$\partial_B\phi\big(\sigma(X)\big) = \left\{ d \in \mathbb{R}^n \;\middle|\; \begin{array}{l} d_i \in \big\{f'_-\big(\sigma_i(X)\big), f'_+\big(\sigma_i(X)\big)\big\} \text{ if } \sigma_i(X) > 0 \\ d_i \in \big\{-f'_+(0), f'_+(0)\big\} \text{ if } \sigma_i(X) = 0 \end{array} \right\}.$$

Then, by taking the convex hull, we easily obtain (2.7). Thus, we complete the proof. □

Theorem 2.5. *Given any vector $w \in \mathbb{R}^n_+$, define $\phi : \mathbb{R}^n \to [-\infty, \infty]$ and $\Phi : \mathbb{M}^{n_1\times n_2} \to [-\infty, \infty]$ as*

$$\phi(x) := \sum_{i=1}^{n} w_i|x|_{[i]}, \quad x \in \mathbb{R}^n \quad \text{and} \quad \Phi(X) := \phi\big(\sigma(X)\big) = \sum_{i=1}^{n} w_i\sigma_i(X).$$

For any matrix $X \in \mathbb{M}^{n_1\times n_2}$, define the index sets $\pi_1, \ldots, \pi_l$ as

$$\pi_k := \{i : \sigma_i(X) = s_k\}, \quad k = 1, \ldots, l,$$

where $s_1 > \ldots > s_l$ are all the l distinct singular values of X. Then, Clarke's generalized gradient $\partial\Phi(X)$ can be characterized as (2.6) with

$$\partial\phi\big(\sigma(X)\big) = \left\{ d \in \mathbb{R}^n \;\middle|\; \begin{array}{l} d_{\pi_k} \prec w_{\pi_k} \;\forall\, k = 1, \ldots, l-1, \\ d_{\pi_l} \prec w_{\pi_l} \text{ if } s_l > 0 \text{ and } |d_{\pi_l}| \prec_w w_{\pi_l} \text{ if } s_l = 0 \end{array} \right\}. \tag{2.8}$$

In particular, for any $(U, V) \in \mathbb{O}^{n_1,n_2}(X)$, one has

$$U\mathrm{Diag}(w)V^{\mathbb{T}} \in \partial\Phi(X). \tag{2.9}$$

Proof. From (2.6), it suffices to characterize $\partial\phi\big(\sigma(X)\big)$. It is not hard to check that $\partial_B\phi\big(\sigma(X)\big)$ can be explicitly expressed as

$$\partial_B\phi\big(\sigma(X)\big) = \left\{ d \in \mathbb{R}^n \;\middle|\; \begin{array}{l} d_{\pi_k} = Q_k w_{\pi_k}, Q_k \in \mathbb{Q}_{|\pi_k|} \text{ for } k = 1, \ldots, l-1, \\ d_{\pi_l} = Q_l w_{\pi_l}, Q_l \in \mathbb{Q}_{|\pi_l|} \text{ if } s_l > 0 \text{ and } Q_l \in \mathbb{Q}^{\pm}_{|\pi_l|} \text{ if } s_l = 0 \end{array} \right\}.$$

Then, by taking the convex hull, from Theorems 2.1 and 2.2, we can write $\partial\phi\big(\sigma(X)\big)$ in terms of (2.8). Then, by further noting the fact that $w \in \partial_B\phi\big(\sigma(X)\big) \subseteq \partial\phi\big(\sigma(X)\big)$, we also have (2.9). Thus, we complete the proof. □

In particular, for any $\mathbb{R}^n_+ \ni w \neq 0$ satisfying $w_1 \geq \ldots \geq w_n \geq 0$, Φ_w defines an orthogonally invariant matrix norm on $\mathbb{M}^{n_1\times n_2}$, called w-weighted norm, denoted by $\|\cdot\|_w$. In this case, by noting the convexity of $\|\cdot\|_w$, Clarke's generalized gradient coincides with the subdifferential of $\|\cdot\|_w$ at X, i.e.,

$$\partial\|X\|_w := \big\{G \in \mathbb{M}^{n_1\times n_2} \mid \|Y\|_w \geq \|X\|_w + \langle G, Y - X\rangle \;\forall\, Y \in \mathbb{M}^{n_1\times n_2}\big\}. \tag{2.10}$$

It is easy to see that $\|\cdot\|_w$ will reduce to the spectral norm, the nuclear norm and the Ky Fan k norm by choosing $w = e_1$, $w = \sum_{i=1}^n e_i$ and $w = \sum_{i=1}^k e_i$ respectively, where $e_i \in \mathbb{R}^n$ denotes the vector whose i-th entry is 1 with all others entries 0. The readers may also refer to [176, 177] for the subdifferentials of these special matrix norms.

2.4 f-version inequalities of singular values

For any $X, Y \in \mathbb{M}^{n_1\times n_2}$, it is well-known that singular values of sum of matrices have the property

$$\sigma(X+Y) \prec_w \sigma(X)+\sigma(Y). \tag{2.11}$$

This weak majorization of singular values, also known as the triangle inequalities for Ky Fan k-norms, was first proved by Ky Fan [48] in 1949. This early result was later extended by Rotfel'd [157, Theorem 1], Thompson [165, Theorem 3] and Uchiyama [171, Theorem 4.4] to f-version subadditive inequalities, finally stated as follows.

Theorem 2.6. *Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be a concave function with $f(0) = 0$. Then, for any $X, Y \in \mathbb{M}^{n_1\times n_2}$, one has*

$$\sum_{i=1}^{k} f\big(\sigma_i(X+Y)\big) \le \sum_{i=1}^{k} f\big(\sigma_i(X)\big) + \sum_{i=1}^{k} f\big(\sigma_i(Y)\big), \quad k = 1, \dots, n.$$

A stronger version of the weak majorization (2.11) also holds, known as the perturbation theorem of singular values (see [13, Theorem IV.3.4]), i.e., for any $X, Y \in \mathbb{M}^{n_1\times n_2}$,

$$|\sigma(X) - \sigma(Y)| \prec_w \sigma(X-Y). \tag{2.12}$$

Notice that it is immediate from Theorem 2.6 that

$$\left|\sum_{i=1}^{k}\Big(f\big(\sigma_i(X)\big) - f\big(\sigma_i(Y)\big)\Big)\right| \le \sum_{i=1}^{k} f\big(\sigma_i(X-Y)\big), \quad k = 1, \dots, n. \tag{2.13}$$

This inequality provides us the possibility to extend the majorization (2.12) to an f-version, by replacing the left-hand side of (2.13) with the sum of absolute values. Recently, Miao (see [146, Conjecture 6]) proposed the following conjecture:

Conjecture 2.7. *Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be a concave function with $f(0) = 0$. Then, for any $X, Y \in \mathbb{M}^{n_1\times n_2}$, one has*

$$\sum_{i=1}^{k} \big|f\big(\sigma_i(X)\big) - f\big(\sigma_i(Y)\big)\big| \le \sum_{i=1}^{k} f\big(\sigma_i(X-Y)\big), \quad k = 1, \dots, n. \tag{2.14}$$

A particular case of the f-version inequality (2.14) when $f(t) = t^q, 0 < q \leq 1$ and $k = n$ was also proposed by Oymak et al. in [140]. Very recently, Lai et al. [94] proved that when $X, Y \in \mathbb{S}_+^n$, this conjecture holds true for the case that $f(t) = t^q, 0 < q \leq 1$. At almost the same time, Yue and So [188, Theorem 5] proved this conjecture for the case that $k = n$ with a stronger requirement that f is continuously differentiable. Here, we slightly extend Yue and So's result to get rid of the continuous differentiability of f.

Theorem 2.8. *Conjecture 2.7 holds for $k = n$.*

Proof. We only need to prove for the continuous case $f(0+) = 0$. For the discontinuous case $f(0+) > 0$, one can simply consider the continuous function $f - f(0+)$.

If $f'_+(0) < \infty$, we define $\widehat{f} : \mathbb{R} \to \mathbb{R}$ by

$$\widehat{f}(t) := \begin{cases} f(t) & \text{if } t \geq 0, \\ f'_+(0)t & \text{if } t < 0. \end{cases}$$

Then consider a sequence of the so-called averaged functions $\widehat{f}^k$ defined by

$$\widehat{f}^k(t) := \int_{-\infty}^{\infty} \widehat{f}(t-s)\psi^k(s)ds,$$

where $\{\psi^k\}$ is a sequence of bounded, measurable functions with $\int_0^\infty \psi^k(s)ds = 1$ such that $B^k = \{s \mid \psi^k(s) > 0\}$ form a bounded sequence converging to $\{0\}$. Notice that $\widehat{f}$ is increasing and concave over $\mathbb{R}$ and thus locally Lipschitz continuous. It then follows from [155, Theorem 9.69] that the averaged functions $\widehat{f}^k$ are continuously differentiable and converge uniformly to $\widehat{f}$ on any compact sets. In particular, $\widehat{f}^k(0) \to \widehat{f}(0) = 0$ as $k \to \infty$. For each k, define

$$\widetilde{f}^k(t) := \widehat{f}^k(t) - \widehat{f}^k(0).$$

By further checking that each $\widetilde{f}^k$ is increasing and concave over $\mathbb{R}_+$ with $\widetilde{f}^k(0) = 0$, from [188, Theorem 5], we obtain that

$$\sum_{i=1}^n \left|\widetilde{f}^k\big(\sigma_i(X)\big) - \widetilde{f}^k\big(\sigma_i(Y)\big)\right| \leq \sum_{i=1}^n \widetilde{f}^k\big(\sigma_i(X-Y)\big) \quad \forall\, k.$$

By letting $k \to \infty$, we obtain that the inequality (2.14) holds for $k = n$.

If $f'_+(0) = \infty$, we further consider $f_\varepsilon : \mathbb{R}_+ \to \mathbb{R}_+$ defined by

$$f_\varepsilon(t) := f(t+\varepsilon) - f(\varepsilon), \quad t \geq 0$$

with $\varepsilon > 0$. As reduced to the previous case, we obtain that

$$\sum_{i=1}^{n} \left| f_\varepsilon\big(\sigma_i(X)\big) - f\big(\sigma_i(Y)\big) \right| \leq \sum_{i=1}^{n} f_\varepsilon\big(\sigma_i(X-Y)\big) \quad \forall\, \varepsilon > 0.$$

By letting $\varepsilon \downarrow 0$, we obtain that the inequality (2.14) holds for $k = n$. Thus, we complete the proof. □

Another extension of the weak majorization (2.11) is the generalized Lidskii inequality proved by Thompson and Freede [166, Theorem 3], stated as follows (see also [118, Theorem 9.C.4]):

Theorem 2.9 (Thompson and Freede [166]). *Let $i_1, \ldots, i_k$ and $j_i, \ldots, j_k$ be integers such that*

$$1 \leq i_1 < \ldots < i_k \leq n, \quad 1 \leq j_1 < \ldots < j_k \leq n \quad \text{and} \quad i_k + j_k \leq k + n. \tag{2.15}$$

Then for any $X, Y \in \mathbb{M}^{n_1 \times n_2}$, we have

$$\sum_{s=1}^{k} \sigma_{i_s + j_s - s}(X+Y) \leq \sum_{s=1}^{k} \sigma_{i_s}(X) + \sum_{s=1}^{k} \sigma_{j_s}(Y).$$

Theorem 2.9 reveals a more specific perspective of the nature of singular values of X,Y and $X+Y$. In particular, this theorem includes the well-known Weyl's inequality [179] and the Lidskii's inequality [106] as special cases. It arouses our curiosity that whether the following conjecture, as a stronger version of Theorem 2.15, holds as well.

Conjecture 2.10. *Let $i_1, \ldots, i_k$ and $j_i, \ldots, j_k$ be integers such that (2.15) holds. For any $X, Y \in \mathbb{M}^{n_1 \times n_2}$ and for each $1 \leq i \leq n$, define*

$$\alpha_i := \max\{\sigma_i(X), \sigma_i(Y)\}, \quad \beta_i := \min\{\sigma_i(X), \sigma_i(Y)\} \quad \text{and} \quad \gamma_i := \sigma_i(X - Y).$$

Then we have

$$\sum_{s=1}^{k} \alpha_{i_s + j_s - s} \leq \sum_{s=1}^{k} \beta_{i_s} + \sum_{s=1}^{k} \gamma_{j_s}. \tag{2.16}$$

Notice that the inequality (2.16) includes (2.12) as a special case. Numerical tests show that Conjecture 2.10 seems to be true. However, to the best of our knowledge, this conjecture has not been proved theoretically yet. It is interesting to notice that Conjecture 2.10 is stronger than Conjecture 2.7, as can be seen as follows:

Theorem 2.11. *If Conjecture 2.10 holds, then Conjecture 2.7 holds as well.*

Proof. It is easy to see from the definition that

$$\alpha_i \geq \alpha_{i+1}, \quad \beta_i \geq \beta_{i+1}, \quad \gamma_i \geq \gamma_{i+1} \quad \text{and} \quad \alpha_i \geq \beta_i \quad \forall\, 1 \leq i \leq n.$$

Moreover, the inequality (2.12) can be equivalently written as

$$\sum_{i=1}^{k} \alpha_i \leq \sum_{i=1}^{k} \beta_i + \sum_{i=1}^{k} \gamma_i, \quad 1 \leq k \leq n.$$

For any fixed k with $1 \leq k \leq n$, let $\theta := \sum_{i=1}^{k} (\beta_i + \gamma_i - \alpha_i) \geq 0$. Now, we aim to show that

$$(\alpha_1 + \theta, \alpha_2, \ldots, \alpha_k, 0, \ldots, 0)^T \succ (\beta_1, \gamma_1, \ldots, \beta_k, \gamma_k)^T. \tag{2.17}$$

It is clear that both sides have the equal sum of all components. Then, by observing the largest $l, 1 \leq l \leq k$ components of vectors on both sides, this majorization (2.17) reduces to

$$\sum_{i=1}^{l} \alpha_i + \theta \geq \max_{0 \leq j \leq l} \left\{ \sum_{i=1}^{j} \beta_i + \sum_{i=1}^{l-j} \gamma_i \right\}, \quad 1 \leq l \leq k. \tag{2.18}$$

For any fixed l with $1 \leq l \leq k$, choose $l+1$ pairs of index sets as follows:

$$\begin{aligned}
&(I_0, J_0) = \big(\{1, \ldots, k-l\}, \{l+1, \ldots, k\}\big),\\
&(I_1, J_1) = \big(\{2, \ldots, k-l+1\}, \{l, \ldots, k-1\}\big),\\
&\cdots\cdots\\
&(I_l, J_l) = \big(\{l+1, \ldots, k\}, \{1, \ldots, k-l\}\big).
\end{aligned}$$

Suppose that Conjecture 2.10 holds. It immediately follows that

$$\sum_{i=l+1}^{k} \alpha_i \leq \sum_{I_j} \beta_i + \sum_{J_j} \gamma_j, \quad 0 \leq j \leq l,$$

which implies that

$$\begin{aligned}
\sum_{i=1}^{l} \alpha_i + \theta &= \sum_{i=1}^{k} (\beta_i + \gamma_i) - \sum_{i=l+1}^{k} \alpha_i\\
&\geq \sum_{i=1}^{k} (\beta_i + \gamma_i) - \left(\sum_{I_j} \beta_i + \sum_{J_j} \gamma_j \right)\\
&= \left(\sum_{i=1}^{j} + \sum_{i=k-l+j+1}^{k} \right) \beta_i + \left(\sum_{i=1}^{l-j} + \sum_{i=k-j+1}^{k} \right) \gamma_i\\
&\geq \sum_{i=1}^{j} \beta_i + \sum_{i=1}^{l-j} \gamma_i, \quad 1 \leq j \leq l.
\end{aligned}$$

Thus, the inequality (2.18) holds and thus the majorization (2.17) holds. Moreover, since $f : \mathbb{R}_+ \to \mathbb{R}_+$ is concave with $f(0) = 0$, from Proposition 2.3, we have

$$\sum_{i=1}^{k} f(\alpha_i) \leq f(\alpha_1 + \theta) + \sum_{i=2}^{k} f(\alpha_k) \leq \sum_{i=1}^{k} f(\beta_i) + \sum_{i=1}^{k} f(\gamma_i), \quad 1 \leq k \leq n,$$

which is equivalent to (2.14). Thus, we complete the proof. □

Finally, we remark that the techniques used by Yue and So [188, Theorem 5] for proving the case $k = n$ can also be slightly modified to prove the case $k \leq n$. This extension, together with the techniques used in the proof of Theorem 2.8,

can prove that Conjecture 2.7 holds true. However, as Theorem 2.8 is enough for what we care for in the sequel, here, we do not include the proof of Conjecture 2.7 with the corresponding modification due to its long length. According to personal communication with So, the proof of Conjecture 2.7 may appear in the next version of [188], in which the way to relax the requirement of continuous differentiability also differs from ours.

2.5 Epi-convergence (in distribution)

Now we introduce the definition of epi-convergence, which yields the convergence of minimizers and optimal values under suitable assumptions.

Definition 2.5. *Let $\{\phi^k\}$ be a sequence of extended real-valued functions on $\mathbb{R}^n$. We say that $\{\phi^k\}$ epi-converges to ϕ, denoted by $\phi^k \xrightarrow{e} \phi$, if for every point $x \in \mathbb{R}^n$,*

$$\begin{aligned} &\liminf_{k\to\infty} \phi^k(x^k) \geq \phi(x) \quad \text{for every sequence } x^k \to x, \text{ and} \\ &\limsup_{k\to\infty} \phi^k(x^k) \leq \phi(x) \quad \text{for some sequence } x^k \to x. \end{aligned}$$

In this case, the function ϕ is called the epi-limit of $\{\phi^k\}$, written as $\phi = \text{e-}\lim_k \phi^k$.

The notion of epi-convergence, albeit under the name of infimal convergence, was first introduced by Wijsman [180] for studying the relationship between the convergence of convex functions and their conjugates. The name is motivated by the fact that this convergence notion is equivalent to the set-convergence of the epigraphs, i.e.,

$$\phi^k \xrightarrow{e} \phi \quad \Longleftrightarrow \quad \operatorname{epi} \phi^k \to \operatorname{epi} \phi,$$

where the set

$$\operatorname{epi} \phi := \big\{(x,t) \in \mathbb{R}^n \times \mathbb{R} : \phi(x) \leq t\big\}$$

denotes the epigraph of the extended real-valued function ϕ, and the set convergence is in the sense of Painlevé-Kuratowski, e.g., see the definition in [155, Chapter 4.B]. This notion is also referred to the name of Γ-convergence introduced by De Giorgi and Franzoni [30] in the calculus of variations. It is well-known that continuous convergence, a fortiori uniform convergence, implies epi-convergence. Moreover, in general, epi-convergence neither implies nor is implied by pointwise convergence, unless certain properties are satisfied.

A function $\phi : \mathbb{R}^n \to [-\infty, \infty]$ is said to be lower semicontinuous (l.s.c. for short) at $\overline{x}$ if

$$\liminf_{x \to \overline{x}} \phi(x) \geq \phi(\overline{x}), \quad \text{or equivalently} \quad \liminf_{x \to \overline{x}} \phi(x) = \phi(\overline{x}),$$

and lower semicontinuous over $\mathbb{R}^n$ if this holds for every $\overline{x} \in \mathbb{R}^n$. Let $\operatorname{cl} \phi$ denote the closure of ϕ, i.e., the greatest of all the l.s.c. functions ψ such that $\psi \leq \phi$. The following results (see [155, Chapter 7]) will be used in the sequel.

Proposition 2.12. *If the sequence $\{\phi^k\}$ is nondecreasing (i.e., $\phi^{k+1} \geq \phi^k$), then* $\operatorname{e-lim}_k \phi^k$ *exists and equals the pointwise supremum of $\{\operatorname{cl} \phi^k\}$, i.e., $\phi^k \xrightarrow{e} \sup_k(\operatorname{cl} \phi^k)$.*

Proposition 2.13. *Let $\{\phi^k\}$ be a sequence of extended real-valued functions and let ψ be a continuous extended real-valued functions. If $\phi^k \xrightarrow{e} \phi$, then $\phi^k + \psi \xrightarrow{e} \phi + \psi$.*

For any $\varepsilon > 0$, we say that x is an ε-minimizer of ϕ if

$$\phi(x) \leq \inf \phi + \varepsilon.$$

Then we have the following fundamental result (see [5], [40, 3.5.Theorem] or [155, Theorem 7.33]).

Theorem 2.14 (Rockafellar and Wets [155])**.** *Let $\{\phi^k\}$ be a sequence of lower semicontinuous extended real-valued functions on $\mathbb{R}^n$. Suppose that $\{\phi^k\}$ is eventually level-bounded and $\phi^k \xrightarrow{e} \phi$. Then*

$$\inf \phi^k \to \inf \phi.$$

In addition, if for each k, x^k is a minimizer of ϕ^k, or generally, an ε^k-minimizer with $\varepsilon^k \downarrow 0$, then any cluster points of $\{x^k\}$ is a minimizers of ϕ. In particular, if ϕ is uniquely minimized at $\overline{x}$, then $x^k \to \overline{x}$.

Epi-convergence can be induced by a metric on the space of l.s.c. functions, i.e.,

$$\mathrm{LSC}(\mathbb{R}^n) := \{\phi : \mathbb{R}^n \to [-\infty, \infty] \mid \phi \text{ is l.s.c. and } \phi \not\equiv \infty\}.$$

This metric $d_{\mathrm{LSC}(\mathbb{R}^n)}(\cdot,\cdot)$ on $\mathrm{LSC}(\mathbb{R}^n)$ for epi-convergence is defined in terms of the metric $d_{\mathcal{C}^{n+1}}(\cdot,\cdot)$ on the space of all nonempty closed sets in $\mathbb{R}^{n+1}$ (denoted by $\mathcal{C}^{n+1}$) for Painlevé-Kuratowski set convergence. Here, $d_{\mathcal{C}^{n+1}}(\cdot,\cdot)$ denotes the so-called (integrated) set distance between two sets (see the definition in [155, Chapter 4.I]). More precisely, for any l.s.c. functions $\phi \not\equiv \infty$ and $\psi \not\equiv \infty$ on $\mathbb{R}^n$,

$$d_{\mathrm{LSC}(\mathbb{R}^n)}(\phi, \psi) := d_{\mathcal{C}^{n+1}}(\operatorname{epi}\phi, \operatorname{epi}\psi).$$

This setting of epigraph topology does not lose generality, as the epi-convergence of general functions can be characterized by the epi-convergence of their closures, due to the fact that (e.g., see [155, Proposition 7.4])

$$\phi^k \xrightarrow{e} \phi \quad \Longleftrightarrow \quad \phi \text{ is lower semicontinuous and } \operatorname{cl}\phi^k \xrightarrow{e} \phi.$$

For more details on epi-convergence, the readers may refer to [28, 155, 4].

The above characterization of epi-convergence is particularly useful for defining epi-convergence in distribution for random l.s.c. functions. Let $(\Omega, \mathcal{F}, P)$ be a probability space and let ϕ be a random l.s.c. function on $\mathbb{R}^n$. Then, $\operatorname{epi}\phi$ induces a probability measure on the Borel sets of the complete metric space $(\mathcal{C}^{n+1}, d_{\mathcal{C}^{n+1}})$. In view of this fact, the epi-convergence in distribution can be well-defined. A sequence of random l.s.c. functions $\{\phi^k\}$ on $\mathbb{R}^n$ is said to epi-converge in distribution to ϕ if the probability measures induced by $\operatorname{epi}\phi^k$ on the metric space $(\mathcal{C}^{n+1}, d_{\mathcal{C}^{n+1}})$

weakly converge to that induced by $\operatorname{epi}\phi$. Alternatively, a more visible definition of epi-convergence in distribution can be stated as follows.

Definition 2.6. *Let $\{\phi^k\}$ be a sequence of random l.s.c. functions on $\mathbb{R}^n$. We say that $\{\phi^k\}$ epi-converges in distribution to ϕ, denoted by $\phi^k \xrightarrow{e-d} \phi$, if for any rectangles $R_1, \ldots, R_l$ with open interiors $R_1^\circ, \ldots, R_l^\circ$ and any real numbers $a_1, \ldots, a_l$,*

$$\begin{aligned}
&\Pr\left\{\inf_{x\in R_1}\phi(x) > a_1, \cdots, \inf_{x\in R_l}\phi(x) > a_l\right\}\\
\le\ &\liminf_{k\to\infty}\Pr\left\{\inf_{x\in R_1}\phi^k(x) > a_1, \cdots, \inf_{x\in R_l}\phi^k(x) > a_l\right\}\\
\le\ &\limsup_{k\to\infty}\Pr\left\{\inf_{x\in R_1^\circ}\phi^k(x) \ge a_1, \cdots, \inf_{x\in R_l^\circ}\phi^k(x) \ge a_l\right\}\\
\le\ &\Pr\left\{\inf_{x\in R_1^\circ}\phi(x) \ge a_1, \cdots, \inf_{x\in R_l^\circ}\phi(x) \ge a_l\right\}
\end{aligned}$$

Epi-convergence in distribution gives us an elegant way of proving the convergence in distribution of minimizers or ε^k-minimizers. The following epi-convergence theorem of Knight [87, Theorem 1] is particularly useful in this regard (see also [73, Proposition 9]).

Theorem 2.15 (Knight [87])**.** *Let $\{\phi^k\}$ be a sequence of random l.s.c. functions such that $\phi^k \xrightarrow{e-d} \phi$. Assume that the following statements hold:*

(i) *x^k is an ε^k-minimizer of ϕ^k with $\varepsilon^k \xrightarrow{p} 0$;*

(ii) *$x^k = O_p(1)$;*

(iii) *the function ϕ has a unique minimizer $\overline{x}$.*

Then, $x^k \xrightarrow{d} \overline{x}$. In addition, if ϕ is a deterministic function, then $x^k \xrightarrow{p} \overline{x}$.

In particular, when all ϕ^k are convex functions and ϕ has a unique minimizer, we know from [65] that $\widehat{x}^k$ is guaranteed to be $O_p(1)$. In order to apply Theorem 2.15 on epi-convergence in distribution to a constrained optimization problem, we

need to transform the constrained optimization problem into an unconstrained one by using the indicator function of the feasible set. This leads to the issue of epi-convergence in distribution of the sum of two sequences of random functions.

The space of l.s.c functions can also be endowed with the topology of uniform convergence on compact sets (or compact convergence). This topology is generated by all the seminorms $\|\cdot\|_K$, defined by

$$\|\phi\|_K := \sup_{x\in K} |\phi(x)|,$$

as K ranges over all compact subsets of $\mathbb{R}^n$. This topology is stronger than both the topology of pointwise convergence and the topology of epi-convergence but weaker than the topology of uniform convergence. Indeed, $\{\phi^k\}$ converges to ϕ in this topology, denoted by $\phi^k \xrightarrow{u} \phi$, if and only if $\{\phi^k\}$ converges uniformly to ϕ on each compact set, i.e.,

$$\sup_{t\in K} |\phi^k(x) - \phi(x)| \to 0 \quad \forall \,\text{compact set } K \subseteq \mathbb{R}^n.$$

We also use "$\xrightarrow{u-d}$" to denote the weak convergence (or convergence in distribution) with respect to the topology of uniform convergence on compact sets. The following result stated in [143, Lemma 1] will be used in the sequel.

Theorem 2.16 (Pflug [143]). *Let $\{\phi^k\}$ be a sequence of random l.s.c. functions and $\{\psi^k\}$ be a sequence of deterministic l.s.c. functions. If either of the following two assumptions holds:*

(i) $\phi^k \xrightarrow{e-d} \phi$ *and* $\psi^k \xrightarrow{u} \psi$;

(ii) $\phi^k \xrightarrow{u-d} \phi$ *and* $\psi^k \xrightarrow{e} \psi$,

then $\phi^k + \psi^k \xrightarrow{e-d} \phi + \psi$.

In particular for a sequence of random convex functions, as a direct extension of Rockafellar [153, Theorem 10.8], Andersen and Gill [2, Theorem II.1] proved an

"in probability" version that the pointwise convergence in probability implies the convergence in probability (and thus in distribution) with respect to the topology of uniform convergence on compact subset, stated as follows:

Theorem 2.17 (Andersen and Gill [2]). *Let E be an open convex subset of $\mathbb{R}^n$ and let $\{\phi^k\}$ be a sequence of real-valued random convex functions on E such that for each $x \in E$, $\phi^k(x) \xrightarrow{p} \phi(x)$ as $k \to \infty$. Then ϕ is also convex and for any compact set $K \subset E$,*

$$\sup_{x \in K} |\phi^k(x) - \phi(x)| \xrightarrow{p} 0 \quad as \quad k \to \infty.$$

For more details on epi-convergence in distribution, the readers may refer to King and Wets [84], Geyer [64], Pflug [142, 143] and Knight [87].

2.6 The majorized proximal gradient method

Before we explore the subject of this section, we first briefly introduce the majorization method, which is a kind of general framework for solving optimization problems. Let $\mathbb{X}$ be a finite-dimensional real Hilbert space equipped with an inner product $\langle \cdot, \cdot \rangle$ and its induced norm $\| \cdot \|$. Let f be a real-valued function to be minimized over some subset $K \subseteq \mathbb{X}$. A function $\widetilde{f}$ is said to majorize the function f at some point $\overline{x}$ over K if

$$\widetilde{f}(\overline{x}) = f(\overline{x}) \quad \text{and} \quad \widetilde{f}(x) \geq f(x) \quad \forall\, x \in K, \tag{2.19}$$

The general principle of majorization methods for minimizing f over K is to generate a sequence $\{x^{k+1}\}$ from an initial (feasible) point x^0, by minimizing f^k instead of f over K in each iteration $k \geq 0$, i.e.,

$$x^{k+1} = \arg\min_{x \in K} \; f^k(x),$$

where the function f^k majorizes the function f at x^k over K. In order words, geometrically, the surface f^k lies above the surface f and touches the later at the point x^k. It directly follows from (2.19) that the majorization method processes the descent property:

$$f(x^{k+1}) \leq f^k(x^{k+1}) \leq f^k(x^k) = f(x^k) \quad \forall\, k \geq 0. \tag{2.20}$$

Roughly speaking, the efficiency of a majorization method depends on the quality of the majorization function in terms of two keys — (1) the deviation of the majorization functions from the original one, and (2) the difficulty of the majorization functions to be minimized computationally. Naturally, these two aims can conflict. Therefore, to some extent, it is an art to construct majorization functions, where varieties of inequalities may be used, depending on the insights into the shape of the function to be minimized. The most common setting is to systematically generate the majorization functions by a single generator $\widehat{f} : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ as

$$f^k(x) := \widehat{f}(x, x^k) \quad \forall\, k \geq 0.$$

Majorization methods under this setting are referred to as the classical Majorization-Minimization (MM) algorithms, which have been extensively studied especially in the statistical literature.

An MM algorithm first appeared as early as in the work of de Leeuw and Heiser [31] for multidimensional scaling problems, while the original idea of using a majorization function was enunciated even earlier by Ortega and Rheinboldt [138] for linear search methods. The well-known Expectation-maximization (EM) algorithm is a prominent example of an MM algorithm to maximum likelihood estimation. For details of development and applications of MM algorithms, the readers may refer to the recent survey paper [184] or several previous ones [32, 75, 11, 99, 78].

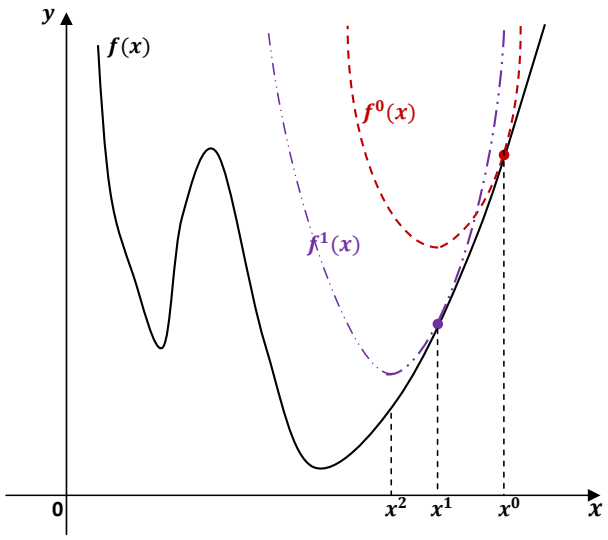


Figure 2.1: The principle of majorization methods

The way to generate the sequence $\{x^k\}$ in MM algorithms can be viewed from a different angle, expressed as

$$x^{k+1} \in \mathcal{M}(x^k),$$

where $\mathcal{M} : K \to 2^K$ is a point-to-set map defined as

$$\mathcal{M}(y) := \arg\min_{x \in K} \ \widehat{f}(x, y).$$

In view of this fact, the convergence analysis for MM algorithms can be tracked back to Zangwill's contribution in [189] on the convergence theory for algorithms derived from point-to-set maps. Later, Meyer [126] strengthened this early result to global convergence for the entire generated sequence under relatively weak hypotheses rather than the subsequential convergence. By noting the close relationship between the (generalized) fixed point of the map $\mathcal{M}(\cdot)$ and the local minimum and maximum of f, based on the results of Zangwill [189], Wu [183] established some convergence results for the EM algorithm under certain conditions. These results can be extended to MM algorithms since MM algorithms are

generalizations of the EM algorithm. Other existing convergence results of EM and MM algorithms can be found in [97, 99, 98, 172, 137, 159], to name only a few. To the best of our knowledge, most convergence results in the literature focus on convergence to interior points of the feasible set only.

The subject of this section — the majorized proximal gradient method is proposed based on the essence of the the majorization method. Let $g : \mathbb{X} \to (-\infty, \infty]$ be a proper closed convex function, $h : \mathbb{X} \to \mathbb{R}$ be a continuously differentiable function on an open set of $\mathbb{X}$ containing the domain of g denoted by $\operatorname{dom} g := \{x \in \mathbb{X} \mid g(x) < \infty\}$, and $p : \mathbb{X} \to \mathbb{R}$ be a continuous (nonconvex) function. A class of nonconvex nonsmooth optimization problems we will consider takes the form

$$\min_{x \in \mathbb{X}} \quad f(x) := h(x) + g(x) + p(x). \tag{2.21}$$

This class of optimization problems apparently look unconstrained, but actually allow constraints. The constraint $x \in K$ can be absorbed into the function g via the characteristic function δ_K provided that K is a closed convex subset of $\mathbb{X}$.

We study this class of optimization problems for preparation of the problems that will be addressed in the sequel. In particular, when $p \equiv 0$, the problem (2.21) has been explored with the proximal gradient method in some references (e.g., see [6, 59, 129, 170]). More closely-related studies can be found in Gao and Sun [62] and Gao [61], in which the proximal subgradient method was proposed to solve the problem (2.21) with the function p being concave. The proximal subgradient method is essentially treated as a majorization method, but with line search allowed if the majorization functions is not easy to construct. In other words, this method ensures the decrease of objective values by using linear search instead of using a global majorization for tractable implementation. Based on the same idea, here, we propose the majorized proximal gradient method to solve the problem (2.21) for general cases.

Algorithm 2.1. (A majorized proximal gradient method)

Step 0. *Input* $x^0 \in \mathbb{X}$. *Choose* $\tau \in (0,1)$ *and* $\delta \in (0,1)$. *Set* $k := 0$.

Step 1. *Choose a self-adjoint positive definite operator* $M^k : \mathbb{X} \to \mathbb{X}$ *and construct a convex function* p^k *that majorizes* p *at* x^k.

Step 2. *Solve the (strongly) convex optimization problem:*

$$d^k := \arg\min_{d \in \mathbb{X}} \left\{ \langle \nabla h(x^k), d \rangle + \frac{1}{2} \langle d, M^k d \rangle + g(x^k + d) + p^k(x^k + d) \right\}. \quad (2.22)$$

Step 3. *Choose* $\overline{\alpha}^k > 0$ *and let* l_k *be the smallest nonnegative integer satisfying*

$$f(x^k + \overline{\alpha}^k \tau^l d^k) \leq f(x^k) + \delta \overline{\alpha}^k \tau^l \Delta^k, \quad (2.23)$$

where

$$\Delta^k := \langle \nabla h(x^k), d^k \rangle + g(x^k + d^k) - g(x^k) + p^k(x^k + d^k) - p^k(x^k). \quad (2.24)$$

Set $\alpha^k := \overline{\alpha}^k \tau^{l_k}$ *and* $x^{k+1} := x^k + \alpha^k d^k$.

Step 4. *If converged, stop; otherwise, set* $k := k + 1$ *and go to* **Step 1**.

In particular, when $p = 0$, Algorithm 2.1 reduces to the proximal gradient method studied in [6, 59, 129, 170], and when p is a concave function, Algorithm 2.1 reduces to the proximal subgradient method studied in [62] and [61] if p^k is chosen to be

$$p^k(x) := p(x^k) + \langle G^k, x - x^k \rangle \quad \forall\, k \geq 0$$

with $G^k \in \partial p(x^k)$. In Algorithm 2.1, the Armijo rule is applied to choosing α^k due to its simplicity and efficiency. Other kinds of line search rules may also be allowed for this algorithm. The following result shows the well-definedness of the Armijo rule in Algorithm 2.1. Its proof, as well as the one for the convergence later, is in

line with the one for the proximal subgradient method in [61].

Lemma 2.18. *Let $\{x^k\}$ and $\{d^k\}$ be two sequences generated from Algorithm 2.1. Then we have that for each $k \geq 0$,*

$$\Delta^k \leq -\langle d^k, M^k d^k \rangle < 0, \tag{2.25}$$

where Δ^k is defined by (2.24) and

$$f(x^k + \alpha d^k) \leq f(x^k) + \alpha \Delta^k + o(\alpha) \quad \forall\, \alpha \in (0, 1]. \tag{2.26}$$

Moreover, assume that ∇h is Lipschitz continuous with constant $\kappa \geq 0$ over $\operatorname{dom} g$ and $\lambda_{\min}(M^k) \geq \nu > 0\ \forall\, k \geq 0$, where $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue. Then for each $k \geq 0$, we have that for any $\delta \in (0, 1)$,

$$f(x^k + \alpha d^k) \leq f(x^k) + \delta \alpha \Delta^k \quad \forall\, 0 < \alpha \leq \min\{1, 2\nu(1-\delta)/\kappa\}. \tag{2.27}$$

Proof. Since d_k is the optimal solution to the problem (2.22), we have

$$\begin{aligned}
&\langle \nabla h(x^k), d^k \rangle + \frac{1}{2}\langle d^k, M^k d^k \rangle + (g + p^k)(x^k + d^k) \\
&\leq \langle \nabla h(x^k), \alpha d^k \rangle + \frac{1}{2}\langle \alpha d^k, M^k (\alpha d^k) \rangle + (g + p^k)(x^k + \alpha d^k) \\
&\leq \alpha \langle \nabla h(x^k), d^k \rangle + \frac{\alpha^2}{2}\langle d^k, M^k d^k \rangle + \alpha (g + p^k)(x^k + d^k) + (1 - \alpha)(g + p^k)(x^k),
\end{aligned}$$

where the last inequality follows from the convexity of $g + p^k$. By rearranging the terms, we obtain

$$\Delta = \langle \nabla h(x^k), d^k \rangle + (g + p^k)(x^k + d^k) - (g + p^k)(x^k) \leq -\frac{1 + \alpha}{2}\langle d^k, M^k d^k \rangle.$$

Then, by letting $\alpha \uparrow 1$, we can obtain (2.25) since M is positive definite. Moreover, since p^k majorizes p at x^k, together with the continuous differentiability of h and

the convexity of $g + p^k$, we have that for any $\alpha \in (0, 1]$,

$$\begin{aligned}
& f(x^k + \alpha d^k) - f(x^k) \\
&= (h + g + p)(x^k + \alpha d^k) - (h + g + p)(x^k) \\
&\le (h + g + p^k)(x^k + \alpha d^k) - (h + g + p^k)(x^k) \\
&\le \langle \nabla h(x^k), \alpha d^k \rangle + o(\alpha) + \alpha(g + p^k)(x^k + d^k) + (1 - \alpha)(g + p^k)(x^k) - (g + p^k)(x^k) \\
&= \alpha \left(\langle \nabla h(x^k), d^k \rangle + (g + p^k)(x^k + d^k) - (g + p^k)(x^k) \right) + o(\alpha) \\
&= \alpha \Delta^k + o(\alpha), \qquad (2.28)
\end{aligned}$$

which proves (2.26). If in addition ∇h is Lipschitz continuous with constant $\kappa \ge 0$ over $\operatorname{dom} g$, then from the fundamental theorem of calculus, we have

$$\begin{aligned}
h(x^k + \alpha d^k) - h(x^k) &= \langle \nabla h(x^k), \alpha d^k \rangle + \int_0^1 \langle \nabla h(x^k + t\alpha d^k) - \nabla h(x^k), \alpha d^k \rangle dt \\
&\le \alpha \langle \nabla h(x^k), d^k \rangle + \int_0^1 \| \nabla h(x^k + t\alpha d^k) - \nabla h(x^k) \| \| \alpha d^k \| dt \\
&\le \alpha \langle \nabla h(x^k), d^k \rangle + \frac{1}{2} \alpha^2 \kappa \| d \|^2.
\end{aligned}$$

This implies that the term $o(\alpha)$ in (2.28) can be replaced by $\frac{1}{2}\alpha^2\kappa\|d^k\|^2$. When $\lambda_{\min}(M^k) \ge \nu > 0 \; \forall \, k \ge 0$, we further have for any $0 < \alpha \le 2\nu(1 - \delta)/\kappa$,

$$\frac{1}{2}\alpha^2\kappa\|d^k\|^2 \le \alpha\nu(1 - \delta)\|d^k\|^2 \le \alpha(1 - \delta)\langle d^k, M^k d^k \rangle \le -\alpha(1 - \delta)\Delta^k.$$

This, together with (2.28), proves (2.27). Thus, we complete the proof. □

To analyze the convergence of the majorized proximal gradient method, we assume that the convex majorization function $p^k, k \ge 0$ are constructed from a single generator $\widehat{p}(x, y) : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ as

$$p^k(x) := \widehat{p}(x, x^k) \quad \forall \, k \ge 0. \qquad (2.29)$$

This assumption provides a global connection between all the majorization functions $p^k, k \ge 0$ to make the convergence analysis possible.

Theorem 2.19. *Let $\{x^k\}$ and $\{d^k\}$ be the sequences generated from Algorithm 2.1. Assume that $0 < \nu \le \lambda_{\min}(M^k) \le \mu < \infty \ \forall\, k \ge 0$. Then $\{f(x^k)\}$ is monotonically decreasing satisfying that for each $k \ge 0$,*

$$f(x^{k+1}) - f(x^k) \le \delta\alpha^k\Delta^k \le -\delta\alpha^k\nu\|d^k\|^2. \tag{2.30}$$

In addition, suppose that $\inf \overline{\alpha}^k > 0$, then the following results holds:

(i) *If $\{x^{k_j}\}$ is a convergent subsequence of $\{x^k\}$, then $\lim_{j\to\infty} d^{k_j} = 0$.*

(ii) *If $p^k, k \ge 0$ are constructed from (2.29) with $\widehat{p}(x, y)$ being continuous, then any limit point $\overline{x}$ of $\{x^k\}$ satisfies*

$$0 \in \partial(h + g + \widehat{p}_{\overline{x}})(\overline{x}) = \nabla h(\overline{x}) + \partial g(\overline{x}) + \partial\widehat{p}_{\overline{x}}(\overline{x}),$$

where $\widehat{p}_{\overline{x}}(x) := \widehat{p}(x, \overline{x})$.

Proof. Since $\lambda_{\min}(M^k) \ge \nu > 0$, from Lemma 2.18, we have

$$\Delta^k \le -\langle d^k, M^k d^k\rangle \le -\nu\|d^k\|^2. \tag{2.31}$$

Then, the inequality (2.30) is immediate from (2.23) and (2.31). This directly implies that $\{f(x^k)\}$ is monotonically decreasing.

(i) Suppose that $\lim_{j\to\infty} x^{k_j} \to \overline{x}$. The semicontinuity of f implies $f(\overline{x}) \le \liminf_{j\to\infty} f(x^{k_j})$. This, together with the monotonic decreasing property of $\{f(x^k)\}$, implies that $\{f(x^k)\}$ converges to a finite limit. Hence, $\{f(x^{k+1}) - f(x^k)\}$ converges to 0. Then, from (2.30), we obtain that

$$\lim_{k\to\infty} \alpha^k\Delta^k = 0. \tag{2.32}$$

Now we prove $\lim_{j\to\infty} d^{k_j} = 0$ by contradiction. Suppose not. Then by passing to a subsequence if necessary, there exists some $\gamma > 0$ such that $\|d^{k_j}\| \ge \gamma \ \forall\, j \ge 0$. It then follows from (2.31) that $\Delta^{k_j} \le -\nu\gamma \ \forall\, j \ge 0$. This, together with (2.32),

implies that $\lim_{j\to\infty}\alpha^{k_j}=0$. Recall that $\alpha^{k_j}=\overline{\alpha}^{k_j}\tau^{l_{k_j}}$ and $\inf_k\overline{\alpha}^k>0$. Then there exists some index $\overline{j}$ such that $\alpha^{k_j}<\overline{\alpha}^{k_j}$ and $\alpha^{k_j}\le\tau$ for all $j\ge\overline{j}$. Furthermore, from the choice of α^{k_j} (2.23), we have

$$f\big(x^{k_j}+(\alpha^{k_j}/\tau)d^{k_j}\big)>f(x^{k_j})+\delta(\alpha^{k_j}/\tau)\Delta^k\quad\ \forall\, j\ge\overline{j}.$$

Thus, for all $j\ge\overline{j}$, we have

$$\begin{aligned}\delta\Delta^{k_j}&<\frac{(h+g+p^{k_j})(x^{k_j}+(\alpha^{k_j}/\tau)d^{k_j})-(h+g+p^{k_j})(x^{k_j})}{\alpha^{k_j}/\tau}\\&\le\frac{h(x^{k_j}+(\alpha^{k_j}/\tau)d^{k_j})-h(x^{k_j})}{\alpha^{k_j}/\tau}+(g+p^{k_j})(x^{k_j}+d^{k_j})-(g+p^{k_j})(x^{k_j}).\end{aligned}$$

Using the definition of Δ^k, the last inequality can be rewritten as

$$\frac{h(x^{k_j}+(\alpha^{k_j}/\tau)d^{k_j})-h(x^{k_j})}{\alpha^{k_j}/\tau}-\langle\nabla h(x^{k_j}),d^{k_j}\rangle\ge-(1-\delta)\Delta^{k_j}.$$

Dividing both sides by $\|d^k\|$ and using (2.31) yields

$$\frac{h(x^{k_j}+\widehat{\alpha}^{k_j}d^{k_j}/\|d^{k_j}\|)-h(x^{k_j})}{\widehat{\alpha}^{k_j}}-\frac{\langle\nabla h(x^{k_j}),d^{k_j}\rangle}{\|d^{k_j}\|}\ge-(1-\delta)\frac{\Delta^{k_j}}{\|d^{k_j}\|}\ge(1-\delta)\|d^{k_j}\|,\tag{2.33}$$

where $\widehat{\alpha}^{k_j}:=(\alpha^{k_j}/\tau)\|d^{k_j}\|$. Note that $-\alpha^{k_j}\Delta^{k_j}\ge\nu\alpha^{k_j}\|d^{k_j}\|^2\ge\nu\gamma\alpha^{k_j}\|d^{k_j}\|$ for all $j\ge\overline{j}$. Thus, (2.32) implies that $\{\alpha^{k_j}\|d^{k_j}\|\}$ converges 0 and so is $\{\widehat{\alpha}^{k_j}\}$. In addition, since $\{d^{k_j}/\|d^{k_j}\|\}$ is bounded, by passing to a subsequence if necessary, we assume that $\{d^{k_j}/\|d^{k_j}\|\}$ converges to some point $\overline{d}$. Now letting $j\to\infty$ in (2.33), we obtain

$$0=\langle\nabla f(\overline{x}),\overline{d}\rangle-\langle\nabla f(\overline{x}),\overline{d}\rangle\ge(1-\delta)\tau>0,$$

which is a clear contradiction. Thus, $\{d^{k_j}\}$ converges to 0.

(ii) Suppose that $\lim_{j\to\infty}x^{k_j}=\overline{x}$. Since d^{k_j} is the optimal solution to the problem (2.22). Then there exists some $G^{k_j}\in\partial(g+p^{k_j})(x^{k_j}+d^{k_j})$ such that

$$0=\nabla h(x^{k_j})+M^{k_j}d^{k_j}+G^{k_j}.\tag{2.34}$$

Moreover, we know from [153, Theorem 24.7] that $\{G^{k_j}\}$ is bounded. By passing to a subsequence if necessary, we assume that $\lim_{j\to\infty} G^{k_j} = \overline{G}$. Then by letting $j \to \infty$ in (2.34), we obtain that

$$0 = \nabla h(\overline{x}) + \overline{G}. \tag{2.35}$$

Moreover, $G^{k_j} \in \partial(g + p^{k_j})(x^{k_j} + d^{k_j})$ implies that

$$(g + p^{k_j})(x) \geq (g + p^{k_j})(x^{k_j} + d^{k_j}) + \langle G^{k_j}, x - (x^{k_j} + d^{k_j})\rangle \quad \forall\, x \in \mathbb{X}.$$

By letting $j \to \infty$, from (2.29), the continuity of $\widehat{p}$ and (ii), we obtain

$$g(x) + \widehat{p}(x, \overline{x}) \geq g(\overline{x}) + \widehat{p}(\overline{x}, \overline{x}) + \langle \overline{G}, x - \overline{x}\rangle \quad \forall\, x \in \mathbb{X},$$

which implies that $\overline{G} \in \partial(g + \widehat{p}_{\overline{x}})(\overline{x})$. This, together with (2.35) and [153, Theorem 23.8], proves (ii). □

To make a closer look at Algorithm 2.1, the Armijo line search rule can be regarded as providing a local quadratic majorization of the function h, which depends on the choice M^k in each iteration. This arouses us to ask if the function h can be globally majorized, whether the line search can be removed from Algorithm 2.1. We turn back to Lemma 2.18. In Lemma 2.18, if ∇h is Lipschitz continuous with constant $\kappa > 0$, then by choosing $M^k \equiv \kappa I$ (here I stands for the identity operator), the inequality (2.27) holds for $\alpha = 1$ provided $\delta \in (0, 1/2]$. More generally, suppose that in each iteration $k \geq 0$,

$$h(x) \leq h^k(x) := h(x^k) + \langle \nabla h(x^k), x - x^k\rangle + \frac{1}{2}\langle x - x^k, M^k(x - x^k)\rangle \quad \forall\, x \in \mathbb{X}. \tag{2.36}$$

In other words, the function h is globally majorized by a quadratic convex function h^k at x^k. In this case, (2.28) can be explicitly written as

$$\begin{aligned} f(x^k + \alpha d^k) &\leq f(x^k) + \alpha\Delta^k + \frac{1}{2}\alpha^2\langle d^k, M^k d^k\rangle \\ &\leq f(x^k) + \Big(\alpha - \frac{1}{2}\alpha^2\Big)\Delta^k \quad \forall\, \alpha \in (0, 1]. \end{aligned}$$

This implies that the update $x^{k+1} = x^k + d^k$ with the step length $\alpha^k \equiv 1$ in Algorithm 2.1 is applicable if (2.36) holds. Therefore, the answer to our question is positive. The simplified majorized proximal gradient method without line search is described as follows:

Algorithm 2.2. (A majorized proximal gradient method without line search)

Step 0. *Input* $x^0 \in \mathbb{X}$. *Set* $k := 0$.

Step 1. *Construct a quadratic convex function* h^k *that majorized* h *at* x^k *as (2.36) and construct a convex function* p^k *that majorizes* p *at* x^k.

Step 2. *Solve the (strongly) convex optimization problem:*

$$x^{k+1} := \arg\min_{x \in \mathbb{X}} \left\{ h^k(x) + g(x) + p^k(x) \right\}. \tag{2.37}$$

Step 3. *If converged, stop; otherwise, set* $k := k + 1$ *and go to* **Step 1**.

Algorithm 2.2 is particularly useful for the case that ∇h is Lipschitz continuous. As the function $h^k + g + p^k$ in (2.37) majorizes the function $h + g + p$ at x^k for all $k \geq 0$, Algorithm 2.2 is actually a majorization method for solving the original nonconvex nonsmooth problem (2.21). In particular, if $h(x)$ vanishes and $p(x)$ is convex, Algorithm 2.2 with the choice $h^k(x) := \frac{\gamma^k}{2}\|x - x^k\|^2, \gamma_k > 0$ and $p^k(x) = p(x)$ reduces to the well-known (primal) proximal point algorithm studied in [119, 154] for solving the convex optimization problem. In addition, the convergence result in Theorem 2.19 for Algorithm 2.1 can also be adapted to Algorithm 2.2.

Chapter 3

Matrix completion with fixed basis coefficients

In this chapter, we address low-rank matrix completion problems with fixed basis coefficients. The unknown matrix for recovery could be rectangular, symmetric/Hermitian, or further symmetric/Hermition positive semidefinite. To discuss all these cases simultaneously, throughout this chapter, we use a unified symbol $\mathbb{V}^{n_1\times n_2}$ to denote the matrix space we concern, i.e., $\mathbb{R}^{n_1\times n_2}$, $\mathbb{C}^{n_1\times n_2}$, $\mathcal{S}^n$ or $\mathcal{H}^n$.

The organization of this chapter is as follows: In Section 3.1, we introduce the observation model of matrix completion with fixed basis coefficients and the formulation of the rank-correction step. In Section 3.2, we establish a non-asymptotic recovery error bound for the rank-correction step and discuss the impact of the rank-correction term on reducing the recovery error. In Section 3.3, we derive necessary and sufficient conditions for rank consistency of the rank-correction step in the sense of Bach [7], in which constraint nondegeneracy for conic optimization problem plays a key role in our analysis. The construction of the rank-correction function is discussed in Section 3.4. Numerical results are reported in Section 3.5 to validate the efficiency of our proposed rank-corrected procedure.

3.1 Problem formulation

In this section, we formulate the model of the matrix completion problem with fixed basis coefficients, and then propose a rank-correction step for solving this class of problems.

3.1.1 The observation model

Let $\{\Theta_1, \ldots, \Theta_d\}$ be a given orthonormal basis of the given real inner product space $\mathbb{V}^{n_1 \times n_2}$. Then, any matrix $X \in \mathbb{V}^{n_1 \times n_2}$ can be uniquely expressed in the form of

$$X = \sum_{k=1}^{d} \langle \Theta_k, X \rangle \Theta_k,$$

where $\langle \Theta_k, X \rangle$ is called the basis coefficient of X relative to Θ_k. Let $\overline{X} \in \mathbb{V}^{n_1 \times n_2}$ be the unknown low-rank matrix to be recovered. In some practical applications, for example, the correlation and density matrix completion, a few basis coefficients of the unknown matrix $\overline{X}$ are fixed (or assumed to be fixed) due to a certain structure or reliable prior information. Throughout this paper, we let $\alpha \subseteq \{1, 2, \ldots, d\}$ denote the set of the indices relative to which the basis coefficients are fixed, and β denote the complement of α in $\{1, 2, \ldots, d\}$, i.e., $\alpha \cap \beta = \emptyset$ and $\alpha \cup \beta = \{1, \ldots, d\}$. We define $d_1 := |\alpha|$ and $d_2 := |\beta|$.

When a few basis coefficients are fixed, one only needs to observe the rest for recovering the unknown matrix $\overline{X}$. Assume that we are given a collection of m noisy observations of the basis coefficients relative to $\{\Theta_k : k \in \beta\}$ in the following form

$$y_i = \left\langle \Theta_{\omega_i}, \overline{X} \right\rangle + \nu \xi_i, \quad i = 1, \ldots, m, \tag{3.1}$$

where ω_i are the indices randomly sampled from the index set β, ξ_i are the independent and identically distributed (i.i.d.) noises with $\mathbb{E}(\xi_i) = 0$ and $\mathbb{E}(\xi_i^2) = 1$, and

$\nu > 0$ controls the magnitude of noise. Unless otherwise stated, we assume a general weighted sampling (with replacement) scheme with the sampling distributions of ω_i as follows.

Assumption 3.1. *The indices $\omega_1, \ldots, \omega_m$ are i.i.d. copies of a random variable ω that has a probability distribution Π over $\{1, \ldots, d\}$ defined by*

$$\Pr(\omega = k) = \begin{cases} 0 & \text{if } k \in \alpha, \\ p_k > 0 & \text{if } k \in \beta. \end{cases}$$

Note that each $\Theta_k, k \in \beta$ is assumed to be sampled with a positive probability in this sampling scheme. In particular, when the sampling probability of all $k \in \beta$ are equal, i.e., $p_k = 1/d_2 \ \forall\, k \in \beta$, we say that the observations are sampled uniformly at random.

Next, we present some examples of low-rank matrix completion problems in the above settings.

- **Correlation matrix completion:**

  A correlation matrix is an $n \times n$ real symmetric or Hermitian positive semidefinite matrix with all diagonal entries being ones. Then, The recovery of a correlation matrix is based on the observations of entries.

  (i) For the real case, $\mathbb{V}^{n_1 \times n_2} = \mathcal{S}^n$, $d = n(n+1)/2$, $d_1 = n$,

  $$\Theta_\alpha = \left\{ e_i e_i^{\mathbb{T}} \;\middle|\; 1 \le i \le n \right\} \quad \text{and} \quad \Theta_\beta = \left\{ \frac{1}{\sqrt{2}}(e_i e_j^{\mathbb{T}} + e_j e_i^{\mathbb{T}}) \;\middle|\; 1 \le i < j \le n \right\}.$$

  (ii) For the complex case, $\mathbb{V}^{n_1 \times n_2} = \mathcal{H}^n$, $d = n^2$, $d_1 = n$,

  $$\Theta_\alpha = \left\{ e_i e_i^{\mathbb{T}} \;\middle|\; 1 \le i \le n \right\} \text{ and } \Theta_\beta = \left\{ \frac{1}{\sqrt{2}}(e_i e_j^{\mathbb{T}} + e_j e_i^{\mathbb{T}}), \frac{\sqrt{-1}}{\sqrt{2}}(e_i e_j^{\mathbb{T}} - e_j e_i^{\mathbb{T}}) \middle| i < j \right\}.$$

  Here, $\sqrt{-1}$ represents the imaginary unit. Of course, one may fix some off-diagonal entries in specific applications.

- **Density matrix completion:**

  A density matrix of dimension $n = 2^l$ for some positive integer l is an $n \times n$ Hermitian positive semidefinite matrix with trace one. In quantum state tomography, one aims to recover a density matrix from Pauli measurements (i.e.,observations of the coefficients relative to the Pauli basis) [71, 53], given by

  $$\Theta_\alpha = \left\{\frac{1}{\sqrt{n}} I_n\right\} \text{ and } \Theta_\beta = \left\{\frac{1}{\sqrt{n}}(\sigma_{s_1} \otimes \cdots \otimes \sigma_{s_l}) \,\middle|\, (s_1, \ldots, s_l) \in \{0,1,2,3\}^l\right\} \Big\backslash \Theta_\alpha,$$

  where "$\otimes$" means the Kronecker product of two matrices and

  $$\sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},\ \sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},\ \sigma_2 = \begin{pmatrix} 0 & -\sqrt{-1} \\ \sqrt{-1} & 0 \end{pmatrix},\ \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

  are the Pauli matrices. In this setting, $\mathbb{V}^{n_1 \times n_2} = \mathcal{H}^n$, $\mathrm{Tr}(\overline{X}) = \langle I_n, \overline{X} \rangle = 1$, $d = n^2$, and $d_1 = 1$.

- **Rectangular matrix completion:**

  Assume that a few entries of a rectangular matrix are known and let $\mathcal{I}$ be the index set of these entries. One aims to recover this rectangular matrix from the observations of the rest entries.

  (i) For the real case, $\mathbb{V}^{n_1 \times n_2} = \mathbb{R}^{n_1 \times n_2}$, $d = n_1 n_2$, $d_1 = |\mathcal{I}|$,

  $$\Theta_\alpha = \left\{e_i e_j^{\mathbb{T}} \,\middle|\, (i,j) \in \mathcal{I}\right\} \quad \text{and} \quad \Theta_\beta = \left\{e_i e_j^{\mathbb{T}} \,\middle|\, (i,j) \notin \mathcal{I}\right\}.$$

  (ii) For the complex case, $\mathbb{V}^{n_1 \times n_2} = \mathbb{C}^{n_1 \times n_2}$, $d = 2n_1 n_2$, $d_1 = 2|\mathcal{I}|$,

  $$\Theta_\alpha = \left\{e_i e_j^{\mathbb{T}}, \sqrt{-1} e_i e_j^{\mathbb{T}} \,\middle|\, (i,j) \in \mathcal{I}\right\} \quad \text{and} \quad \Theta_\beta = \left\{e_i e_j^{\mathbb{T}}, \sqrt{-1} e_i e_j^{\mathbb{T}} \,\middle|\, (i,j) \notin \mathcal{I}\right\}.$$

For convenience of discussion, we first introduce some linear operators that are frequently used in the subsequent sections. For any given index set $\pi \subseteq \{1, \ldots, d\}$,

e.g., $\pi = \alpha$ or $\pi = \beta$, we define the linear operators $\mathcal{R}_\pi$: $\mathbb{V}^{n_1\times n_2} \to \mathbb{R}^{|\pi|}$ and $\mathcal{P}_\pi$: $\mathbb{V}^{n_1\times n_2} \to \mathbb{V}^{n_1\times n_2}$, respectively, by

$$\mathcal{R}_\pi(X) := \big(\langle \Theta_k, X\rangle\big)_{k\in\pi}^{\mathbb{T}} \quad \text{and} \quad \mathcal{P}_\pi(X) := \sum_{k\in\pi}\langle \Theta_k, X\rangle\Theta_k, \qquad X \in \mathbb{V}^{n_1\times n_2}. \tag{3.2}$$

It is easy to see that $\mathcal{P}_\pi = \mathcal{R}_\pi^*\mathcal{R}_\pi$. Define the self-adjoint operators $\mathcal{Q}_\beta : \mathbb{V}^{n_1\times n_2} \to \mathbb{V}^{n_1\times n_2}$ and $\mathcal{Q}_\beta^\dagger : \mathbb{V}^{n_1\times n_2} \to \mathbb{V}^{n_1\times n_2}$ associated with the sampling probability, respectively, by

$$\mathcal{Q}_\beta(X) := \sum_{k\in\beta} p_k\langle \Theta_k, X\rangle\Theta_k \quad \text{and} \quad \mathcal{Q}_\beta^\dagger(X) := \sum_{k\in\beta}\frac{1}{p_k}\langle \Theta_k, X\rangle\Theta_k, \quad X\in\mathbb{V}^{n_1\times n_2}. \tag{3.3}$$

One may easily verify that the operators $\mathcal{Q}_\beta$, $\mathcal{Q}_\beta^\dagger$ and $\mathcal{P}_\beta$ satisfy the following relations

$$\mathcal{Q}_\beta\mathcal{Q}_\beta^\dagger = \mathcal{Q}_\beta^\dagger\mathcal{Q}_\beta = \mathcal{P}_\beta, \quad \mathcal{P}_\beta\mathcal{Q}_\beta = \mathcal{Q}_\beta\mathcal{P}_\beta = \mathcal{Q}_\beta \quad \text{and} \quad \mathcal{Q}_\beta^\dagger\mathcal{R}_\alpha^* = 0. \tag{3.4}$$

Let Ω be the multiset of all the sampled indices from the index set β, i.e.,

$$\Omega := \{\omega_1, \ldots, \omega_m\}.$$

With a slight abuse on notation, we define the sampling operator $\mathcal{R}_\Omega$: $\mathbb{V}^{n_1\times n_2} \to \mathbb{R}^m$ associated with Ω by

$$\mathcal{R}_\Omega(X) := \big(\langle \Theta_{\omega_1}, X\rangle, \ldots, \langle \Theta_{\omega_m}, X\rangle\big)^{\mathbb{T}}, \quad X \in \mathbb{V}^{n_1\times n_2}.$$

Then, the observation model (3.1) can be expressed in the following vector form

$$y = \mathcal{R}_\Omega(\overline{X}) + \nu\xi, \tag{3.5}$$

where $y = (y_1, \ldots, y_m)^{\mathbb{T}} \in \mathbb{R}^m$ and $\xi = (\xi_1, \ldots, \xi_m)^{\mathbb{T}} \in \mathbb{R}^m$ denote the observation vector and the noise vector, respectively.

3.1.2 The rank-correction step

It is obvious that a small portion of (noisy) observations may be generated from lots of different matrices. However, the low-rank structure of the unknown matrix could drastically reduce the amount of candidate matrices. This property makes it possible to reconstruct the unknown low-rank matrix by minimizing the deviation from observations and the rank simultaneously. This natural idea leads to the rank penalized least squares estimator for recovery as follows:

$$\begin{aligned} \min_{X\in\mathbb{V}^{n_1\times n_2}} \quad & \frac{1}{2m}\left\|y-\mathcal{R}_\Omega(X)\right\|_2^2+\rho_m \mathrm{rank}(X) \\ \text{s.t.} \quad & \mathcal{R}_\alpha(X)=\mathcal{R}_\alpha(\overline{X}), \end{aligned} \tag{3.6}$$

where $\rho_m > 0$ is a parameter depending on the number of observations to control the tradeoff between the deviation and the rank. However, the rank function is discontinuous and nonconvex, which makes the optimization problem (3.6) NP-hard in general.

The nuclear norm, i.e., the sum of all singular values, is the convex envelope of the rank function over a unit ball of spectral norm [49]. In many situations, the nuclear norm has been demonstrated to be a successful alternative to the rank function for matrix recovery, see e.g., [49, 150, 21, 22, 70, 19, 149, 91, 135, 86]. Applying this technique to our matrix completion model brings us the nuclear norm penalized least square estimator instead of (3.6) as follows:

$$\begin{aligned} \min_{X\in\mathbb{V}^{n_1\times n_2}} \quad & \frac{1}{2m}\left\|y-\mathcal{R}_\Omega(X)\right\|_2^2+\rho_m \|X\|_* \\ \text{s.t.} \quad & \mathcal{R}_\alpha(X)=\mathcal{R}_\alpha(\overline{X}). \end{aligned} \tag{3.7}$$

This convex optimization problem is computationally tractable. However, its efficiency for encouraging a low-rank solution is not universal. The efficiency may be challenged if the observations are sampled at random obeying a general distribution, particularly for the case considered in [158] where certain rows and/or

columns are sampled with high probability. The setting of fixed basis coefficients in our matrix completion model can be regarded to be under an extreme sampling scheme. In particular, for the correlation and density matrix completion, the nuclear norm completely loses its efficiency for low rank since in this case it reduces to a constant. In order to overcome the shortcomings of the nuclear norm penalization, we propose a rank-correction step to generate an estimator with a better recovery performance.

Given a spectral operator $F : \mathbb{V}^{n_1\times n_2} \to \mathbb{V}^{n_1\times n_2}$ (see Definition 2.2) and an initial estimator $\widetilde{X}_m$ for the unknown matrix $\overline{X}$, say the nuclear norm penalized least squares estimator or one of its analogies, our rank-correction step is to solve the convex optimization problem

$$\begin{array}{ll} \min\limits_{X\in\mathbb{V}^{n_1\times n_2}} & \dfrac{1}{2m}\|y-\mathcal{R}_\Omega(X)\|_2^2+\rho_m\left(\|X\|_*-\langle F(\widetilde{X}_m),X\rangle+\dfrac{\gamma_m}{2}\|X-\widetilde{X}_m\|_F^2\right)\\ \text{s.t.} & \mathcal{R}_\alpha(X)=\mathcal{R}_\alpha(\overline{X}), \end{array} \tag{3.8}$$

where $\rho_m > 0$ and $\gamma_m \geq 0$ are the regularization parameters depending on the number of observations. The last quadratic proximal term is added to guarantee the boundness of the solution to (3.8). If the function $\|X\|_* - \langle F(\widetilde{X}_m), X\rangle$ is level-bounded, one may simply set $\gamma_m = 0$. Clearly, when $F \equiv 0$ and $\gamma_m = 0$, the problem (3.8) reduces to the nuclear norm penalized least squares problem. In the sequel, we call $-\langle F(\widetilde{X}_m), X\rangle$ the rank-correction term. If the true matrix is known to be positive semidefinite, we add the constraint $X \in \mathbb{S}^n_+$ to (3.8). Thus, the rank-correction step is to solve the convex conic optimization problem

$$\begin{array}{ll} \min\limits_{X\in\mathbb{S}^n} & \dfrac{1}{2m}\|y-\mathcal{R}_\Omega(X)\|_2^2+\rho_m\left(\langle I-F(\widetilde{X}_m),X\rangle+\dfrac{\gamma_m}{2}\|X-\widetilde{X}_m\|_F^2\right)\\ \text{s.t.} & \mathcal{R}_\alpha(X)=\mathcal{R}_\alpha(\overline{X}),\quad X\in\mathbb{S}^n_+. \end{array} \tag{3.9}$$

For this case, we assume that the initial estimator $\widetilde{X}_m$ belongs to $\mathbb{S}^n_+$ as the projection of any estimator onto $\mathbb{S}^n_+$ can approximate the true matrix $\overline{X}$ better.

The rank-correction step above is inspired by the majorized penalty approach recently proposed by Gao and Sun [62] for solving the rank constrained matrix optimization problem:

$$\begin{aligned} \min \;\; & h(X) \\ \text{s.t.} \;\; & \operatorname{rank}(X) \le r, \quad X \in K, \end{aligned} \tag{3.10}$$

where $r \ge 1$, $h : \mathbb{V}^{n_1 \times n_2} \to \mathbb{R}$ is a given continuous function and $K \in \mathbb{V}^{n_1 \times n_2}$ is a closed convex set. Note that for any $X \in \mathbb{V}^{n_1 \times n_2}$, the constraint $\operatorname{rank}(X) \le r$ is equivalent to

$$0 = \sigma_{r+1}(X) + \cdots + \sigma_n(X) = \|X\|_* - \|X\|_{(r)},$$

where $\|X\|_{(r)} := \sigma_1(X) + \cdots + \sigma_r(X)$ denotes the Ky Fan r-norm. The central idea of the majorized penalty approach is to solve the following penalized version of (3.10):

$$\begin{aligned} \min \;\; & h(X) + \rho\big(\|X\|_* - \|X\|_{(r)}\big) \\ \text{s.t.} \;\; & X \in K, \end{aligned}$$

where $\rho > 0$ is the penalty parameter. With the current iterate X^k, the majorized penalty approach yields the next iterate X^{k+1} by solving the convex optimization problem

$$\begin{aligned} \min \;\; & \widehat{h}^k(X) + \rho\Big(\|X\|_* - \langle G^k, X\rangle + \frac{\gamma_k}{2}\|X - X^k\|_F^2\Big) \\ \text{s.t.} \;\; & X \in K, \end{aligned} \tag{3.11}$$

where $\gamma_k \ge 0$, G^k is a subgradient of the convex function $\|X\|_{(r)}$ at X^k, and $\widehat{h}^k$ is a convex majorization function of h at X^k. Comparing with (3.8), one may notice that our proposed rank-correction step is close to one step of the majorized penalty approach.

Due to the structured randomness of matrix completion, we expect that the estimator generated from the rank-correction step possesses some favorable properties for recovery. The key issue is how to construct the rank-correction function

F to make such improvements possible. In the next two sections, we provide theoretical supports to our proposed rank-correction step, from which some important guidelines on the construction of F can be captured.

Henceforth, we let $\widehat{X}_m$ denote the estimator generated from the rank-correction step (3.8) or (3.9) for the corresponding cases and let $r = \operatorname{rank}(\overline{X}) \geq 1$. Throughout this chapter, for any $X \in \mathbb{V}^{n_1 \times n_2}$ and any $(U, V) \in \mathbb{O}^{n_1,n_2}(X)$, we write $U = [U_1 \ U_2]$ and $V = [V_1 \ V_2]$ with $U_1 \in \mathbb{O}^{n_1 \times r}$, $U_2 \in \mathbb{O}^{n_1 \times (n_1 - r)}$, $V_1 \in \mathbb{O}^{n_2 \times r}$ and $V_2 \in \mathbb{O}^{n_2 \times (n_2 - r)}$. Meanwhile, for any $X \in \mathbb{S}^n_+$ and any $P \in \mathbb{O}^n(X)$, we write $P = [P_1 \ P_2]$ with $P_1 \in \mathbb{O}^{n \times r}$ and $P_2 \in \mathbb{O}^{n \times (n-r)}$.

3.2 Error bounds

In this section, we aim to derive a recovery error bound in the Frobenius norm for the rank-correction step and discuss the impact of the rank-correction term on the obtained bound.The following analysis focuses on the cases for recovering a rectangular matrix or a symmetric/Hermitian matrix. All the results obtained in this section are applicable to the positive semidefinite case since adding more prior information can only improve recoverability.

We first introduce the orthogonal decomposition $\mathbb{V}^{n_1 \times n_2} = T \oplus T^\perp$ with

$$\begin{cases} T := \big\{ X \in \mathbb{V}^{n_1 \times n_2} \ \big| \ X = X_1 + X_2 \text{ with } \operatorname{col}(X_1) \subseteq \operatorname{col}(\overline{X}), \operatorname{row}(X_2) \subseteq \operatorname{row}(\overline{X}) \big\}, \\ T^\perp := \big\{ X \in \mathbb{V}^{n_1 \times n_2} \ \big| \ \operatorname{row}(X) \perp \operatorname{row}(\overline{X}) \text{ and } \operatorname{col}(X) \perp \operatorname{col}(\overline{X}) \big\}, \end{cases}$$

where $\operatorname{row}(X)$ and $\operatorname{col}(X)$ denote the row space and column space of the matrix X, respectively. Let $\mathcal{P}_T : \mathbb{V}^{n_1 \times n_2} \to \mathbb{V}^{n_1 \times n_2}$ and $\mathcal{P}_{T^\perp} : \mathbb{V}^{n_1 \times n_2} \to \mathbb{V}^{n_1 \times n_2}$ be the orthogonal projection operators onto the subspaces T and $T^\perp$, respectively. It is

not hard to verify that

$$\begin{cases} \mathcal{P}_T(X) = \overline{U}_1\overline{U}_1^{\mathbb{T}}X + X\overline{V}_1\overline{V}_1^{\mathbb{T}} - \overline{U}_1\overline{U}_1^{\mathbb{T}}X\overline{V}_1\overline{V}_1^{\mathbb{T}}, \\ \mathcal{P}_{T^\perp}(X) = \overline{U}_2\overline{U}_2^{\mathbb{T}}X\overline{V}_2\overline{V}_2^{\mathbb{T}} \end{cases} \tag{3.12}$$

for any $X \in \mathbb{V}^{n_1\times n_2}$ and $(\overline{U}, \overline{V}) \in \mathbb{O}^{n_1,n_2}(\overline{X})$. Define a_m and b_m, respectively, by

$$\begin{cases} a_m := \min\left\{\|\overline{U}_1\overline{V}_1^{\mathbb{T}} - \mathcal{P}_T(F(\widetilde{X}_m) + \gamma_m\widetilde{X}_m))\|, \|\overline{U}_1\overline{V}_1^{\mathbb{T}} - (F(\widetilde{X}_m) + \gamma_m\widetilde{X}_m))\|\right\} \\ b_m := 1 - \|\mathcal{P}_{T^\perp}(F(\widetilde{X}_m) + \gamma_m\widetilde{X}_m))\|. \end{cases} \tag{3.13}$$

Note that the first term in the objective function of (3.8) can be rewritten as

$$\frac{1}{2m}\|y - \mathcal{R}_\Omega(X)\|_2^2 = \frac{1}{2m}\left\|\mathcal{R}_\Omega(X - \overline{X})\right\|_2^2 - \frac{\nu}{m}\left\langle\mathcal{R}_\Omega^*(\xi), X\right\rangle.$$

Using the optimality of $\widehat{X}_m$ to the problem (3.8), we obtain the following result.

Theorem 3.1. *Assume that* $\|\mathcal{P}_{T^\perp}(F(\widetilde{X}_m) + \gamma_m\widetilde{X}_m)\| < 1$. *For any* $\kappa > 1$, *if*

$$\rho_m \geq \frac{\kappa\nu}{b_m}\left\|\frac{1}{m}\mathcal{R}_\Omega^*(\xi)\right\|, \tag{3.14}$$

then the following inequality holds:

$$\frac{1}{2m}\left\|\mathcal{R}_\Omega(\widehat{X}_m - \overline{X})\right\|_2^2 \leq \sqrt{2r}\left(a_m + \frac{b_m}{\kappa}\right)\rho_m\|\widehat{X}_m - \overline{X}\|_F + \frac{\rho_m\gamma_m}{2}\left(\|\overline{X}\|_F^2 - \|\widehat{X}_m\|_F^2\right). \tag{3.15}$$

Proof. Let $\Delta_m := \widehat{X}_m - \overline{X}$. Since $\widehat{X}_m$ is optimal to (3.8) and $\overline{X}$ is feasible to (3.8), it follows that

$$\begin{aligned} \frac{1}{2m}\|\mathcal{R}_\Omega(\Delta_m)\|_2^2 \leq \left\langle\frac{\nu}{m}\mathcal{R}_\Omega^*(\xi), \Delta_m\right\rangle - \rho_m\big(\|\widehat{X}_m\|_* - \|\overline{X}\|_* - \langle F(\widetilde{X}_m) + \gamma_m\widetilde{X}_m, \Delta_m\rangle\big) \\ + \frac{\rho_m\gamma_m}{2}\big(\|\overline{X}\|_F^2 - \|\widehat{X}_m\|_F^2\big). \end{aligned} \tag{3.16}$$

Then, it follows from (3.14) that

$$\begin{aligned} \left\langle\frac{\nu}{m}\mathcal{R}_\Omega^*(\xi), \Delta_m\right\rangle &\leq \nu\left\|\frac{1}{m}\mathcal{R}_\Omega^*(\xi)\right\|\big(\|\mathcal{P}_T(\Delta_m)\|_* + \|\mathcal{P}_{T^\perp}(\Delta_m)\|_*\big) \\ &\leq \frac{\rho_m b_m}{\kappa}\big(\|\mathcal{P}_T(\Delta_m)\|_* + \|\mathcal{P}_{T^\perp}(\Delta_m)\|_*\big). \end{aligned} \tag{3.17}$$

From the directional derivative of the nuclear norm at $\overline{X}$ (see [176, Theorem 1]), we have

$$\|\widehat{X}_m\|_* - \|\overline{X}\|_* \geq \langle \overline{U}_1\overline{V}_1^{\mathbb{T}}, \Delta_m\rangle + \|\overline{U}_2^{\mathbb{T}}\Delta_m\overline{V}_2\|_*.$$

This, together with equations (3.12) and (3.13), implies that

$$\begin{aligned}
&\|\widehat{X}_m\|_* - \|\overline{X}\|_* - \langle F(\widetilde{X}_m)+\gamma_m\widetilde{X}_m, \Delta_m\rangle \\
&\geq \langle \overline{U}_1\overline{V}_1^{\mathbb{T}}, \Delta_m\rangle + \|\overline{U}_2^{\mathbb{T}}\Delta_m\overline{V}_2\|_* - \langle F(\widetilde{X}_m) + \gamma_m\widetilde{X}_m, \Delta_m\rangle \\
&= \langle \overline{U}_1\overline{V}_1^{\mathbb{T}} - \mathcal{P}_T(F(\widetilde{X}_m)+\gamma_m\widetilde{X}_m), \Delta_m\rangle + \|\mathcal{P}_{T^\perp}(\Delta_m)\|_* - \langle \mathcal{P}_{T^\perp}(F(\widetilde{X}_m)+\gamma_m\widetilde{X}_m), \Delta_m\rangle \\
&= \langle \overline{U}_1\overline{V}_1^{\mathbb{T}} - \mathcal{P}_T(F(\widetilde{X}_m)+\gamma_m\widetilde{X}_m), \mathcal{P}_T(\Delta_m)\rangle + \|\mathcal{P}_{T^\perp}(\Delta_m)\|_* \\
&\qquad\qquad - \langle \mathcal{P}_{T^\perp}(F(\widetilde{X}_m)+\gamma_m\widetilde{X}_m), \mathcal{P}_{T^\perp}(\Delta_m)\rangle \\
&\geq -\min\left\{\|\overline{U}_1\overline{V}_1^{\mathbb{T}} - \mathcal{P}_T(F(\widetilde{X}_m)+\gamma_m\widetilde{X}_m)\|, \|\overline{U}_1\overline{V}_1^{\mathbb{T}} - (F(\widetilde{X}_m)+\gamma_m\widetilde{X}_m)\|\right\}\|\mathcal{P}_T(\Delta_m)\|_* \\
&\qquad\qquad + \left(1 - \|\mathcal{P}_{T^\perp}(F(\widetilde{X}_m+\gamma_m\widetilde{X}_m)\|\right)\|\mathcal{P}_{T^\perp}(\Delta_m)\|_* \\
&= -a_m\|\mathcal{P}_T(\Delta_m)\|_* + b_m\|\mathcal{P}_{T^\perp}(\Delta_m)\|_*. \qquad (3.18)
\end{aligned}$$

By substituting (3.18) and (3.17) into (3.16), we obtain that

$$\begin{aligned}
\frac{1}{2m}\|\mathcal{R}_\Omega(\Delta_m)\|_2^2 \leq \rho_m\left(\left(a_m + \frac{b_m}{\kappa}\right)\|\mathcal{P}_T(\Delta_m)\|_* - \left(b_m - \frac{b_m}{\kappa}\right)\|\mathcal{P}_{T^\perp}(\Delta_m)\|_*\right) \\
+ \frac{\rho_m\gamma_m}{2}(\|\overline{X}\|_F^2 - \|\widehat{X}_m\|_F^2). \qquad (3.19)
\end{aligned}$$

Note that $\mathrm{rank}(\mathcal{P}_T(\Delta_m)) \leq 2r$. Hence,

$$\|\mathcal{P}_T(\Delta_m)\|_* \leq \sqrt{2r}\|\mathcal{P}_T(\Delta_m)\|_F \leq \sqrt{2r}\|\Delta_m\|_F,$$

and the desired result follows from (3.19). Thus, we complete the proof. □

Theorem 3.1 shows that, to derive an error bound on $\|\widehat{X}_m - \overline{X}\|_F$, we only need to establish the relation between $\|\widehat{X}_m - \overline{X}\|_F^2$ and $\frac{1}{m}\|\mathcal{R}_\Omega(\widehat{X}_m - \overline{X})\|_2^2$. It is well-known that the sampling operator $\mathcal{R}_\Omega$ does not satisfy the RIP, but it has a similar property with high probability under certain conditions (see, e.g.,

[135, 91, 86, 110]). For deriving such a property, here, we impose a bound restriction on the true matrix $\overline{X}$ in the form of $\|\mathcal{R}_\beta(\overline{X})\|_\infty \leq h$. This condition is very mild since a bound is often known in some applications such as in the correlation and density matrix completion. Correspondingly, we add the bound constraint $\|\mathcal{R}_\beta(X)\|_\infty \leq h$ to the problem (3.8) in the rank-correction step. Since the feasible set is bounded in this case, we simply set $\gamma_m = 0$ and let $\widehat{X}_m^h$ denote the estimator generated from the rank-correction step in this case.

The above boundedness setting is similar to the one adopted by Klopp [86] for the nuclear norm penalized least squares estimator. A slight difference is that the upper bound is imposed on the basis coefficients of $\overline{X}$ relative to $\{\Theta_k : k \in \beta\}$ rather than all the entries of $\overline{X}$. It is easy to see that if the bound is not too tight, the estimator $\widehat{X}_m^h$ is the same as $\widehat{X}_m$. Therefore, we next derive the recovery error bound of $\widehat{X}_m^h$ instead of $\widehat{X}_m$, by following Klopp's arguments in [86], which are also in line with the work done by Negahban and Wainwright [135].

Let μ_1 be a constant to control the smallest sampling probability for observations as

$$p_k \geq (\mu_1 d_2)^{-1} \quad \forall\, k \in \beta. \tag{3.20}$$

It follows from Assumption 3.1 that $\mu_1 \geq 1$ and in particular $\mu_1 = 1$ for the uniform sampling. Note that the magnitude of μ_1 does not depend on d_2 or the matrix size. By the definition of $\mathcal{Q}_\beta$, we then have

$$\langle \mathcal{Q}_\beta(\Delta), \Delta \rangle \geq (\mu_1 d_2)^{-1} \|\Delta\|_F^2 \quad \forall\, \Delta \in \{\Delta \in \mathbb{V}^{n_1 \times n_2} \mid \mathcal{R}_\alpha(\Delta) = 0\}. \tag{3.21}$$

Let $\{\epsilon_1, \ldots, \epsilon_m\}$ be an i.i.d. Rademacher sequence, i.e., an i.i.d. sequence of Bernoulli random variables taking the values 1 and -1 with probability $1/2$. Define

$$\vartheta_m := \mathbb{E}\left\| \frac{1}{m} \mathcal{R}_\Omega^*(\epsilon) \right\| \quad \text{with} \quad \epsilon = (\epsilon_1, \ldots, \epsilon_m)^{\mathbb{T}}. \tag{3.22}$$

Then, we can obtain a similar result to [86, Lemma 12] by showing that the sampling operator $\mathcal{R}_\Omega$ satisfies some approximate RIP for the matrices in some specified

sets.

Lemma 3.2. *Given any $s > 0$ and $t > 0$, define*

$$K(s,t) := \left\{ \Delta \in \mathbb{V}^{n_1 \times n_2} \;\middle|\; \mathcal{R}_\alpha(\Delta) = 0, \|\mathcal{R}_\beta(\Delta)\|_\infty = 1,\ \|\Delta\|_* \le s\|\Delta\|_F,\ \langle \mathcal{Q}_\beta(\Delta), \Delta \rangle \ge t \right\}.$$

Then, for any θ, τ_1 and τ_2 satisfying

$$\theta > 1, \quad 0 < \tau_1 < 1 \quad \text{and} \quad 0 < \tau_2 < \frac{\tau_1}{\theta}, \tag{3.23}$$

with probability at least $1 - \frac{\exp(-(\tau_1 - \theta\tau_2)^2 m t^2/2)}{1 - \exp(-(\theta^2-1)(\tau_1-\theta\tau_2)^2 m t^2/2)}$,

$$\frac{1}{m}\|\mathcal{R}_\Omega(\Delta)\|_2^2 \ge (1-\tau_1)\langle \mathcal{Q}_\beta(\Delta), \Delta \rangle - \frac{16}{\tau_2} s^2 \mu_1 d_2 \vartheta_m^2 \quad \forall\, \Delta \in K(s,t). \tag{3.24}$$

In particular, given any constant $c > 0$, with probability at least $1 - \frac{(n_1+n_2)^{-c}}{1 - 2^{-(\theta^2-1)c}}$, the inequality (3.24) holds with $t = \sqrt{\frac{2c\log(n_1+n_2)}{(\tau_1-\theta\tau_2)^2 m}}$.

Proof. The proof is similar to that of [86, Lemma 12]. For any $s, t > 0$ and τ_1, τ_2, θ satisfying (3.23), we need to show that the event

$$E = \left\{ \exists \Delta \in K(s,t) \text{ such that } \left| \frac{1}{m}\|\mathcal{R}_\Omega(\Delta)\|_2^2 - \langle \mathcal{Q}_\beta(\Delta), \Delta \rangle \right| \ge \tau_1 \langle \mathcal{Q}_\beta(\Delta), \Delta \rangle + \frac{16}{\tau_2} s^2 \mu_1 d_2 \vartheta_m^2 \right\}$$

occurs with probability less than $\frac{\exp(-(\tau_1-\theta\tau_2)^2 m t^2/2)}{1-\exp(-(\theta^2-1)(\tau_1-\theta\tau_2)^2 m t^2/2)}$. We first decompose $K(s,t)$ as

$$K(s,t) = \bigcup_{k=1}^{\infty} \left\{ \Delta \in K(s,t) \;\middle|\; \theta^{k-1} t \le \langle \mathcal{Q}_\beta(\Delta), \Delta \rangle \le \theta^k t \right\}.$$

For any $a \ge t$, we further define

$$K(s,t,a) := \{ \Delta \in K(s,t) \mid \langle \mathcal{Q}_\beta(\Delta), \Delta \rangle \le a \}.$$

Then we get $E \subseteq \bigcup_{k=1}^{\infty} E_k$ with

$$E_k = \left\{ \exists \Delta \in K(s,t,\theta^k t) \text{ such that } \left| \frac{1}{m}\|\mathcal{R}_\Omega(\Delta)\|_2^2 - \langle \mathcal{Q}_\beta(\Delta), \Delta \rangle \right| \ge \theta^{k-1}\tau_1 t + \frac{16}{\tau_2} s^2 \mu_1 d_2 \vartheta_m^2 \right\}.$$

Then, we need to estimate the probability of each event E_k. Define

$$Z_a := \sup_{\Delta \in K(s,t,a)} \left| \frac{1}{m} \|\mathcal{R}_\Omega(\Delta)\|_2^2 - \langle \mathcal{Q}_\beta(\Delta), \Delta \rangle \right|.$$

Notice that for any $\Delta \in \mathbb{V}^{n_1 \times n_2}$,

$$\frac{1}{m} \|\mathcal{R}_\Omega(\Delta)\|_2^2 = \frac{1}{m} \sum_{i=1}^{m} \langle \Theta_{\omega_i}, \Delta \rangle^2 \overset{a.s.}{\to} \mathbb{E}(\langle \Theta_{\omega_i}, \Delta \rangle^2) = \langle \mathcal{Q}_\beta(\Delta), \Delta \rangle.$$

Since $\|\mathcal{R}_\beta(\Delta)\|_\infty \leq 1$ for all $\Delta \in K(s,t)$, from Massart's Hoeffding type concentration inequality [120, Theorem 9] for suprema of empirical processes, we have

$$\Pr\big(Z_a \geq \mathbb{E}(Z_a) + \varepsilon\big) \leq \exp\left(-\frac{m\varepsilon^2}{2}\right) \quad \forall\, \varepsilon > 0. \tag{3.25}$$

Next, we use the standard Rademacher symmetrization in the theory of empirical processes to further derive an upper bound of $\mathbb{E}(Z_a)$. Let $\{\epsilon_1, \ldots, \epsilon_m\}$ be a Rademacher sequence. Then, we have

$$\begin{aligned}
\mathbb{E}(Z_a) &= \mathbb{E}\left(\sup_{\Delta \in K(s,t,a)} \left| \frac{1}{m} \sum_{i=1}^{m} \langle \Theta_{\omega_i}, \Delta \rangle^2 - \mathbb{E}\big(\langle \Theta_{\omega_i}, \Delta \rangle^2\big) \right| \right) \\
&\leq 2\mathbb{E}\left(\sup_{\Delta \in K(s,t,a)} \left| \frac{1}{m} \sum_{i=1}^{m} \epsilon_i \langle \Theta_{\omega_i}, \Delta \rangle^2 \right| \right) \\
&\leq 8\mathbb{E}\left(\sup_{\Delta \in K(s,t,a)} \left| \frac{1}{m} \sum_{i=1}^{m} \epsilon_i \langle \Theta_{\omega_i}, \Delta \rangle \right| \right) \\
&= 8\mathbb{E}\left(\sup_{\Delta \in K(s,t,a)} \left| \frac{1}{m} \sum_{i=1}^{m} \langle \mathcal{R}_\Omega^*(\epsilon), \Delta \rangle \right| \right) \\
&\leq 8\mathbb{E}\left\| \frac{1}{m} \mathcal{R}_\Omega^*(\epsilon) \right\| \left(\sup_{\Delta \in K(s,t,a)} \|\Delta\|_* \right),
\end{aligned} \tag{3.26}$$

where the first inequality follows from the symmetrization theorem (e.g., see [173, Lemma 2.3.1] and [16, Theorem 14.3]) and the second inequality follows from the contraction theorem (e.g., see [101, Theorem 4.12] and [16, Theorem 14.4]). Moreover, from (3.21), we have

$$\|\Delta\|_* \leq s\|\Delta\|_F \leq s\sqrt{\mu_1 d_2 \langle \mathcal{Q}_\beta(\Delta), \Delta \rangle} \leq s\sqrt{\mu_1 d_2 a} \quad \forall\, \Delta \in K(s,t,a). \tag{3.27}$$

Combining (3.26) and (3.27) with the definition of ϑ_m in (3.22), we obtain that

$$\begin{aligned}\mathbb{E}(Z_a) + \Big(\frac{\tau_1}{\theta} - \tau_2\Big)a &\le 8s\vartheta_m\sqrt{\mu_1 d_2 a} + \Big(\frac{\tau_1}{\theta} - \tau_2\Big)a \\ &= s\vartheta_m\sqrt{\frac{32\mu_1 d_2}{\tau_2}} \cdot \sqrt{2a\tau_2} + \Big(\frac{\tau_1}{\theta} - \tau_2\Big)a \\ &\le \frac{16}{\tau_2}s^2\mu_1 d_2\vartheta_m^2 + \frac{\tau_1}{\theta}a,\end{aligned}$$

where the second inequality follows from the simple fact $x_1x_2 \le (x_1^2 + x_2^2)/2$ for any $x_1, x_2 \ge 0$. Then, it follows from (3.25) that

$$\Pr\Big(Z_a \ge \frac{\tau_1}{\theta}a + \frac{16}{\tau_2}s^2\mu_1 d_2\vartheta_m^2\Big) \le \Pr\Big(Z_a \ge \mathbb{E}(Z_a) + \Big(\frac{\tau_1}{\theta} - \tau_2\Big)a\Big) \le \exp\left(-\Big(\frac{\tau_1}{\theta} - \tau_2\Big)^2\frac{ma^2}{2}\right).$$

This implies that

$$\Pr(E_k) \le \exp\Big(-\frac{1}{2}\theta^{2(k-1)}(\tau_1 - \theta\tau_2)^2 mt^2\Big).$$

Then, since $\theta > 1$, by using $\theta^k \ge 1 + k(\theta - 1)$ for any $k \ge 1$, we have

$$\begin{aligned}\Pr(E) &\le \sum_{k=1}^{\infty}\Pr(E_k) \le \sum_{k=1}^{\infty}\exp\Big(-\frac{1}{2}\theta^{2(k-1)}(\tau_1 - \theta\tau_2)^2 mt^2\Big) \\ &\le \exp\Big(-\frac{1}{2}(\tau_1 - \theta\tau_2)^2 mt^2\Big)\sum_{k=1}^{\infty}\exp\Big(-\frac{1}{2}(\theta^{2(k-1)} - 1)(\tau_1 - \theta\tau_2)^2 mt^2\Big) \\ &\le \exp\Big(-\frac{1}{2}(\tau_1 - \theta\tau_2)^2 mt^2\Big)\sum_{k=1}^{\infty}\exp\Big(-\frac{1}{2}(k-1)(\theta^2 - 1)(\tau_1 - \theta\tau_2)^2 mt^2\Big) \\ &\le \frac{\exp\big(-(\tau_1 - \theta\tau_2)^2 mt^2/2\big)}{1 - \exp\big(-(\theta^2 - 1)(\tau_1 - \theta\tau_2)^2 mt^2/2\big)}.\end{aligned}$$

In particular, for any constant $c > 0$, let $t = \sqrt{\frac{2c\log(n_1+n_2)}{(\tau_1 - \theta\tau_2)^2 m}}$. Then, direct calculation yields

$$\frac{\exp\big(-(\tau_1 - \theta\tau_2)^2 mt^2/2\big)}{1 - \exp\big(-(\theta^2 - 1)(\tau_1 - \theta\tau_2)^2 mt^2/2\big)} = \frac{(n_1 + n_2)^{-c}}{1 - (n_1 + n_2)^{-(\theta^2 - 1)c}} \le \frac{(n_1 + n_2)^{-c}}{1 - 2^{-(\theta^2 - 1)c}}.$$

Thus, we complete the proof. □

Now, combining Theorem 3.1 and Lemma 3.2, we obtain the following result.

Theorem 3.3. *Assume that* $\|\mathcal{P}_{T^\perp}(F(\widetilde{X}_m))\| < 1$ *and* $\|\mathcal{R}_\beta(\overline{X})\|_\infty \le h$ *for some* h. *Then, there exist some positive absolute constants* c_0, c_1, c_2 *and* C_0 *such that for any* $\kappa > 1$, *if* ρ_m *is chosen as (3.14), then with probability at least* $1 - c_1(n_1+n_2)^{-c_2}$,

$$\frac{\|\widehat{X}_m^h - \overline{X}\|_F^2}{d_2} \le C_0 \max\left\{\mu_1^2 d_2 r\left(c_0^2\Big(a_m + \frac{b_m}{\kappa}\Big)^2\rho_m^2 + \Big(\frac{\kappa}{\kappa-1}\Big)\Big(1+\frac{a_m}{b_m}\Big)^2\vartheta_m^2 h^2\right), \right.$$
$$\left. h^2\mu_1\sqrt{\frac{\log(n_1+n_2)}{m}}\right\}. \tag{3.28}$$

Proof. The proof is similar to that of [86, Theorem 3]. Let $\Delta_m^h := \widehat{X}_m^h - \overline{X}$. By noting that $\gamma_m = 0$ in this case, from (3.19), we have

$$\Big(a_m + \frac{b_m}{\kappa}\Big)\|\mathcal{P}_T(\Delta_m^h)\|_* - \Big(b_m - \frac{b_m}{\kappa}\Big)\|\mathcal{P}_{T^\perp}(\Delta_m^h)\|_* \ge 0.$$

Then, by setting $w_m := \frac{\kappa}{\kappa-1}\big(1+\frac{a_m}{b_m}\big)$, together with the above inequality, we obtain that

$$\|\Delta_m^h\|_* \le \|\mathcal{P}_T(\Delta_m^h)\|_* + \|\mathcal{P}_{T^\perp}(\Delta_m^h)\|_* \le w_m\|\mathcal{P}_T(\Delta_m^h)\|_* \le \sqrt{2r}w_m\|\Delta_m^h\|_F. \tag{3.29}$$

Let $h_m := \|\mathcal{R}_\beta(\Delta_m^h)\|_\infty$. Clearly, $h_m \le 2h$. For any fixed constants $c > 0$ and any fixed θ, τ_1, τ_2 satisfying (3.23), we proceed the discussions by two cases:

Case 1. Suppose that $\langle \mathcal{Q}_\beta(\Delta_m^h), \Delta_m^h\rangle \le h_m^2\sqrt{\frac{2c\log(n_1+n_2)}{(\tau_1-\theta\tau_2)^2 m}}$. From (3.21), we obtain that

$$\frac{\|\Delta_m^h\|_F^2}{d_2} \le 4h^2\mu_1\sqrt{\frac{2c\log(n_1+n_2)}{(\tau_1-\theta\tau_2)^2 m}}. \tag{3.30}$$

Case 2. Suppose that $\langle \mathcal{Q}_\beta(\Delta_m^h), \Delta_m^h\rangle > h_m^2\sqrt{\frac{2c\log(n_1+n_2)}{(\tau_1-\theta\tau_2)^2 m}}$. Then, from (3.29), we have $\Delta_m^h/h_m \in K(s,t)$ with $s = \sqrt{2r}w_m$ and $t = \sqrt{\frac{2c\log(n_1+n_2)}{(\tau_1-\theta\tau_2)^2 m}}$. Together with Lemma 3.2, we obtain that with probability at least $1 - \frac{(n_1+n_2)^{-c}}{1-2^{-(\theta^2-1)c}}$,

$$(1-\tau_1)\langle \mathcal{Q}_\beta(\Delta_m^h), \Delta_m^h\rangle \le \frac{1}{m}\|\mathcal{R}_\Omega(\Delta_m^h)\|_2^2 + \frac{32}{\tau_2}w_m^2\mu_1 d_2 r\vartheta_m^2 h_m^2.$$

Combining the above inequality with Theorem 3.1 and the equation (3.21), we obtain that for any given τ_3 satisfying $0 < \tau_3 < 1$,

$$\begin{aligned}\frac{\|\Delta_m^h\|_F^2}{d_2} &\leq \mu_1 \langle \mathcal{Q}_\beta(\Delta_m^h), \Delta_m^h \rangle \\ &\leq \frac{\mu_1}{1-\tau_1}\Big(\frac{1}{m}\|\mathcal{R}_\Omega(\Delta_m^h)\|_2^2 + \frac{32}{\tau_2} w_m^2 \mu_1 d_2 r \vartheta_m^2 h_m^2\Big) \\ &\leq \frac{2\sqrt{2r}}{1-\tau_1}\Big(a_m + \frac{b_m}{\kappa}\Big)\mu_1\rho_m\|\Delta_m^h\|_F + \frac{32}{(1-\tau_1)\tau_2} w_m^2\mu_1^2 d_2 r\vartheta_m^2 h_m^2 \\ &\leq \tau_3 \frac{\|\Delta_m^h\|_F^2}{d_2} + \frac{2}{(1-\tau_1)^2\tau_3}\Big(a_m + \frac{b_m}{\kappa}\Big)^2 \mu_1^2\rho_m^2 r d_2 + \frac{32}{(1-\tau_1)\tau_2} w_m^2\mu_1^2 d_2 r\vartheta_m^2 h_m^2.\end{aligned}$$

By plugging in w_m, we have that

$$\frac{\|\Delta_m^h\|_F^2}{d_2} \leq \frac{\mu_1^2 d_2 r}{1-\tau_3}\left(\frac{2}{(1-\tau_1)^2\tau_3}\Big(a_m + \frac{b_m}{\kappa}\Big)^2\rho_m^2 + \frac{128}{(1-\tau_1)\tau_2}\Big(\frac{\kappa}{\kappa-1}\Big)\Big(1+\frac{a_m}{b_m}\Big)^2\vartheta_m^2 h^2\right). \tag{3.31}$$

Finally, by choosing τ_1, τ_2, τ_3 and θ to be absolute constants in (3.30) and (3.31), we complete the proof. □

We remind the readers that in the proof of Theorem 3.3, when the probability is fixed, (i.e., the constants c and θ are fixed), a tighter error bound can be achieved by choosing τ_1, τ_2 and τ_3 beyond absolute constants. More precisely, the error bound (3.31) can be minimized over $0 < \tau_3 < 1$, and after that, the joint error bound of (3.30) and (3.31) can be further minimized over $0 < \tau_1 < 1$ and $0 < \tau_2 < \theta/\tau_1$. Nevertheless, for simplicity of our discussions, we stay with the the error bound (3.28) in Theorem 3.3 in the sequel.

In order to choose a parameter ρ_m such that (3.14) holds, we need to estimate $\|\frac{1}{m}\mathcal{R}_\Omega^*(\xi)\|$. For this purpose, we make the following assumption on the noises.

Assumption 3.2. *The i.i.d. noise variables ξ_i are sub-exponential, i.e., there exist positive constants c_1, c_2 such that for all $t > 0$,*

$$\Pr\big(|\xi_i| \geq t\big) \leq c_1 \exp(-c_2 t).$$

The noncommutative Bernstein inequality is a useful tool for the study of matrix completion problems. It provides bounds of the probability that the sum of random matrices deviates from its mean in the operator norm (see, e.g., [149, 169, 70]). Recently, the noncommutative Bernstein inequality was extended by replacing bounds of the operator norm of matrices with bounds of the Orlicz norms (see [90, 91]). Given any $s \geq 1$, the ψ_s Orlicz norm of a random variable z is defined by

$$\|z\|_{\psi_s} := \inf \left\{ t > 0 \;\middle|\; \mathbb{E} \exp\left(\frac{|z|^s}{t^s} \right) \leq 2 \right\}.$$

The Orlicz norms are useful to characterize the tail behavior of random variables. The following noncommutative Bernstein inequality is taken from [88, Corollary 2.1].

Theorem 3.4 (Koltchinskii [88]). *Let $Z_1, \ldots, Z_m \in \mathbb{V}^{n_1 \times n_2}$ be independent random matrices with mean zero. Suppose that $\max\left\{ \left\| \|Z_i\| \right\|_{\psi_s}, 2\mathbb{E}^{\frac{1}{2}}(\|Z_i\|^2) \right\} < \varpi_s$ for some constant ϖ_s. Define*

$$\sigma_Z := \max\left\{ \left\| \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}(Z_i Z_i^{\mathbb{T}}) \right\|^{1/2}, \; \left\| \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}(Z_i^{\mathbb{T}} Z_i) \right\|^{1/2} \right\}.$$

Then, there exists a constant C such that for all $t > 0$, with probability at least $1 - \exp(-t)$,

$$\left\| \frac{1}{m} \sum_{i=1}^{m} Z_i \right\| \leq C \max\left\{ \sigma_Z \sqrt{\frac{t + \log(n_1 + n_2)}{m}}, \varpi_s \left(\log \frac{\varpi_s}{\sigma_Z} \right)^{1/s} \frac{t + \log(n_1 + n_2)}{m} \right\}.$$

It is known that a random variable is sub-exponential if and only its ψ_1 Orlicz norm is finite [173]. To apply the noncommutative Bernstein inequality, we let μ_2 be a constant such that

$$\max\left\{ \left\| \sum_{k \in \beta} p_k \Theta_k \Theta_k^{\mathbb{T}} \right\|, \; \left\| \sum_{k \in \beta} p_k \Theta_k^{\mathbb{T}} \Theta_k \right\| \right\} \leq \frac{\mu_2}{n}. \tag{3.32}$$

Notice that

$$\mathrm{Tr}\left(\sum_{k \in \beta} p_k \Theta_k \Theta_k^{\mathbb{T}} \right) = \mathrm{Tr}\left(\sum_{k \in \beta} p_k \Theta_k^{\mathbb{T}} \Theta_k \right) = 1.$$

Thus, the lower bound of the term on the left-hand side is $1/n$. This implies that $\mu_2 \geq 1$. In the following, we also assume that the magnitude of μ_2 does not depend on the matrix size. For example, $\mu_2 = 1$ for the correlation matrix completion under uniform sampling and the density matrix completion described in Section 3.1. The following result extends [91, Lemma 2] and [86, Lemmas 5 & 6] from the standard basis to an arbitrary orthonormal basis. A similar result can also be found in [135, Lemma 6].

Lemma 3.5. *Under Assumption 3.2, there exists a positive constant C_* (only depending on the ψ_1 Orlicz norm of ξ_k) such that for all $t > 0$, with probability at least $1 - \exp(-t)$,*

$$\left\| \frac{1}{m} \mathcal{R}^*_\Omega(\xi) \right\| \leq C_* \max \left\{ \sqrt{\frac{\mu_2(t + \log(n_1 + n_2))}{mn}}, \frac{\log(n)(t + \log(n_1 + n_2))}{m} \right\}. \quad (3.33)$$

In particular, when $m \geq n \log^3(n_1 + n_2)/\mu_2$, we also have

$$\mathbb{E} \left\| \frac{1}{m} \mathcal{R}^*_\Omega(\xi) \right\| \leq C_* \sqrt{\frac{2e\mu_2 \log(n_1 + n_2)}{mn}}, \quad (3.34)$$

where e is the exponential constant.

Proof. Recall that

$$\frac{1}{m} \mathcal{R}^*_\Omega(\xi) = \frac{1}{m} \sum_{i=1}^{m} \xi_i \Theta_{\omega_i}.$$

Let $Z_i := \xi_i \Theta_{\omega_i}$. Since $\mathbb{E}(\xi_i) = 0$, the independence of ξ_i and Θ_{ω_i} implies that $\mathbb{E}(Z_i) = 0$. Since $\|\Theta_{\omega_i}\|_F = 1$, we have that

$$\|Z_i\| \leq \|Z_i\|_F = |\xi_i| \|\Theta_{\omega_i}\|_F = |\xi_i|.$$

It follows that $\big\|\|Z_i\|\big\|_{\psi_1} \leq \|\xi_i\|_{\psi_1}$. Thus, $\big\|\|Z_i\|\big\|_{\psi_1}$ is finite since ξ_i is sub-exponential. Meanwhile,

$$\mathbb{E}^{\frac{1}{2}}(\|Z_i\|^2) \leq \mathbb{E}^{\frac{1}{2}}(\|Z_i\|_F^2) = \mathbb{E}^{\frac{1}{2}}(\xi_i^2) = 1.$$

We also have

$$\mathbb{E}\big(Z_i Z_i^{\mathbb{T}}\big) = \mathbb{E}\big(\xi_i^2 \Theta_{\omega_i} \Theta_{\omega_i}^{\mathbb{T}}\big) = \mathbb{E}\big(\Theta_{\omega_i} \Theta_{\omega_i}^{\mathbb{T}}\big) = \sum_{k \in \beta} p_k \Theta_k \Theta_k^{\mathbb{T}}.$$

The calculation of $\mathbb{E}\big(Z_i^{\mathbb{T}} Z_i\big)$ is similar. From (3.32), we obtain that $\sqrt{1/n} \le \sigma_Z \le \sqrt{\mu_2/n}$. Then, applying the noncommutative Bernstein inequality yields (3.33). The proof of (3.34) is exactly the same as the proof of Lemma 6 in [86]. For simplicity, we omit the proof. □

Since Bernoulli random variables are sub-exponential, the right-hand side of (3.34) provides an upper bound of ϑ_m defined by (3.22). Now, we choose $t = c_2 \log(n_1 + n_2)$ in Lemma 3.5 for achieving an optimal order bound with probability at least $1 - (n_1 + n_2)^{-c_2}$, where c_2 is the same as that in Theorem 3.3. With this choice, when $m \ge (1 + c_2) n \log^2(n) \log(n_1 + n_2)/\mu_2$, the first term in the maximum of (3.33) dominates the second term. Hence, for any given $\kappa > 1$, by choosing

$$\rho_m = \frac{\kappa \nu}{b_m} C_* \sqrt{\frac{(1 + c_2)\mu_2 \log(n_1 + n_2)}{mn}}, \tag{3.35}$$

from Theorem 3.3 and Lemma 3.5, we obtain the following main result for recovery error bound.

Theorem 3.6. *Assume that* $\|\mathcal{P}_{T^\perp}(F(\widetilde{X}_m))\| < 1$, $\|\mathcal{R}_\beta(\overline{X})\|_\infty \le h$ *for some* h, *and Assumption 3.2 holds. Then, there exist some positive absolute constants* c_0', c_1', c_2', c_3' *and some positive constants* C_0', C_1' *(only depending on the* ψ_1 *Orlicz norm of* ξ_k*) such that when* $m \ge c_3' n \log^3(n_1 + n_2)/\mu_2$, *for any* $\kappa > 1$, *if* ρ_m *is chosen as*

$$\rho_m = \frac{\kappa \nu}{b_m} C_1' \sqrt{\frac{\mu_2 \log(n_1 + n_2)}{mn}},$$

then with probability at least $1 - c_1'(n_1 + n_2)^{-c_2'}$,

$$\frac{\|\widehat{X}_m^h - \overline{X}\|_F^2}{d_2} \le C_0' \max\left\{\left[{c_0'}^2 \Big(1 + \kappa \frac{a_m}{b_m}\Big)^2 \nu^2 + \Big(\frac{\kappa}{\kappa - 1}\Big)^2 \Big(1 + \frac{a_m}{b_m}\Big)^2 h^2\right] \frac{\mu_1^2 \mu_2 d_2 r \log(n_1 + n_2)}{mn}, \right.$$
$$\left. h^2 \mu_1 \sqrt{\frac{\log(n_1 + n_2)}{m}} \right\}. \tag{3.36}$$

Proof. It is easy to see that combining Theorem 3.3 and Lemma 3.5 yields (3.36) with $C_0' = C_0 \max\{2eC_*^2, 1\}$, $C_1' = C_*\sqrt{1+c_2}$, $c_0' = c_0\sqrt{\frac{1+c_2}{2e}}$, $c_1' = 1 + c_1$, $c_2' = c_2$, $c_3' = 1 + c_2$, where C_0, c_0, c_1, c_2 are the same as that in Theorem 3.3 and C_* is the same as that in Lemma 3.5. □

When the matrix size is large, the second term in the maximum of (3.36) is negligible compared with the first term. Thus, Theorem 3.6 indicates that for any rank-correction function such that $\|\mathcal{P}_{T^\perp}(F(\widetilde{X}_m))\| < 1$, one needs only samples with size of order $d_2 r \log(n_1 + n_2)/n$ to control the recovery error. Note that d_2 is of order $n_1 n_2$ in general. Hence, the order of sample size needed is roughly the degree of freedom of a rank r matrix up to a logarithmic factor in the matrix size. In addition, it is very interesting to notice that the value of κ (or the value of ρ_m) has a substantial influence on the recovery error bound. The first term in the maximum of (3.36) is a sum of two parts related to the magnitude of noise ν and the upper bound of entries h, respectively. The part related to ν increases as κ increases provided $a_m/b_m > 0$, while the part related to h slightly decreases to its limit as κ increases.

Theorem 3.6 also reveals the impact of the rank-correction term on recovery error. It is easy to see from (3.36) that with κ chosen to be the same, a smaller value of a_m/b_m brings a smaller error bound and potentially leads to a smaller recovery error for the rank-correction step. Note that the value of a_m/b_m fully depends on the rank-correction function F when an initial estimator $\widetilde{X}_m$ is given. Note that for any given $\varepsilon_1 \geq 0$ and $0 \leq \varepsilon_2 < 1$, we have

$$\frac{a_m}{b_m} \leq \frac{\varepsilon_1}{1-\varepsilon_2} \quad \text{if} \quad \begin{cases} \left\|\mathcal{P}_T\big(F(\widetilde{X}_m)\big) - \overline{U}_1\overline{V}_1^{\mathbb{T}}\right\| \leq \varepsilon_1, \\ \left\|\mathcal{P}_{T^\perp}\big(F(\widetilde{X}_m)\big)\right\| \leq \varepsilon_2. \end{cases}$$

In particular, if $F \equiv 0$, then the estimator of the rank-correction step reduces to the nuclear norm penalized least squares estimator with $a_m/b_m = 1$. Thus, Theorem

3.6 shows that, with a suitable rank-correction function F, the estimator generated from the rank-correction step for recovery is very likely to perform better than the nuclear norm penalized least squares estimator. In addition, this observation also provides us clues on how to construct a good rank-correction function, to be discussed in Section 3.4.

To disclose the power of the rank-correction term in more details, for any value of a_m/b_m, we intend to find the smallest one among all the error bounds (3.36) with $\kappa > 1$. Here, for simplicity of discussions, instead of (3.36), we consider a slightly relaxed version:

$$\frac{\|\widehat{X}_m^h - \overline{X}\|_F^2}{d_2} \le C_0' \max\left\{\eta_m^2 \frac{\mu_1^2\mu_2 d_2 r \log(n_1+n_2)}{mn}, h^2\mu_1\sqrt{\frac{\log(n_1+n_2)}{m}}\right\},$$

with

$$\eta_m := c_0'\Big(1+\kappa\frac{a_m}{b_m}\Big)\nu + \Big(\frac{\kappa}{\kappa-1}\Big)\Big(1+\frac{a_m}{b_m}\Big)h.$$

It is easy to see from the derivative that over η_m attains its minimum $\overline{\eta}_m$ over $\kappa > 1$ at

$$\overline{\kappa} = 1 + \sqrt{\Big(1+\frac{b_m}{a_m}\Big)\frac{h}{c_0'\nu}},$$

with the minimum value

$$\overline{\eta}_m = \Big(1+\frac{a_m}{b_m}\Big)(c_0'\nu + h) + 2\sqrt{\frac{a_m}{b_m}\Big(1+\frac{a_m}{b_m}\Big)c_0'\nu h}.$$

It is interesting to notice that when $a_m/b_m \ll 1$, we have $\overline{\kappa} = O\big(1/\sqrt{a_m/b_m}\big)$, which means that the optimal choice of κ is inversely proportional to $\sqrt{a_m/b_m}$. In other words, for achieving the best possible recovery error, the parameter ρ_m chosen for the rank-correction step with $a_m/b_m < 1$ should be larger than that for the nuclear norm penalized least squares estimator. In addition,

$$\overline{\eta}_m = \begin{cases} \overline{\eta}^1 := 2(c_0'\nu + h) + 2\sqrt{2}\sqrt{c_0'\nu h} & \text{if } \dfrac{a_m}{b_m} = 1, \\ \overline{\eta}^0 := c_0'\nu + h & \text{if } \dfrac{a_m}{b_m} = 0. \end{cases}$$

By direct calculation, we obtain $\overline{\eta}^1/\overline{\eta}^0 \in [2, 2+\sqrt{2}]$, where the upper bound is attained when $c_0'\nu = h$ and the lower bound is approached when $c_0'\nu/h \to 0$ or $c_0'\nu/h \to \infty$. This finding motivates us to wonder whether the recovery error can be reduced by around half in practice. The numerical experiments of supporting this can be found in Section 3.5.

3.3 Rank consistency

In this section we study the asymptotic behavior of the rank of the estimator $\widehat{X}_m$ for both the rectangular case and the positive semidefinite case. Theorem 3.6 shows that under mild conditions, the distribution of $\widehat{X}_m$ becomes more and more concentrated to the true matrix $\overline{X}$. Due to the low-rank structure of $\overline{X}$, we expect that the estimator $\widehat{X}_m$ has the same low-rank property as $\overline{X}$. For this purpose, we consider the rank consistency in the sense of Bach [7] under the setting that the matrix size is fixed.

Definition 3.1. *An estimator X_m of the true matrix $\overline{X}$ is said to be rank consistent if*

$$\lim_{m\to\infty} \Pr\big(\operatorname{rank}(X_m) = \operatorname{rank}(\overline{X})\big) = 1.$$

Throughout this section we make the following assumptions:

Assumption 3.3. *The spectral operator F is continuous at $\overline{X}$.*

Assumption 3.4. *The initial estimator $\widetilde{X}_m$ satisfies $\widetilde{X}_m \xrightarrow{p} \overline{X}$ as $m \to \infty$.*

In addition, we also need the following properties of the operator $\mathcal{R}_\Omega$ and its adjoint $\mathcal{R}_\Omega^*$.

Lemma 3.7. **(i)** *For any $X \in \mathbb{V}^{n_1\times n_2}$, the random matrix $\frac{1}{m}\mathcal{R}_\Omega^*\mathcal{R}_\Omega(X) \xrightarrow{a.s.} \mathcal{Q}_\beta(X)$.*
(ii) *The random vector $\frac{1}{\sqrt{m}}\mathcal{R}_{\alpha\cup\beta}\mathcal{R}_\Omega^*(\xi) \xrightarrow{d} N\big(0, \operatorname{Diag}(p)\big)$, where $p = (p_1, \ldots, p_d)^{\mathbb{T}}$.*

Proof. (i) From the definition of the sampling operator $\mathcal{R}_\Omega$ and its adjoint $\mathcal{R}_\Omega^*$, we have

$$\frac{1}{m}\mathcal{R}_\Omega^*\mathcal{R}_\Omega(X) = \frac{1}{m}\sum_{i=1}^{m}\langle\Theta_{\omega_i}, X\rangle\,\Theta_{\omega_i}.$$

This is an average value of m i.i.d. random matrices $\langle\Theta_{\omega_i}, X\rangle\Theta_{\omega_i}$. It is easy to see that $\mathbb{E}\big(\langle\Theta_{\omega_i}, X\rangle\Theta_{\omega_i}\big) = \mathcal{Q}_\beta(X)$. The result then follows directly from the strong law of large numbers.

(ii) From the definition of $\mathcal{R}_\Omega^*$ and $\mathcal{R}_{\alpha\cup\beta}$, it is immediate to obtain that

$$\frac{1}{\sqrt{m}}\mathcal{R}_{\alpha\cup\beta}\mathcal{R}_\Omega^*(\xi) = \frac{1}{\sqrt{m}}\mathcal{R}_{\alpha\cup\beta}\left(\sum_{i=1}^{m}\xi_i\Theta_{\omega_i}\right) = \frac{1}{\sqrt{m}}\sum_{i=1}^{m}\xi_i\mathcal{R}_{\alpha\cup\beta}(\Theta_{\omega_i}).$$

Since $\mathbb{E}(\xi_i) = 0$ and $\mathbb{E}(\xi_i^2) = 1$, from the independence of ξ_i and $\mathcal{R}_{\alpha\cup\beta}(\Theta_{\omega_i})$, we have $\mathbb{E}\big(\xi_i\mathcal{R}_{\alpha\cup\beta}(\Theta_{\omega_i})\big) = 0$ and $\mathrm{cov}\big(\xi_i, \mathcal{R}_{\alpha\cup\beta}(\Theta_{\omega_i})\big) = p_i$. Applying the central limit theorem then yields the desired result. □

Based on the above epi-convergence results in Section 2.5, we can analyze the asymptotic behavior of optimal solutions of a sequence of constrained optimization problems. The following result is a direct consequence of the above epi-convergence theorems and Lemma 3.7.

Theorem 3.8. *If $\rho_m \to 0$ and $\gamma_m = O_p(1)$, then $\widehat{X}_m \xrightarrow{p} \overline{X}$ as $m \to \infty$.*

Proof. Let Φ_m denote the objective function of (3.8) and K denote the feasible set. Then, the problem (3.8) can be concisely written as

$$\min_{X\in\mathbb{V}^{n_1\times n_2}} \quad \Phi_m(X) + \delta_K(X).$$

By Assumptions 3.3 and 3.4 and Lemma 3.7, we have that the convex functions Φ_m converges pointwise in probability to the convex function Φ, where $\Phi(X) := \frac{1}{2}\|\mathcal{Q}_\beta(X-\overline{X})\|_2^2$ for any $X \in \mathbb{V}^{n_1\times n_2}$. Then from Theorems 2.16 and 2.17, we obtain that

$$\Phi_m + \delta_K \xrightarrow{e-d} \Phi + \delta_K.$$

Note that $\overline{X}$ is the unique minimizer of $\Phi(X) + \delta_K(X)$ since $\Phi(X)$ is strongly convex over the feasible set K. Thus, we complete the proof by applying Theorem 2.15 on epi-convergence in distribution. □

Theorem 3.8 actually implies that $\widehat{X}_m$ has a higher rank than $\overline{X}$ with probability converging to 1 if $\rho_m \to 0$ and $\gamma_m = O_p(1)$, due to the following straightforward result.

Lemma 3.9. *If* $X_m \xrightarrow{p} \overline{X}$, *then* $\lim_{m\to\infty} \Pr\big(\mathrm{rank}(X_m) \geq \mathrm{rank}(\overline{X})\big) = 1$.

Proof. It follows the Lipschitz continuity of singular values that

$$\sigma_k(X_m) \xrightarrow{p} \sigma_k(X) \quad \forall\, 1 \leq k \leq n.$$

Thus, for a fixed ε satisfying $0 < \varepsilon < 1$,

$$\mathbb{P}\big(\mathrm{rank}(X_n) \geq \mathrm{rank}(\overline{X})\big) \geq \mathbb{P}\big(|\sigma_r(X_m) - \sigma_r(\overline{X})| \leq \varepsilon\sigma_r(\overline{X})\big) \to 1 \quad \text{as} \quad m \to \infty.$$

Alternatively, this result can also be proved by the lower semicontinuity of the rank function. □

In what follows, we focus on the characterization of necessary and sufficient conditions for rank consistency of $\widehat{X}_m$. The idea is similar to that of [7] for the nuclear norm penalized least squares estimator. Note that, unlike for the recover error bound, adding more constraints may break the rank consistency. Therefore, we separate the discussion into the rectangular case (recovering a rectangular matrix or a symmetric/Hermitian matrix) and the positive semidefinite case (recovering a symmetric/Hermitian positive semidefinite matrix) below.

3.3.1 The rectangular case

Since we have established that $\widehat{X}_m \xrightarrow{p} \overline{X}$, we only need to focus on some neighborhood of $\overline{X}$ for the discussion about the rank consistency of $\widehat{X}_m$. First, we take a

look at a local property of the rank function via the directional derivative of the singular value functions.

Let $\sigma_i'(X;\cdot)$ denote the directional derivative function of the i-th largest singular value function $\sigma_i(\cdot)$ at X. From [105, Section 5.1] and [34, Proposition 6], for $\mathbb{V}^{n_1\times n_2} \ni H \to 0$,

$$\sigma_i(X+H) - \sigma_i(X) - \sigma_i'(X;H) = O(\|H\|_F^2), \quad i = 1,\ldots,n. \tag{3.37}$$

Recall that $r = \operatorname{rank}(\overline{X})$. From [34, Proposition 6], we have

$$\sigma_{r+1}'(\overline{X};H) = \|\overline{U}_2^{\mathbb{T}} H \overline{V}_2\|, \qquad H \in \mathbb{V}^{n_1\times n_2}.$$

This leads to the following result for the perturbation of the rank function. A similar result can also be found in [7, Proposition 18], whose proof is more involved.

Lemma 3.10. *Let $\overline{\Delta} \in \mathbb{V}^{n_1\times n_2}$ satisfy $\overline{U}_2^{\mathbb{T}}\overline{\Delta}\,\overline{V}_2 \neq 0$. Then for all $\rho \neq 0$ sufficiently small and Δ sufficiently close to $\overline{\Delta}$, we have*

$$\operatorname{rank}(\overline{X} + \rho\Delta) > \operatorname{rank}(\overline{X}).$$

Proof. By replacing X and H in (3.37) with $\overline{X}$ and $\rho\Delta$, respectively, and noting that $\sigma_{r+1}(\overline{X}) = 0$, we have

$$\sigma_{r+1}(\overline{X} + \rho\Delta) - \|\overline{U}_2^{\mathbb{T}}(\rho\Delta)\overline{V}_2\| = O(\|\rho\Delta\|_F^2).$$

Since $\overline{U}_2^{\mathbb{T}}\overline{\Delta}\,\overline{V}_2 \neq 0$, for any $\rho \neq 0$ sufficiently small and Δ sufficiently close to $\overline{\Delta}$,

$$\begin{aligned}\frac{\sigma_{r+1}(\overline{X}+\rho\Delta)}{|\rho|} &= \|\overline{U}_2^{\mathbb{T}}\Delta\overline{V}_2\| + O(|\rho|\|\Delta\|_F^2)\\ &\geq \|\overline{U}_2^{\mathbb{T}}\overline{\Delta}\,\overline{V}_2\| - \|\overline{U}_2^{\mathbb{T}}(\Delta-\overline{\Delta})\overline{V}_2\| + O(|\rho|\|\Delta\|_F^2)\\ &\geq \frac{1}{2}\|\overline{U}_2^{\mathbb{T}}\overline{\Delta}\,\overline{V}_2\| > 0.\end{aligned}$$

This implies that $\operatorname{rank}(\overline{X} + \rho\Delta) > r$. □

To guarantee the efficiency of the rank-correction term on encouraging a low-rank solution, the parameter ρ_m should not decay too fast. Define

$$\widehat{\Delta}_m := \rho_m^{-1}(\widehat{X}_m - \overline{X}).$$

Then, for a slow decay on ρ_m, we can establish the following result.

Proposition 3.11. *If $\rho_m \to 0, \sqrt{m}\rho_m \to \infty$ and $\gamma_m = O_p(1)$, then $\widehat{\Delta}_m \xrightarrow{p} \widehat{\Delta}$, where $\widehat{\Delta}$ is the unique optimal solution to the following convex optimization problem*

$$\begin{aligned} \min_{\Delta \in \mathbb{V}^{n_1 \times n_2}} \quad & \frac{1}{2}\langle \mathcal{Q}_\beta(\Delta), \Delta \rangle + \langle \overline{U}_1 \overline{V}_1^{\mathbb{T}} - F(\overline{X}), \Delta \rangle + \|\overline{U}_2^{\mathbb{T}} \Delta \overline{V}_2\|_* \\ \text{s.t.} \quad & \mathcal{R}_\alpha(\Delta) = 0. \end{aligned} \tag{3.38}$$

Proof. By letting $\Delta := \rho_m^{-1}(X - \overline{X})$ in the optimization problem (3.38), one can easily see that $\widehat{\Delta}_m$ is the optimal solution to

$$\begin{aligned} \min_{\Delta \in \mathbb{V}^{n_1 \times n_2}} \quad & \frac{1}{2m}\|\mathcal{R}_\Omega(\Delta)\|_2^2 - \frac{\nu}{m\rho_m}\langle \mathcal{R}_\Omega^*(\xi), \Delta \rangle + \frac{1}{\rho_m}\big(\|\overline{X} + \rho_m \Delta\|_* - \|\overline{X}\|_*\big) \\ & - \langle F(\widetilde{X}_m), \Delta \rangle + \frac{\rho_m \gamma_m}{2}\|\Delta\|_F^2 + \gamma_m \langle \overline{X} - \widetilde{X}_m, \Delta \rangle \\ \text{s.t.} \quad & \mathcal{R}_\alpha(\Delta) = 0. \end{aligned} \tag{3.39}$$

Let Φ_m and Φ denote the objective functions of (3.39) and (3.38), respectively. Let K denote the feasible set of (3.38). By the definition of directional derivative and [176, Theorem 1], we have

$$\lim_{\rho_m \to 0} \frac{1}{\rho_m}\big(\|\overline{X} + \rho_m \Delta\|_* - \|\overline{X}\|_*\big) = \langle \overline{U}_1 \overline{V}_1^{\mathbb{T}}, \Delta \rangle + \|\overline{U}_2^{\mathbb{T}} \Delta \overline{V}_2\|_*.$$

Then, by combining Assumptions 3.3 and 3.4 with Lemma 3.7, we obtain that Φ_m converges pointwise in probability to Φ. Then from Theorems 2.16 and 2.17, we obtain that

$$\Phi_m + \delta_K \xrightarrow{e-d} \Phi + \delta_K.$$

Moreover, the optimal solution to (3.38) is unique due to the strong convexity of Φ over the feasible set K. Therefore, we complete the proof by applying Theorem 2.15 on epi-convergence in distribution. □

Note that $\widehat{X}_m = \overline{X} + \rho_m \widehat{\Delta}_m$. From Theorem 3.8, Lemmas 3.9 and 3.10 and Proposition 3.11, we see that the condition $\overline{U}_2^{\mathbb{T}} \widehat{\Delta} \overline{V}_2 = 0$ is necessary for the rank consistency of $\widehat{X}_m$. From the following property of the unique solution $\widehat{\Delta}$ to (3.38), we can derive a more detailed necessary condition for rank consistency as stated in Theorem 3.13 below.

Lemma 3.12. *Let $\widehat{\Delta}$ be the optimal solution to (3.38). Then $\overline{U}_2^{\mathbb{T}} \widehat{\Delta} \overline{V}_2 = 0$ if and only if the linear system*

$$\overline{U}_2^{\mathbb{T}} \mathcal{Q}_\beta^\dagger (\overline{U}_2 \Gamma \overline{V}_2^{\mathbb{T}}) \overline{V}_2 = \overline{U}_2^{\mathbb{T}} \mathcal{Q}_\beta^\dagger \big(\overline{U}_1 \overline{V}_1^{\mathbb{T}} - F(\overline{X}) \big) \overline{V}_2 \tag{3.40}$$

has a solution $\widehat{\Gamma} \in \mathbb{V}^{(n_1-r)\times(n_2-r)}$ with $\|\widehat{\Gamma}\| \leq 1$. Moreover, in this case,

$$\widehat{\Delta} = \mathcal{Q}_\beta^\dagger \big(\overline{U}_2 \widehat{\Gamma}\, \overline{V}_2^{\mathbb{T}} - \overline{U}_1 \overline{V}_1^{\mathbb{T}} + F(\overline{X}) \big). \tag{3.41}$$

Proof. Assume that $\overline{U}_2^{\mathbb{T}} \widehat{\Delta} \overline{V}_2 = 0$. Since $\widehat{\Delta}$ is the optimal solution to (3.38), from the optimality condition, the subdifferential of $\|X\|_*$ at 0, and [153, Theorem 23.7], we obtain that there exist some $\widehat{\Gamma} \in \mathbb{V}^{(n_1-r)\times(n_2-r)}$ with $\|\widehat{\Gamma}\| \leq 1$ and $\widehat{\eta} \in \mathbb{R}^{d_1}$ such that

$$\begin{cases} \mathcal{Q}_\beta(\widehat{\Delta}) + \overline{U}_1 \overline{V}_1^{\mathbb{T}} - F(\overline{X}) - \mathcal{R}_\alpha^*(\widehat{\eta}) - \overline{U}_2 \widehat{\Gamma}\, \overline{V}_2^{\mathbb{T}} = 0, \\ \mathcal{R}_\alpha(\widehat{\Delta}) = 0. \end{cases} \tag{3.42}$$

Then, according to (3.4), we can easily obtain (3.41) by applying the operator $\mathcal{Q}_\beta^\dagger$ to the first equation of (3.42) and using the second equation. By further combining (3.41) and $\overline{U}_2^{\mathbb{T}} \widehat{\Delta} \overline{V}_2 = 0$, we obtain that $\widehat{\Gamma}$ is a solution to the linear system (3.40).

Conversely, if the linear system (3.40) has a solution $\widehat{\Gamma}$ with $\|\widehat{\Gamma}\| \leq 1$, then it is easy to check that (3.42) is satisfied with $\widehat{\Delta}$ taking the form of (3.41) and $\widehat{\eta} = \mathcal{R}_\alpha \big(\overline{U}_1 \overline{V}_1^{\mathbb{T}} - F(\overline{X}) - \overline{U}_2 \widehat{\Gamma}\, \overline{V}_2^{\mathbb{T}} \big)$. Consequently, $\overline{U}_2^{\mathbb{T}} \widehat{\Delta} \overline{V}_2 = 0$ follows directly from the equations (3.40) and (3.41). □

Theorem 3.13. *If $\rho_m \to 0$, $\sqrt{m}\rho_m \to \infty$ and $\gamma_m = O_p(1)$, then a necessary condition for the rank consistency of $\widehat{X}_m$ is that the linear system (3.40) has a solution $\widehat{\Gamma} \in \mathbb{V}^{(n_1-r)\times(n_2-r)}$ with $\|\widehat{\Gamma}\| \leq 1$.*

Proof. This result is immediate from Lemma 3.12 and the fact that $\overline{U}_2^{\mathbb{T}}\widehat{\Delta}\overline{V}_2 = 0$ is necessary for rank consistency. □

By making a slight modification for the necessary condition in Theorem 3.13, we provide a sufficient condition for the rank consistency of the estimator $\widehat{X}_m$ as follows.

Theorem 3.14. *If $\rho_m \to 0, \sqrt{m}\rho_m \to \infty$ and $\gamma_m = O_p(1)$, then a sufficient condition for the rank consistency of the estimator $\widehat{X}_m$ is that the linear system (3.40) has a unique solution $\widehat{\Gamma} \in \mathbb{V}^{(n_1-r)\times(n_2-r)}$ with $\|\widehat{\Gamma}\| < 1$.*

Proof. The estimator $\widehat{X}_m$ is an optimal solution to (3.8) if and only if there exist a subgradient $\widehat{G}_m$ of the nuclear norm at $\widehat{X}_m$ and a vector $\widehat{\eta}_m \in \mathbb{R}^{d_1}$ such that $(\widehat{X}_m, \widehat{\eta}_m)$ satisfies the KKT conditions:

$$\begin{cases} \dfrac{1}{m}\mathcal{R}_\Omega^*\big(\mathcal{R}_\Omega(\widehat{X}_m) - y\big) + \rho_m\big(\widehat{G}_m - F(\widetilde{X}_m) + \gamma_m(\widehat{X}_m - \widetilde{X}_m)\big) - \mathcal{R}_\alpha^*(\widehat{\eta}_m) = 0, \\ \mathcal{R}_\alpha(\widehat{X}_m) = \mathcal{R}_\alpha(\overline{X}). \end{cases} \tag{3.43}$$

Let $(\widehat{U}_m, \widehat{V}_m) \in \mathbb{O}^{n_1,n_2}(\widehat{X}_m)$. From Theorem 3.8 and Lemma 3.9, we know that $\mathrm{rank}(\widehat{X}_m) \geq r$ with probability one. When $\mathrm{rank}(\widehat{X}_m) \geq r$ holds, from the characterization of the subdifferential of the nuclear norm [176, 177], we have that

$$\widehat{G}_m = \widehat{U}_{m,1}\widehat{V}_{m,1}^{\mathbb{T}} + \widehat{U}_{m,2}\widehat{\Gamma}_m\widehat{V}_{m,2}^{\mathbb{T}}$$

for some $\widehat{\Gamma}_m \in \mathbb{V}^{(n_1-r)\times(n_2-r)}$ satisfying $\|\widehat{\Gamma}_m\| \leq 1$. Moreover, if $\|\widehat{\Gamma}_m\| < 1$, then $\mathrm{rank}(\widehat{X}_m) = r$. Since $\widehat{X}_m \xrightarrow{p} \overline{X}$, by [34, Proposition 8] we have $\widehat{U}_{m,1}\widehat{V}_{m,1}^{\mathbb{T}} \xrightarrow{p} \overline{U}_1\overline{V}_1^{\mathbb{T}}$. Together with Lemma 3.7, the equation (3.5) and Lemma 3.12, it is not hard to obtain that

$$\begin{aligned} &\frac{1}{m\rho_m}\mathcal{R}_\Omega^*\big(\mathcal{R}_\Omega(\widehat{X}_m) - y\big) + \widehat{U}_{m,1}\widehat{V}_{m,1}^{\mathbb{T}} - F(\widetilde{X}_m) + \gamma_m(\widehat{X}_m - \widetilde{X}_m) \\ &\qquad \xrightarrow{p} \mathcal{Q}_\beta(\widehat{\Delta}) + \overline{U}_1\overline{V}_1^{\mathbb{T}} - F(\overline{X}) = \overline{U}_2\widehat{\Gamma}\overline{V}_2^{\mathbb{T}}, \end{aligned} \tag{3.44}$$

where the equality follows from (3.41) and $\widehat{\Gamma}$ is the unique optimal solution to (3.40). Then, by applying the operator $\mathcal{Q}_\beta^\dagger$ to (3.43), we obtain from (3.44) that

$$\overline{U}_2^{\mathbb{T}}\mathcal{Q}_\beta^\dagger(\widehat{U}_{m,2}\widehat{\Gamma}_m\widehat{V}_{m,2}^{\mathbb{T}})\overline{V}_2 \xrightarrow{p} \overline{U}_2^{\mathbb{T}}\mathcal{Q}_\beta^\dagger(\overline{U}_2\widehat{\Gamma}\overline{V}_2^{\mathbb{T}})\overline{V}_2. \tag{3.45}$$

Since $\widehat{X}_m \xrightarrow{p} \overline{X}$, according to [34, Proposition 7], there exist two sequences of matrices $Q_{m,U} \in \mathbb{O}^{n_1-r}$ and $Q_{m,V} \in \mathbb{O}^{n_2-r}$ such that

$$\widehat{U}_{m,2}Q_{m,U} \xrightarrow{p} \overline{U}_2 \quad \text{and} \quad \widehat{V}_{m,2}Q_{m,V} \xrightarrow{p} \overline{V}_2. \tag{3.46}$$

Moreover, the uniqueness of the solution to the linear system (3.40) is equivalent to the non-singularity of its linear operator $\overline{U}_2^{\mathbb{T}}\mathcal{Q}_\beta^\dagger(\overline{U}_2(\cdot)\overline{V}_2^{\mathbb{T}})\overline{V}_2$. By combining (3.45) and (3.46), we obtain that

$$Q_{m,U}^{\mathbb{T}}\widehat{\Gamma}_m Q_{m,V} \xrightarrow{p} \widehat{\Gamma}.$$

Hence, we obtain that $\|\widehat{\Gamma}_m\| < 1$ with probability one since $\|\widehat{\Gamma}\| < 1$. As discussed above, it follows that $\operatorname{rank}(\widehat{X}_m) = r$ with probability one. □

3.3.2 The positive semidefinite case

For the positive semidefinite case, we first need the following Slater condition.

Assumption 3.5. *There exists some $X^0 \in \mathbb{S}_{++}^n$ such that $\mathcal{R}_\alpha(X^0) = \mathcal{R}_\alpha(\overline{X})$.*

Proposition 3.15. *If $\rho_m \to 0, \sqrt{m}\rho_m \to \infty$ and $\gamma_m = O_p(1)$, then $\widehat{\Delta}_m \xrightarrow{p} \widehat{\Delta}$, where $\widehat{\Delta}$ is the unique optimal solution to the following convex optimization problem*

$$\begin{array}{ll} \min\limits_{\Delta\in\mathbb{S}^n} & \frac{1}{2}\langle\mathcal{Q}_\beta(\Delta),\Delta\rangle + \langle I_n - F(\overline{X}),\Delta\rangle \\ \text{s.t.} & \mathcal{R}_\alpha(\Delta) = 0, \quad \overline{P}_2^{\mathbb{T}}\Delta\overline{P}_2 \in \mathbb{S}_+^{n-r}. \end{array} \tag{3.47}$$

Proof. It is easy to verify that $\widehat{\Delta}_m$ is the optimal solution to

$$\begin{aligned}\min_{\Delta\in\mathbb{S}^n}\ & \frac{1}{2m}\|\mathcal{R}_\Omega(\Delta)\|_2^2 - \frac{\nu}{m\rho_m}\langle\mathcal{R}_\Omega^*(\xi),\Delta\rangle + \langle I_n - F(\widetilde{X}_m),\Delta\rangle + \frac{\rho_m\gamma_m}{2}\|\Delta\|_F^2 \\ & + \gamma_m\langle\overline{X} - \widetilde{X}_m,\Delta\rangle \qquad (3.48)\\ \text{s.t.}\ & \Delta \in K_m := \rho_m^{-1}(E\cap\mathbb{S}_+^n - \overline{X}),\end{aligned}$$

where $E := \{X \in \mathbb{S}^n \mid \mathcal{R}_\alpha(X) = \mathcal{R}_\alpha(\overline{X})\}$. Let Φ_m and Φ denote the objective functions of (3.48) and (3.47), respectively. Then Φ_m converges pointwise in probability to Φ. Moreover, by considering the upper limit and lower limit of the family of feasible sets K_m, we know that K_m converges in the sense of Painlevé-Kuratowski to the tangent cone $\mathcal{T}_{E\cap\mathbb{S}_+^n}(\overline{X})$ (see [155, 15]). Note that the Slater condition implies that E and $\mathbb{S}_+^n$ cannot be separated. Then, from [155, Theorem 6.42], we have

$$\mathcal{T}_{E\cap\mathbb{S}_+^n}(\overline{X}) = \mathcal{T}_E(\overline{X})\cap\mathcal{T}_{\mathbb{S}_+^n}(\overline{X}).$$

Elearly, $\mathcal{T}_E(\overline{X}) = \{\Delta\in\mathbb{S}^n \mid \mathcal{R}_\alpha(\Delta) = 0\}$. Moreover, by Arnold [3],

$$\mathcal{T}_{\mathbb{S}_+^n}(\overline{X}) = \left\{\Delta\in\mathbb{S}^n \;\middle|\; \overline{P}_2^{\mathbb{T}}\Delta\overline{P}_2 \in \mathbb{S}_+^{n-r}\right\}.$$

Since epi-convergence of functions corresponds to set convergence of their epigraphs [155], we obtain that $\delta_{K_m} \xrightarrow{e} \delta_{\mathcal{T}_{E\cap\mathbb{S}_+^n}} = \delta_{\mathcal{T}_E} + \delta_{\mathcal{T}_{\mathbb{S}_+^n}}$. Then, from Theorem 2.16,

$$\Phi_m + \delta_{K_m} \xrightarrow{e-d} \Phi + \delta_{\mathcal{T}_E} + \delta_{\mathcal{T}_{\mathbb{S}_+^n}}.$$

In addition, the optimal solution to (3.47) is unique due to the strong convexity of Φ over the feasible set $E\cap\mathbb{S}_+^n$. Therefore, we complete the proof by applying Theorem 2.15 on epi-convergence in distribution. □

For the optimal solution $\widehat{\Delta}$ to (3.47), we also have the following further characterization.

Lemma 3.16. *Let $\widehat{\Delta}$ be the optimal solution to (3.47). Then $\overline{P}_2^{\mathbb{T}}\widehat{\Delta}\overline{P}_2 = 0$ if and only if the linear system*

$$\overline{P}_2^{\mathbb{T}}\mathcal{Q}_\beta^\dagger(\overline{P}_2\Lambda\overline{P}_2^{\mathbb{T}})\overline{P}_2 = \overline{P}_2^{\mathbb{T}}\mathcal{Q}_\beta^\dagger\big(I_n - F(\overline{X})\big)\overline{P}_2 \tag{3.49}$$

has a solution $\widehat{\Lambda} \in \mathbb{S}_+^{n-r}$. Moreover, in this case,

$$\widehat{\Delta} = \mathcal{Q}_\beta^\dagger\big(\overline{P}_2\widehat{\Lambda}\,\overline{P}_2^{\mathbb{T}} - I_n + F(\overline{X})\big). \tag{3.50}$$

Proof. Note that the Slater condition also holds for the problem (3.47). (One may check the point $X^0 - \overline{X}$.) Hence, $\widehat{\Delta}$ is the optimal solution to (3.47) if and only if there exists $(\widehat{\zeta}, \widehat{\Lambda}) \in \mathbb{R}^{d_1} \times \mathbb{S}^{n-r}$ such that

$$\begin{cases} \mathcal{Q}_\beta(\widehat{\Delta}) + I_n - F(\overline{X}) - \mathcal{R}_\alpha^*(\widehat{\zeta}) - \overline{P}_2\widehat{\Lambda}\overline{P}_2^{\mathbb{T}} = 0, \\ \mathcal{R}_\alpha(\widehat{\Delta}) = 0, \\ \overline{P}_2^{\mathbb{T}}\widehat{\Delta}\overline{P}_2 \in \mathbb{S}_+^{n-r},\ \widehat{\Lambda} \in \mathbb{S}_+^{n-r},\ \langle \overline{P}_2^{\mathbb{T}}\widehat{\Delta}\overline{P}_2, \widehat{\Lambda}\rangle = 0. \end{cases} \tag{3.51}$$

Applying the operator $\mathcal{Q}_\beta^\dagger$ to the first equation of (3.51) yields the equality (3.50). Suppose that $\overline{P}_2^{\mathbb{T}}\widehat{\Delta}\overline{P}_2 = 0$. Then, it is immediate to obtain from (3.51) that $\widehat{\Lambda}$ is a solution to the linear system (3.49).

Conversely, if the linear system (3.49) has a solution $\widehat{\Lambda} \in \mathbb{S}_+^{n-r}$, then it is easy to check that (3.51) is satisfied with $\widehat{\Delta}$ taking the form of (3.50) and $\widehat{\zeta} = \mathcal{R}_\alpha\big(I_n - F(\overline{X}) - \overline{P}_2\widehat{\Lambda}\,\overline{P}_2^{\mathbb{T}}\big)$. Then, $\overline{P}_2^{\mathbb{T}}\widehat{\Delta}\overline{P}_2 = 0$ directly follows from (3.50) and the first equation of (3.51). □

Note that Lemma 3.10 still holds for the positive semidefinite case if $\overline{U}_2^{\mathbb{T}}\Delta\overline{V}_2$ is replaced by $\overline{P}_2^{\mathbb{T}}\Delta\overline{P}_2$. Therefore, in line with the rectangular case, from Lemma 3.16, we have the following necessary condition for rank consistency.

Theorem 3.17. *If $\rho_m \to 0, \sqrt{m}\rho_m \to \infty$ and $\gamma_m = O_p(1)$, then a necessary condition for the rank consistency of $\widehat{X}_m$ is that the linear system (3.49) has a solution $\widehat{\Lambda} \in \mathbb{S}_+^{n-r}$.*

Proof. This result is immediate from Lemma 3.16 and the fact that $\overline{P}_2^{\mathbb{T}}\widehat{\Delta}\overline{P}_2 = 0$ is necessary for rank consistency. □

Similarly to Theorem 3.14, we have the following sufficient condition for rank consistency for the positive semidefinite case.

Theorem 3.18. *If $\rho_m \to 0, \sqrt{m}\rho_m \to \infty$ and $\gamma_m = O_p(1)$, then a sufficient condition for the rank consistency of $\widehat{X}_m$ is that the linear system (3.49) has a unique solution $\widehat{\Lambda} \in \mathbb{S}^{n-r}_{++}$.*

Proof. The Slater condition implies that $\widehat{X}_m$ is the optimal solution to (3.9) if and only if there exists multipliers $(\widehat{\zeta}_m, \widehat{S}_m) \in \mathbb{R}^{d_1} \times \mathbb{S}^n$ such that $(\widehat{X}_m, \widehat{\zeta}_m, \widehat{S}_m)$ satisfy the KKT conditions:

$$\begin{cases} \dfrac{1}{m}\mathcal{R}^*_\Omega\big(\mathcal{R}_\Omega(\widehat{X}_m) - y\big) + \rho_m\big(I_n - F(\widetilde{X}_m) + \gamma_m(\widehat{X}_m - \widetilde{X}_m)\big) - \mathcal{R}^*_\alpha(\widehat{\zeta}_m) - \widehat{S}_m = 0, \\ \mathcal{R}_\alpha(\widehat{X}_m) = \mathcal{R}_\alpha(\overline{X}), \\ \widehat{X}_m \in \mathbb{S}^n_+,\ \widehat{S}_m \in \mathbb{S}^n_+,\ \langle \widehat{X}_m, \widehat{S}_m\rangle = 0. \end{cases} \tag{3.52}$$

The third equation of (3.52) implies that $\widehat{X}_m$ and $\widehat{S}_m$ can have a simultaneous eigenvalue decomposition. Let $\widehat{P}_m \in \mathbb{O}^n(\widehat{X}_m)$. From Theorem 3.8 and Lemma 3.9, we know that $\mathrm{rank}(\widehat{X}_m) \geq r$ with probability one. When $\mathrm{rank}(\widehat{X}_m) \geq r$ holds, we can write

$$\widehat{S}_m = \widehat{P}_{m,2}\widehat{\Lambda}_m\widehat{P}^{\mathbb{T}}_{m,2}$$

for some diagonal matrix $\widehat{\Lambda}_m \in \mathbb{S}^{n-r}_+$. In addition, if $\widehat{\Lambda}_m \in \mathbb{S}^{n-r}_{++}$, then $\mathrm{rank}(\widehat{X}_m) = r$. Since $\widehat{X}_m \xrightarrow{p} \overline{X}$, according to [34, Proposition 1], there exists a sequence of matrices $Q_m \in \mathbb{O}^{n-r}$ such that $\widehat{P}_{m,2}Q_m \xrightarrow{p} \overline{P}_2$. Then, using the similar arguments to the proof of Theorem 3.14, we obtain that

$$Q_m^{\mathbb{T}}\widehat{\Lambda}_m Q_m \xrightarrow{p} \widehat{\Lambda}.$$

Since $\widehat{\Lambda} \in \mathbb{S}^n_{++}$, we have $\widehat{\Lambda}_m \in \mathbb{S}^n_{++}$ with probability one. Thus, we complete the proof. □

3.3.3 Constraint nondegeneracy and rank consistency

In this subsection, with the help of constraint nondegeneracy, we provide conditions to guarantee that the linear systems (3.40) and (3.49) have a unique solution. The concept of constraint nondegeneracy was pioneered by Robinson [152] and later extensively developed by Bonnans and Shapiro [15]. Consider the following constrained optimization problem

$$\begin{aligned} \min_{X \in \mathbb{V}^{n_1 \times n_2}} \quad & \Phi(X) + \Psi(X) \\ \text{s.t.} \quad & \mathcal{A}(X) - b \in K, \end{aligned} \tag{3.53}$$

where $\Phi : \mathbb{V}^{n_1 \times n_2} \to \mathbb{R}$ is a continuously differentiable function, $\Psi : \mathbb{V}^{n_1 \times n_2} \to \mathbb{R}$ is a convex function, $\mathcal{A} : \mathbb{V}^{n_1 \times n_2} \to \mathbb{R}^l$ is a linear operator, $b \in \mathbb{R}^l$ is a given vector and $K \subseteq \mathbb{R}^l$ is a closed convex set. Let $\widehat{X}$ be a given feasible point of (3.53) and $\widehat{z} := \mathcal{A}(\widehat{X}) - b$.

When Ψ is differentiable at $\widehat{X}$, we say that the constraint nondegeneracy holds at $\widehat{X}$ if

$$\mathcal{A}\,\mathbb{V}^{n_1 \times n_2} + \mathrm{lin}\big(\mathcal{T}_K(\widehat{z})\big) = \mathbb{R}^l, \tag{3.54}$$

where $\mathcal{T}_K(\widehat{z})$ denotes the tangent cone of K at $\widehat{z}$ and $\mathrm{lin}(\mathcal{T}_K(\widehat{z}))$ denotes the largest linearity space contained in $\mathcal{T}_K(\widehat{z})$, i.e.,

$$\mathrm{lin}(\mathcal{T}_K(\widehat{z})) = \mathcal{T}_K(\widehat{z}) \cap (-\mathcal{T}_K(\widehat{z})).$$

When the function Ψ is nondifferentiable, we can rewrite the optimization problem (3.53) equivalently as

$$\begin{aligned} \min_{(X,t) \in \mathbb{V}^{n_1 \times n_2} \times \mathbb{R}} \quad & \Phi(X) + t \\ \text{s.t.} \quad & \widetilde{\mathcal{A}}(X,t) \in K \times \mathrm{epi}\Psi, \end{aligned}$$

where epiΨ denotes the epigraph of Ψ and $\widetilde{\mathcal{A}}: \mathbb{V}^{n_1\times n_2}\times\mathbb{R}\to\mathbb{R}^l\times\mathbb{V}^{n_1\times n_2}\times\mathbb{R}$ is a linear operator defined by

$$\widetilde{\mathcal{A}}(X,t) := \begin{pmatrix} \mathcal{A}(X)-b \\ X \\ t \end{pmatrix}, \quad (X,t)\in\mathbb{V}^{n_1\times n_2}\times\mathbb{R}.$$

From (3.54) and [155, Theorem 6.41], the constraint nondegeneracy holds at $(\widehat{X},\widehat{t})$ with $\widehat{t}=\Psi(\widehat{X})$ if

$$\widetilde{\mathcal{A}}\begin{pmatrix}\mathbb{V}^{n_1\times n_2}\\ \mathbb{R}\end{pmatrix} + \begin{pmatrix}\operatorname{lin}\big(\mathcal{T}_K(\widehat{X})\big)\\ \operatorname{lin}\big(\mathcal{T}_{\mathrm{epi}\Psi}(\widehat{X},\widehat{t})\big)\end{pmatrix} = \begin{pmatrix}\mathbb{R}^l\\ \mathbb{V}^{n_1\times n_2}\\ \mathbb{R}\end{pmatrix}.$$

By the definition of $\widetilde{\mathcal{A}}$, it is not difficult to verify that this condition is equivalent to

$$[\mathcal{A}\ \ 0]\big(\operatorname{lin}(\mathcal{T}_{\mathrm{epi}\Psi}(\widehat{X},\widehat{t}))\big) + \operatorname{lin}\big(\mathcal{T}_K(\widehat{X})\big) = \mathbb{R}^l. \tag{3.55}$$

By letting $\Psi=\|\cdot\|_*, \mathcal{A}=\mathcal{R}_\alpha$ and $K=\{0\}$, one can see that the problem (3.8) takes the form of (3.53). By the expression of $\mathcal{T}_{\mathrm{epi}\Psi}(\overline{X},\bar{t})$ with $\bar{t}=\|\overline{X}\|_*$ (e.g., see [79]), we see that for the problem (3.8), the condition (3.55) reduces to

$$\mathcal{R}_\alpha\big(\mathcal{T}(\overline{X})\big)=\mathbb{R}^{d_1}, \tag{3.56}$$

where

$$\mathcal{T}(\overline{X}) = \big\{H\in\mathbb{V}^{n_1\times n_2}\ \big|\ \overline{U}_2^{\mathbb{T}}H\overline{V}_2=0\big\}. \tag{3.57}$$

Hence, we say that the constraint nondegeneracy holds at $\overline{X}$ to the problem (3.8) if the condition (3.56) holds. By letting $\Psi=\delta_{\mathbb{S}^n_+}, \mathcal{A}=\mathcal{R}_\alpha$ and $K=\{0\}$, we can see that the problem (3.9) takes the form of (3.53) , and now that the condition (3.55) reduces to

$$\mathcal{R}_\alpha\big(\operatorname{lin}(\mathcal{T}_{\mathbb{S}^n_+}(\overline{X}))\big)=\mathbb{R}^{d_1}. \tag{3.58}$$

Thus, we say that the constraint nondegeneracy holds at $\overline{X}$ to the problem (3.9) if the condition (3.58) holds. From Arnold's characterization of the tangent cone in [3]:

$$\mathcal{T}_{\mathbb{S}_+^n}(\overline{X}) = \big\{H \in \mathbb{S}^n \;\big|\; \overline{P}_2^{\mathbb{T}} H \overline{P}_2 \in \mathbb{S}_+^{n-r}\big\},$$

we can write the linearity space $\mathrm{lin}(\mathcal{T}_{\mathbb{S}_+^n}(\overline{X}))$ explicitly as

$$\mathrm{lin}(\mathcal{T}_{\mathbb{S}_+^n}(\overline{X})) = \big\{H \in \mathbb{S}^n \;\big|\; \overline{P}_2^{\mathbb{T}} H \overline{P}_2 = 0\big\}.$$

Interestingly, for some special matrix completion problems, the constraint nondegeneracy automatically hold at $\overline{X}$, as stated in the following proposition.

Proposition 3.19. *For the following matrix completion problems:*

(i) *the covariance matrix completion with partial positive diagonal entries being fixed , in particular, the correlation matrix completion with all diagonal entries being fixed as ones;*

(ii) *the density matrix completion with its trace being fixed as one,*

the constraint nondegeneracy (3.58) holds at $\overline{X}$.

Proof. For the real covariance matrix case, the proof is given in [144, Lemma 3.3] and [145, Proposition 2.1]. For the complex covariance matrix case, one can use the similar arguments to prove the result.

We next consider the density matrix case. Suppose that $\overline{X}$ satisfies the density constraint, i.e., $\mathcal{R}_\alpha(\overline{X}) = \mathrm{Tr}(\overline{X}) = 1$. Note that for any $t \in \mathbb{R}$, we have $t\overline{X} \in \mathrm{lin}(\mathcal{T}_{\mathcal{H}_+^n}(\overline{X}))$. This, along with $\mathrm{Tr}(\overline{X}) = 1$, implies that

$$\mathrm{Tr}\big(\mathrm{lin}(\mathcal{T}_{\mathcal{H}_+^n}(\overline{X}))\big) = \mathcal{R}_\alpha\big(\mathrm{lin}(\mathcal{T}_{\mathcal{H}_+^n}(\overline{X}))\big) = \mathbb{R}.$$

This means that the constraint nondegeneracy condition (3.58) holds. □

Next, we take a closer look at the solutions to the linear systems (3.40) and (3.49). Define linear operators $\mathcal{B}_1 : \mathbb{V}^{r\times r} \to \mathbb{V}^{(n_1-r)\times(n_2-r)}$ and $\mathcal{B}_2 : \mathbb{V}^{(n_1-r)\times(n_2-r)} \to \mathbb{V}^{(n_1-r)\times(n_2-r)}$ associated with $\overline{X}$, respectively, by

$$\mathcal{B}_1(Y) := \overline{U}_2^{\mathbb{T}}\mathcal{Q}_\beta^\dagger(\overline{U}_1 Y \overline{V}_1^T)\overline{V}_2 \quad \text{and} \quad \mathcal{B}_2(Z) := \overline{U}_2^{\mathbb{T}}\mathcal{Q}_\beta^\dagger(\overline{U}_2 Z \overline{V}_2^T)\overline{V}_2, \tag{3.59}$$

where $Y \in \mathbb{V}^{r\times r}$ and $Z \in \mathbb{V}^{(n_1-r)\times(n_2-r)}$. Note that the operator $\mathcal{B}_2$ is self-adjoint and positive semidefinite according to the definition of $\mathcal{Q}_\beta^\dagger$. Let $\widehat{g}(\overline{X})$ be the vector in $\mathbb{R}^r$ defined by

$$\widehat{g}(\overline{X}) := \big(1 - f_1(\sigma(\overline{X})), \ldots, 1 - f_r(\sigma(\overline{X}))\big)^{\mathbb{T}}. \tag{3.60}$$

Then, by the definition of the spectral operator F, we can rewrite (3.40) in the following concise form

$$\mathcal{B}_2(\Gamma) = \mathcal{B}_1\big(\mathrm{Diag}(\widehat{g}(\overline{X}))\big), \quad \Gamma \in \mathbb{V}^{(n_1-r)\times(n_1-r)}. \tag{3.61}$$

For the positive semidefinite case $\mathbb{V}^{n_1\times n_2} = \mathbb{S}^n$ and $\overline{X} \in \mathbb{S}_+^n$, both $\overline{U}_i$ and $\overline{V}_i$ reduce to $\overline{P}_i$ for $i = 1, 2$. In this case, the linear system (3.49) can be concisely written as

$$\mathcal{B}_2(\Lambda) = \mathcal{B}_2(I_{n-r}) + \mathcal{B}_1\big(\mathrm{Diag}(\widehat{g}(\overline{X}))\big), \quad \Lambda \in \mathbb{S}^{n-r}. \tag{3.62}$$

Proposition 3.20. **(i)** *For the rectangular case, if the constraint nondegeneracy (3.56) holds at $\overline{X}$ to the problem (3.8), then the linear operators $\mathcal{B}_2$ defined by (3.59) is self-adjoint and positive definite.*

(ii) *For the positive semidefinite case, if the constraint nondegeneracy (3.58) holds at $\overline{X}$ to the problem (3.9), then the linear operators $\mathcal{B}_2$ is also self-adjoint and positive definite.*

Proof. We prove for the rectangular case by contradiction. Assume that there exists some $\mathbb{V}^{(n_1-r)\times(n_2-r)} \ni \overline{\Gamma} \neq 0$ such that $\mathcal{B}_2(\overline{\Gamma}) = \overline{U}_2^{\mathbb{T}}\mathcal{Q}_\beta^\dagger(\overline{U}_2\overline{\Gamma}\,\overline{V}_2^{\mathbb{T}})\overline{V}_2 = 0$. By noting that $\mathcal{Q}_\beta^\dagger$ is a self-adjoint and positive semidefinite operator, we obtain

$(\mathcal{Q}_\beta^\dagger)^{1/2}(\overline{U}_2\overline{\Gamma}\,\overline{V}_2^{\mathbb{T}}) = 0$. It follows that $\mathcal{P}_\beta(\overline{U}_2\overline{\Gamma}\,\overline{V}_2^{\mathbb{T}}) = 0$. This, together with $\overline{\Gamma} \neq 0$, implies that $\mathcal{P}_\alpha(\overline{U}_2\overline{\Gamma}\,\overline{V}_2^{\mathbb{T}}) = \overline{U}_2\overline{\Gamma}\,\overline{V}_2^{\mathbb{T}} \neq 0$. It further implies that $\mathcal{R}_\alpha(\overline{U}_2\overline{\Gamma}\,\overline{V}_2^{\mathbb{T}}) \neq 0$. However, for any $H \in \mathcal{T}(\overline{X})$, we have

$$\langle \mathcal{R}_\alpha(\overline{U}_2\overline{\Gamma}\,\overline{V}_2^{\mathbb{T}}), \mathcal{R}_\alpha(H)\rangle = \langle \mathcal{P}_\alpha(\overline{U}_2\overline{\Gamma}\,\overline{V}_2^{\mathbb{T}}), H\rangle = \langle \overline{U}_2\overline{\Gamma}\,\overline{V}_2^{\mathbb{T}}, H\rangle = \langle \overline{\Gamma}, \overline{U}_2^{\mathbb{T}} H \overline{V}_2\rangle = 0.$$

Thus, the constraint nondegeneracy condition (3.56) implies that $\mathcal{R}_\alpha(\overline{U}_2\overline{\Gamma}\,\overline{V}_2^{\mathbb{T}}) = 0$. This leads to a contradiction. Therefore, the linear operator $\mathcal{B}_2$ is positive definite. The proof for the positive semidefinite case is similar. □

According to Proposition 3.20, the constraint nondegeneracy at $\overline{X}$ to the problem (3.8) and (3.9), respectively, implies that the linear system (3.40) has a unique solution

$$\widehat{\Gamma} = \mathcal{B}_2^{-1}\mathcal{B}_1\big(\mathrm{Diag}(\widehat{g}(\overline{X}))\big) \tag{3.63}$$

and the linear system (3.49) has a unique solution

$$\widehat{\Lambda} = I_{n-r} + \mathcal{B}_2^{-1}\mathcal{B}_1\big(\mathrm{Diag}(\widehat{g}(\overline{X}))\big). \tag{3.64}$$

Then we can obtain the following main result for rank consistency.

Theorem 3.21. *Suppose that $\rho_m \to 0$, $\sqrt{m}\rho_m \to \infty$ and $\gamma_m = O_p(1)$. For the rectangular case, if the constraint nondegeneracy (3.56) holds at $\overline{X}$ to the problem (3.8) and*

$$\big\|\mathcal{B}_2^{-1}\mathcal{B}_1\big(\mathrm{Diag}(\widehat{g}(\overline{X}))\big)\big\| < 1, \tag{3.65}$$

then the estimator $\widehat{X}_m$ generated from the rank-correction step (3.8) is rank consistent. For the positive semidefinite case, if the constraint nondegeneracy (3.58) holds at $\overline{X}$ to the problem (3.9) and

$$I_{n-r} + \mathcal{B}_2^{-1}\mathcal{B}_1\big(\mathrm{Diag}(\widehat{g}(\overline{X}))\big) \in \mathbb{S}_{++}^{n-r}, \tag{3.66}$$

then the estimator $\widehat{X}_m$ generated from the rank-correction step (3.9) is rank consistent.

Proof. It is immediate from Theorems 3.14 and 3.18, Proposition 3.20 together with (3.63) and (3.64). □

From Theorem 3.21, it is not difficult to see that there exists some threshold $\overline{\varepsilon} > 0$ (depending on $\overline{X}$) such that the condition (3.65) holds if

$$|1 - f_i(\sigma(\overline{X}))| \leq \overline{\varepsilon} \quad \forall\, 1 \leq i \leq r.$$

In other words, when $F(\overline{X})$ is sufficiently close to $\overline{U}_1\overline{V}_1^{\mathbb{T}}$, the condition (3.65) holds automatically and so does the rank consistency. Thus, Theorem 3.21 provides us a guideline to construct a suitable rank-correction function for rank consistency. This is another important aspect of what we can benefit from the rank-correction step, besides the reduction of recovery error discussed in Section 3.2.

The next theorem shows that for the covariance (correlation) and density matrix completion problems with fixed basis coefficients described in Proposition 3.19, if observations are sampled uniformly at random, the rank consistency can be guaranteed for a broad class of rank-correction functions F.

Theorem 3.22. *For the covariance (correlation) and density matrix completion problems defined in Proposition 3.19 under uniform sampling, if* $\rho_m \to 0$, $\sqrt{m}\rho_m \to \infty$, $\gamma_m = O_p(1)$ *and* F *is a spectral operator associated with a symmetric function* $f : \mathbb{R}^n \to \mathbb{R}^n$ *such that*

$$\begin{cases} f_i(x) > 0 & \text{if } x_i > 0, \\ f_i(x) = 0 & \text{if } x_i = 0, \end{cases} \qquad \forall\, x \in \mathbb{R}^n_+ \text{ and } \forall\, i = 1, \ldots, n, \tag{3.67}$$

then the estimator $\widehat{X}_m$ *generated from the rank-correction step is rank consistent.*

Proof. From Propositions 3.19 and 3.20, for both cases, the linear system (3.49) has a unique solution $\widehat{\Lambda}$. Moreover, uniform sampling yields $\mathcal{Q}_\beta^\dagger = \mathcal{P}_\beta/d_2$. Thus, from (3.49), we get

$$\widehat{\Lambda} - \overline{P}_2^{\mathbb{T}}\mathcal{P}_\alpha(\overline{P}_2\widehat{\Lambda}\overline{P}_2^{\mathbb{T}})\overline{P}_2 = \overline{P}_2^{\mathbb{T}}\mathcal{P}_\beta(\overline{P}_2\widehat{\Lambda}\overline{P}_2^{\mathbb{T}})\overline{P}_2 = \overline{P}_2^{\mathbb{T}}\mathcal{P}_\beta\big(I_n - F(\overline{X})\big)\overline{P}_2. \tag{3.68}$$

We first prove the covariance matrix completion by contradiction. Without loss of generality, we assume that the first l diagonal entries are fixed and positive. Then, for any $X \in \mathbb{S}_+^n$, $\mathcal{P}_\alpha(X)$ is the diagonal matrix whose first l diagonal entries are $X_{ii}, 1 \le i \le l$ respectively and the other entries are 0. Assume that $\widehat{\Lambda} \notin \mathbb{S}_{++}^{n-r}$, i.e., $\lambda_{\min}(\widehat{\Lambda}) \le 0$, where $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue. Then, we have

$$\lambda_{\min}(\widehat{\Lambda}) = \lambda_{\min}(\overline{P}_2\widehat{\Lambda}\overline{P}_2^{\mathbb{T}}) \le \lambda_{\min}\big(\mathcal{P}_\alpha(\overline{P}_2\widehat{\Lambda}\overline{P}_2^{\mathbb{T}})\big) \le \lambda_{\min}\big(\overline{P}_2^{\mathbb{T}}\mathcal{P}_\alpha(\overline{P}_2\widehat{\Lambda}\overline{P}_2^{\mathbb{T}})\overline{P}_2\big),$$

where the equality follows from the fact that $\widehat{\Lambda}$ and $\overline{P}_2\widehat{\Lambda}\overline{P}_2^{\mathbb{T}}$ have the same nonzero eigenvalues, the first inequality follows from the fact that the vector of eigenvalues is majorized by the vector of diagonal entries (e.g., see [118, Theorem 9.B.1]), and the second inequality follows from the Courant-Fischer minmax theorem (e.g., see [118, Theorem 20.A.1]). As a result, the left-hand side of (3.68) is not positive definite. However, the right-hand side of (3.68) can be written as

$$\begin{aligned}\overline{P}_2^{\mathbb{T}}\mathcal{P}_\beta\big(I_n - F(\overline{X})\big)\overline{P}_2 &= \overline{P}_2^{\mathbb{T}}\left(\mathcal{P}_\beta(I_n) - F(\overline{X}) + \mathcal{P}_\alpha(F(\overline{X}))\right)\overline{P}_2 \\ &= \overline{P}_2^{\mathbb{T}}\big(\mathcal{P}_\beta(I_n) + \mathcal{P}_\alpha(F(\overline{X}))\big)\overline{P}_2,\end{aligned}$$

where the second equality follows from the fact that $\overline{P}_2^{\mathbb{T}}F(\overline{X})\overline{P}_2 = 0$. Since $\operatorname{rank}(\overline{X}) = r$, with the choice (3.67) of F, we have that for any $1 \le i \le l$,

$$\overline{X}_{ii} = \sum_{j=1}^{r}\lambda_j(\overline{X})|\overline{P}_{ij}|^2 > 0 \quad \Longrightarrow \quad \big(F(\overline{X})\big)_{ii} = \sum_{j=1}^{r} f_i\big(\lambda_j(\overline{X})\big)|\overline{P}_{ij}|^2 > 0.$$

Moreover, $\mathcal{P}_\beta(I_n)$ is the diagonal matrix with the last $n-r$ diagonal entries being ones and the other entries being zeros. Thus, $\mathcal{P}_\beta(I_n) + \mathcal{P}_\alpha\big(F(\overline{X})\big)$ is a diagonal matrix with all positive diagonal entries. It follows that the right-hand side of (3.68) is positive definite. Thus, we obtain a contradiction. Hence, $\widehat{\Lambda} \in \mathbb{S}_{++}^{n-r}$. Then, from Theorem 3.18, we obtain the rank consistency.

For the density matrix completion, $\mathcal{P}_\alpha(\cdot) = \frac{1}{n}\operatorname{Tr}(\cdot)I_n$. By further using the fact that $\overline{P}_2^{\mathbb{T}}F(\overline{X})\overline{P}_2 = 0$ and $\mathcal{P}_\beta(I_n) = 0$, we can rewrite (3.68) as

$$\widehat{\Lambda} - \frac{1}{n}\operatorname{Tr}(\widehat{\Lambda})I_{n-r} = \frac{1}{n}\operatorname{Tr}(F(\overline{X}))I_{n-r}.$$

By taking the trace on both sides, we obtain that $\widehat{\Lambda} = \frac{1}{r}\mathrm{Tr}(F(\overline{X}))I_{n-r}$. Since $\overline{X}$ is a density matrix of rank r, with the choice (3.67) of F, we have that

$$\mathrm{Tr}(\overline{X}) = \sum_{i=1}^{n}\sum_{j=1}^{r}\lambda_j(\overline{X})|\overline{P}_{ij}|^2 = 1 \quad \Longrightarrow \quad \mathrm{Tr}\big(F(\overline{X})\big) = \sum_{i=1}^{n}\sum_{j=1}^{r} f_i\big(\lambda_j(\overline{X})\big)|\overline{P}_{ij}|^2 > 0.$$

It follows that $\widehat{\Lambda} \in \mathbb{S}_{++}^{n-r}$ and thus we obtain the rank consistency. □

3.4 Construction of the rank-correction function

In this section, we focus on the construction of a suitable rank-correction function F based on the results obtained in Sections 3.2 and 3.3. As can be seen from Theorem 3.6, a smaller value of a_m/b_m potentially leads to a smaller recovery error. Thus, we desire a construction of the rank-correction function such that $F(\widetilde{X}_m)$ is close to $\overline{U}_1\overline{V}_1^{\mathbb{T}}$. Meanwhile, according to Theorem 3.21, we also desire that $F(\overline{X})$ is close to $\overline{U}_1\overline{V}_1^{\mathbb{T}}$ for rank consistency. Notice that a reasonable initial estimator $\widetilde{X}_m$ should not deviate too much from the true matrix $\overline{X}$. Therefore, the above two criteria consistently suggest a natural idea to construct a rank-correction function F, if possible, such that

$$F(X) \to \overline{U}_1\overline{V}_1^{\mathbb{T}} \quad \text{as} \quad X \to \overline{X}. \tag{3.69}$$

Next, we proceed the construction of the rank-correction function F for the rectangular case. For the positive semidefinite case, one may just replace the singular value decomposition with the eigenvalue decomposition and conduct exactly the same analysis.

3.4.1 The rank is known

If the rank of the true matrix $\overline{X}$ is known in advance, we construct the rank-correction function F by

$$F(X) := U_1 V_1^{\mathbb{T}}, \tag{3.70}$$

where $(U, V) \in \mathbb{O}^{n_1,n_2}(X)$ and $X \in \mathbb{V}^{n_1\times n_2}$. Note that F defined by (3.70) is not a spectral operator over the whole space of $\mathbb{V}^{n_1\times n_2}$, but in a neighborhood of $\overline{X}$ it is indeed a spectral operator and is actually twice continuously differentiable (see, e.g., [34, Proposition 8]). Hence, it satisfies the criterion (3.69). With this rank-correction function, the rank-correction step is essentially the same as one step of the majorized penalty method developed in [62]. Then we obtain the following result.

Corollary 3.23. *Suppose that the rank of the true matrix $\overline{X}$ is known and the constraint nondegeneracy holds at $\overline{X}$. If $\rho_m \to 0$, $\sqrt{m}\rho_m \to \infty$, $\gamma_m = O_p(1)$ and F is chosen by (3.70), then the estimator $\widehat{X}_m$ generated from the rank-correction step is rank consistent.*

Proof. This is immediate from Theorem 3.21 since the choice (3.70) of the rank-correction function F yields $\widehat{g}(\overline{X}) = 0$. □

3.4.2 The rank is unknown

If the rank of the true matrix $\overline{X}$ is unknown, then the rank-correction function F cannot be defined by (3.70). What we will do is to construct a spectral operator F to imitate the case when the rank is known. Here, we propose F to be a spectral operator

$$F(X) := U\mathrm{Diag}\big(f(\sigma(X))\big)V^{\mathbb{T}} \tag{3.71}$$

associated with the symmetric function $f : \mathbb{R}^n \to \mathbb{R}^n$ defined by

$$f_i(x) = \begin{cases} \phi\left(\dfrac{x_i}{\|x\|_\infty}\right) & \text{if } x \in \mathbb{R}^n \backslash \{0\}, \\ 0 & \text{if } x = 0, \end{cases} \tag{3.72}$$

where $(U, V) \in \mathbb{O}^{n_1,n_2}(X)$, $X \in \mathbb{V}^{n_1 \times n_2}$, and the scalar function $\phi : \mathbb{R} \to \mathbb{R}$ takes the form

$$\phi(t) := \operatorname{sgn}(t)(1 + \varepsilon^\tau)\frac{|t|^\tau}{|t|^\tau + \varepsilon^\tau}, \quad t \in \mathbb{R}, \tag{3.73}$$

for some $\tau > 0$ and $\varepsilon > 0$. Then we have the following result.

Corollary 3.24. *Suppose that the constraint nondegeneracy holds at $\overline{X}$. If $\rho_m \to 0$, $\sqrt{m}\rho_m \to \infty$, $\gamma_m = O_p(1)$, then for any given $\tau > 0$, there exists some $\overline{\varepsilon} > 0$ such that for any F defined by (3.71), (3.72) and (3.73) with $0 < \varepsilon \leq \overline{\varepsilon}$, the estimator $\widehat{X}_m$ generated from the rank-correction step is rank consistent.*

Proof. Note that for each t, $\phi(t) \to \operatorname{sgn}(t)$ as $\varepsilon \downarrow 0$. It implies that $f_i(\overline{X}) \to 1, \forall\, 1 \leq i \leq r$ and thus $\widehat{g}(\overline{X}) \to 0$ as $\varepsilon \downarrow 0$. Then the result is immediate from Theorem 3.21. □

Corollary 3.24 indicates that one needs to choose a small $\varepsilon > 0$ in pursuit of rank consistency. Meanwhile, we also need to take care of the influence of a small $\varepsilon > 0$ on the recovery error bound which depends on the value of a_m/b_m. Certainly, we desire $a_m \approx 0$ and $b_m \approx 1$. This motivates us to choose a function ϕ, if possible, such that

$$\phi\left(\frac{\sigma_i(\widetilde{X}_m)}{\sigma_1(\widetilde{X}_m)}\right) \approx \begin{cases} 1 & \text{if } 1 \leq i \leq r, \\ 0 & \text{if } r+1 \leq i \leq n. \end{cases} \tag{3.74}$$

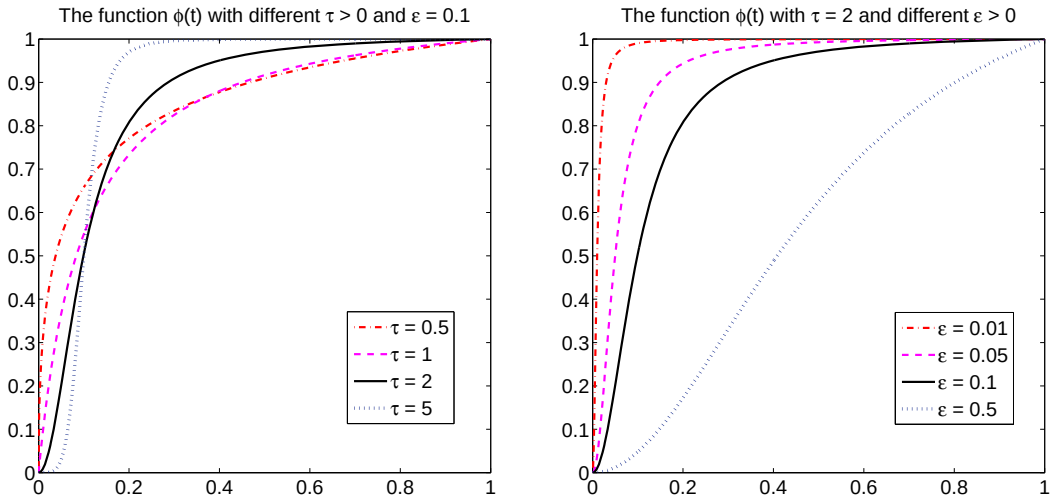
This is also why we normalize the function ϕ defined by (3.73) in the interval $t \in [0, 1]$ such that $\phi(0) = 0$ and $\phi(1) = 1$. However, as indicated by Lemma 3.9, the initial estimator $\widetilde{X}_m$ is very possible to have a higher rank than $\overline{X}$ when it

approaches to $\overline{X}$. It turns out that when $\varepsilon > 0$ is tiny,

$$\phi\left(\frac{\sigma_i(\widetilde{X}_m)}{\sigma_1(\widetilde{X}_m)}\right) \approx 1 \quad \forall\, r+1 \le i \le \operatorname{rank}(\widetilde{X}_m),$$

which violates our desired property (3.74). As a result, $\varepsilon > 0$ should be chosen to be small but balanced. Notice that $\phi(\varepsilon) = (1+\varepsilon^\tau)/2 \approx 1/2$ if $\varepsilon > 0$ is small and $\tau > 0$ is not too small. Thus, the value of ε can be regarded as a divide of confidence on whether $\sigma_i(\widetilde{X}_m)$ is believed to come from a nonzero singular values of $\overline{X}$ with perturbation — positive confidence if $\sigma_i(\widetilde{X}_m) > \varepsilon\sigma_1(\widetilde{X}_m)$ and negative confidence if $\sigma_i(\widetilde{X}_m) < \varepsilon\sigma_1(\widetilde{X}_m)$. On the other hand, the parameter $\tau > 0$ mainly controls the shape of the function ϕ over $t \in [0,1]$. The function ϕ is concave if $0 < \tau \le 1$ and S-shaped with a single inflection point at $\left(\frac{\tau-1}{\tau+1}\right)^{1/\tau}\varepsilon$ if $\tau > 1$. Moreover, the steepness of the function ϕ increases when τ increases. In particular, if $0 < \varepsilon < 1$ and τ is very large, ϕ is very close to the step function taking the value 0 if $0 \le t < \varepsilon$ and the value 1 if $\varepsilon < t \le 1$. In this case, there exists some ε such that the desired property (3.74) can be achieved and that the corresponding rank-correction function F is very close to the one defined by (3.70). Thus, it seems to be a good idea to choose an S-shaped function ϕ with a large τ. However, in practice, the parameter ε should be pre-determined. Since $\operatorname{rank}(\overline{X})$ is unknown and the singular values of $\widetilde{X}_m$ are unpredictable, it is hard to choose a suitable ε in advance, and hence, it will be too risky to choose a large τ for recovery. As a result, one has to be somewhat conservative to choose τ, sacrificing some optimality of recovery in exchange for robustness strategically. If the initial estimator is generated from the nuclear norm penalized least squares problem, we recommend the choices $\tau = 1$ or 2 and $\varepsilon = 0.01 \sim 0.1$ as these choices show stable performance for plenty of problems, as validated in Section 6.

We also remark that for the positive semidefinite case, the rank-correction function defined by (3.71), (3.72) and (3.73) is related to the reweighted trace norm



(a) $\varepsilon = 0.1$ with different $\tau > 0$ (b) $\tau = 2$ with different $\varepsilon > 0$

Figure 3.1: Shapes of the function ϕ with different $\tau > 0$ and $\varepsilon > 0$

for the matrix rank minimization proposed by Fazel et al. [51, 132]. The reweighted trace norm in [51, 132] for the positive semidefinite case is $\langle (X^k + \varepsilon I_n)^{-1}, X \rangle$, which arises from the derivative of the surrogate function $\log\det(X + \varepsilon I_n)$ of the rank function at an iterate X^k, where ε is a small positive constant. Meanwhile, in our proposed rank-correction step, if we choose $\tau = 1$, then $I_n - \frac{1}{1+\varepsilon} F(\widetilde{X}_m) = \varepsilon'(\widetilde{X}_m + \varepsilon' I_n)^{-1}$ with $\varepsilon' = \varepsilon \|\widetilde{X}_m\|$. Superficially, similarity occurs; however, it is notable that ε' depends on $\widetilde{X}_m$, which is different from the constant ε in [51, 132]. More broadly speaking, the rank-correction function F defined by (3.71), (3.72) and (3.73) is not a gradient of any real-valued function. This distinguishes our proposed rank-correction step from the reweighted trace norm minimization in [51, 132] even for the positive semidefinite case.

3.5 Numerical experiments

In this section, we validate the power of our proposed rank-corrected procedure on the recovery by applying it to different kinds of matrix completion problems. In solving the optimization problem in the rank-correction step (3.9) for the positive semidefinite matrix completion, we adopted the code developed by Jiang et al. [79] for large scale linearly constrained convex semidefinite programming problems. The implemented code is based on an inexact version of the accelerated proximal gradient method [136, 10]. We also modified this code to make it adaptive to the optimization problem in the rank-correction step (3.8) for the rectangular matrix completion. (The subproblem in each iteration was solved by a semismooth Newton-CG method.) All tests were run in MATLAB under Windows 7.0 operating system on an Intel Core(TM) i7-2720 QM 2.20GHz processor with 8.00GB RAM.

For convenience, in the sequel, the NNPLS estimator and the RCS estimator, respectively, stand for the estimators from the nuclear norm penalized least squares problem (3.7) (with additional constraint $X \in \mathbb{S}^n_+$ for the positive semidefinite matrix completion) and the rank-correction step (3.8) (the rank-correction step (3.9) for the positive semidefinite matrix completion). Let X_m be an estimator. The **relative error** (**relerr** for short) of X_m and the **relative deviation** (**reldev** for short) are defined, respectively, by

$$\text{relerr} := \frac{\|X_m - \overline{X}\|_F}{\max(10^{-8}, \|\overline{X}\|_F)} \quad \text{and} \quad \text{reldev} := \frac{\|y - \mathcal{R}_\Omega(\widetilde{X}_m)\|_2}{\max(10^{-8}, \|y\|_2)}.$$

3.5.1 Influence of fixed basis coefficients on the recovery

In this subsection, we take the correlation matrix completion for example to test the performance of the NNPLS estimator and the RCS estimator with different

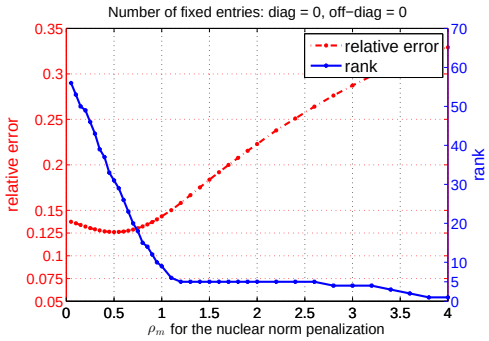
patterns of fixed basis coefficients. We randomly generated the true matrix $\overline{X}$ by the following command:

```
M = randn(n,r);  ML = weight*M(:,1:k);  M(:,1:k) = ML; Xtemp = M*M';
        D = diag(1./sqrt(diag(Xtemp))); X_bar = D*Gtemp*D
```

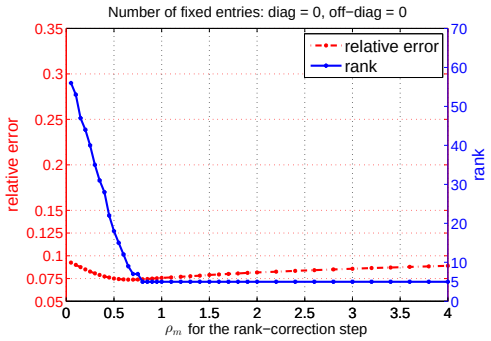
where the parameter `weight` is used to control the relative magnitude difference between the first `k` largest eigenvalues and the other nonzero eigenvalues. In our experiments, we set `weight` $= 5$ and `k` $= 1$, and took $\overline{X}$ = `X_bar` with dimension `n` $= 1000$ and rank `r` $= 5$. We randomly fixed partial diagonal and off-diagonal entries of $\overline{X}$ and sampled the rest entries uniformly at random with i.i.d. Gaussian noise at the noise level 10%.

In Figure 3.2, we plot the curves of the relative error and the rank of the NNPLS estimator and the RCS estimator with different patterns of fixed entries. In the captions of the subfigures, **diag** means the number of fixed diagonal entries and **non-diag** means the number of fixed off-diagonal entries. The subfigures on the left-hand side and the right-hand side show the performance of the NNPLS estimator and the RCS estimator, respectively. For the RCS estimator, the rank-correction function F is defined by (3.71), (3.72) and (3.73) with $\tau = 2$ and $\varepsilon = 0.02$, and the initial $\widetilde{X}_m$ is chosen from those points of the corresponding subfigures on the left-hand side such that the absolute difference between the relative derivation the the noise level attains the minimum.

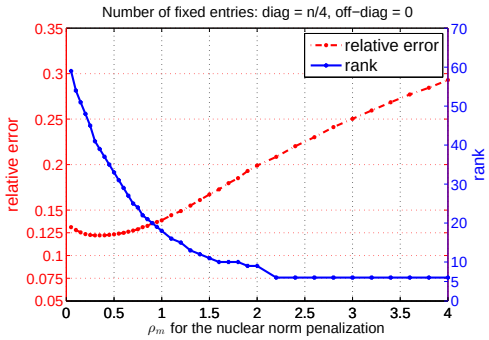
From the subfigures on the left-hand side, we observe that as the number of fixed diagonal entries increases, the parameter ρ_m for the smallest recovery error deviates more and more from the one for attaining the true rank. In particular, when **diag** $= n$, the NNPLS estimator reduces to the (constrained) least squares estimator so that one cannot benefit from the NNPLS estimator for encouraging a low-rank solution. This implies that the NNPLS estimator does not possess the



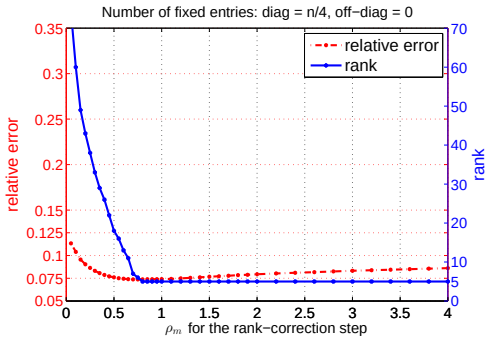
(a) NNPLS: diag $= 0$, off-diag $= 0$



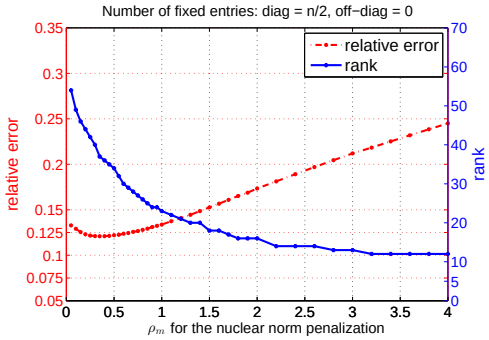
(b) RCS: diag $= 0$, off-diag $= 0$



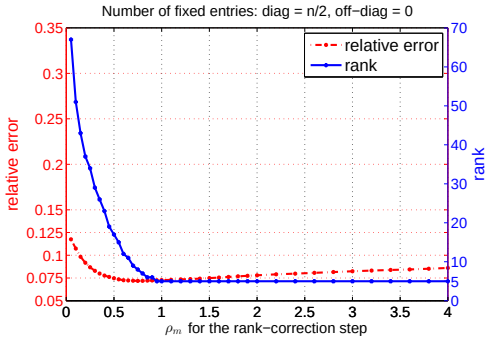
(c) NNPLS: diag $= n/4$, off-diag $= 0$



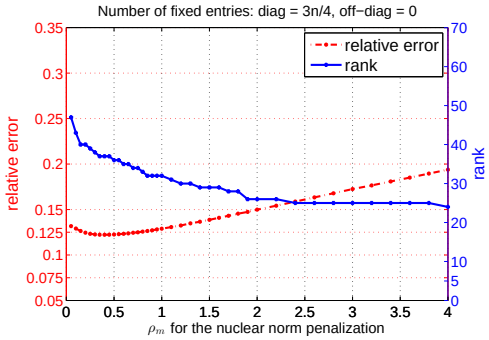
(d) RCS: diag $= n/4$, off-diag $= 0$



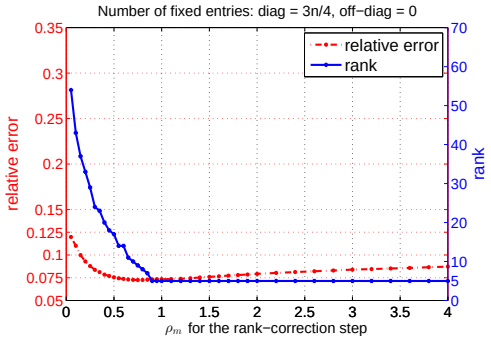
(e) NNPLS: diag $= n/2$, off-diag $= 0$



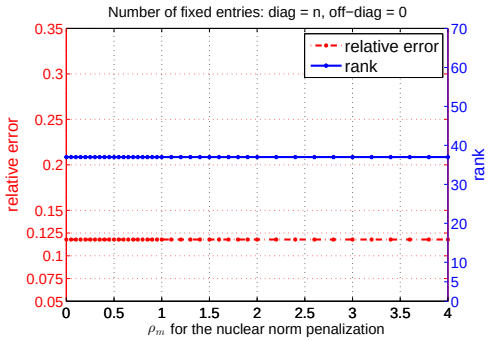
(f) RCS: diag $= n/2$, off-diag $= 0$



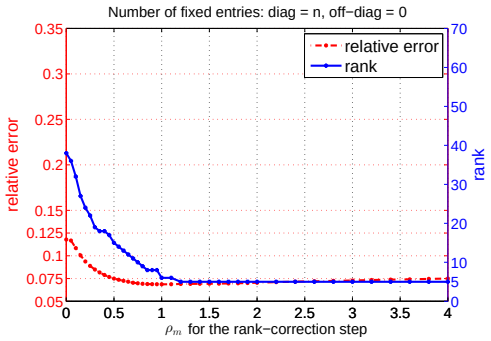
(g) NNPLS: diag $= 3n/4$, off-diag $= 0$



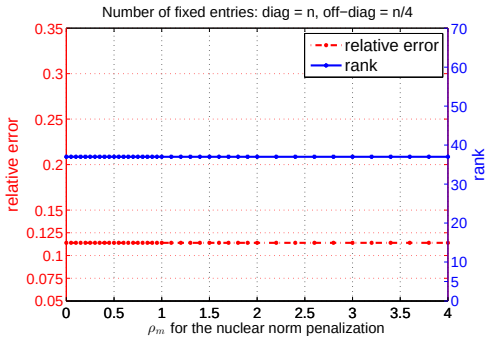
(h) RCS: diag $= 3n/4$, off-diag $= 0$



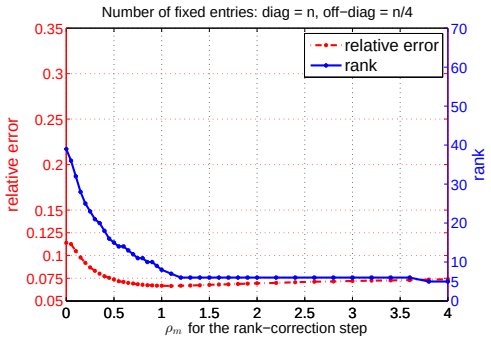
(q) NNPLS: diag $= n$, off-diag $= 0$



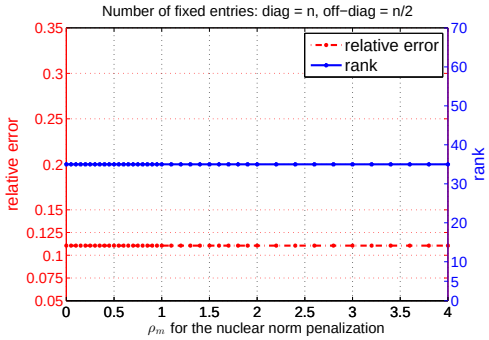
(r) RCS: diag $= n$, off-diag $= 0$



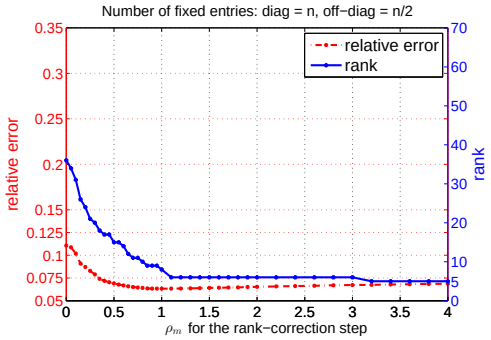
(s) NNPLS: diag $= n$, off-diag $= n/4$



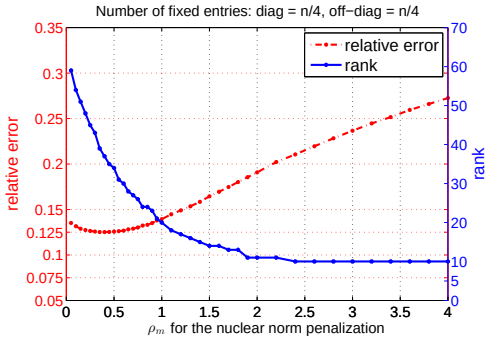
(t) RCS: diag $= n$, off-diag $= n/4$



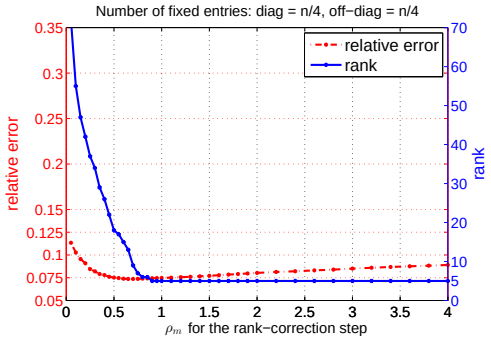
(u) NNPLS: diag $= n$, off-diag $= n/2$



(v) RCS: diag $= n$, off-diag $= n/2$



(w) NNPLS: diag $= n/4$, off-diag $= n/4$



(x) RCS: diag $= n/4$, off-diag $= n/4$

Figure 3.2: Influence of fixed basis coefficients on recovery (sample ratio = 6.38%)

rank consistency when some entries are fixed. However, the subfigures on the right-hand side indicate that the RCS estimator can yield a solution with the correct rank as well as a desired small recovery error simultaneously, with the parameter ρ_m in a large interval. This exactly validates the theoretical result of Theorem 3.22 for rank consistency.

3.5.2 Performance of different rank-correction functions for recovery

In this subsection, we test the performance of different rank-correction functions for recovering a correlation matrix. We randomly generated the true matrix $\overline{X}$ by the command in Subsection 6.1 with $\texttt{n} = 1000$, $\texttt{r} = 10$, $\texttt{weight} = 2$ and $\texttt{k} = 5$. We fixed all the diagonal entries of $\overline{X}$ and sampled partial off-diagonal entries uniformly at random with i.i.d. Gaussian noise at the noise level 10%. We chose the (nuclear norm penalized) least squares estimator to be the initial estimator $\widetilde{X}_m$. In Figure 3.3, we plot four curves corresponding to the rank-correction functions F defined by (3.71), (3.72) and (3.73) with $\tau = 2$ and different ε, and another two curves corresponding to the rank-correction functions F defined by (3.70) at $\widetilde{X}_m$ (i.e., $\widetilde{U}_1\widetilde{V}_1^{\mathbb{T}}$) and $\overline{X}$ (i.e., $\overline{U}_1\overline{V}_1^{\mathbb{T}}$), respectively. The values of a_m, b_m and the optimal recovery error with different ρ_m are listed in Table 3.1.

As can be seen from Figure 3.3, when ρ_m increases, the recovery error decreases with the rank and then increases after the correct rank is attained, except for the case $\overline{U}_1\overline{V}_1^{\mathbb{T}}$. This validates our discussion about the recovery error at the end of Section 3.2. Moreover, for a smaller ε, the curve of recovery error changes more gently, though a certain optimality in the sense of recovery error is sacrificed. This means that the choice of a relatively small ε, say 0.01 or 0.02, is more robust for those ill-conditioned problems. From Table 3.1, we see that a smaller a_m/b_m

corresponds to a better optimal recovery error. It is worthwhile to point out that, even if a_m/b_m is larger than 1, the performance of the RCS estimator for recovery is still much better than that of the NNPLS estimator.

Table 3.1: Influence of the rank-correction term on the recovery error

rank-correction function	a_m	b_m	a_m/b_m	optimal relerr
zero function	1	1	1	10.85%
$\varepsilon = 0.01, \tau = 2$	0.1420	0.2351	0.6038	5.96%
$\varepsilon = 0.02, \tau = 2$	0.1459	0.5514	0.2646	5.80%
$\varepsilon = 0.05, \tau = 2$	0.1648	0.8846	0.1863	5.75%
$\varepsilon = 0.1, \tau = 2$	0.2399	0.9681	0.2478	5.77%
$\widetilde{U}_1\widetilde{V}_1^{\mathbb{T}}$ (initial)	0.1445	0.9815	0.1472	5.75%
$\overline{U}_1\overline{V}_1^{\mathbb{T}}$ (true)	0	1	0	2.25%

3.5.3 Performance for different matrix completion problems

In this subsection, we test the performance of the RCS estimator for the covariance and density matrix completion problems. As can be seen from Figure 3.2, a good choice of the parameter ρ_m for the RCS estimator could be the smallest one such that the rank becomes stable. Such a parameter ρ_m can be found by the bisection search method. This is actually what we benefit from rank consistency. In the following numerical experiments, we apply the above strategy to find a suitable ρ_m for the RCS estimator, and choose the rank-correction function F defined by (3.71), (3.72) and (3.73) with $\tau = 2$ and $\varepsilon = 0.02$. However, it is difficult to

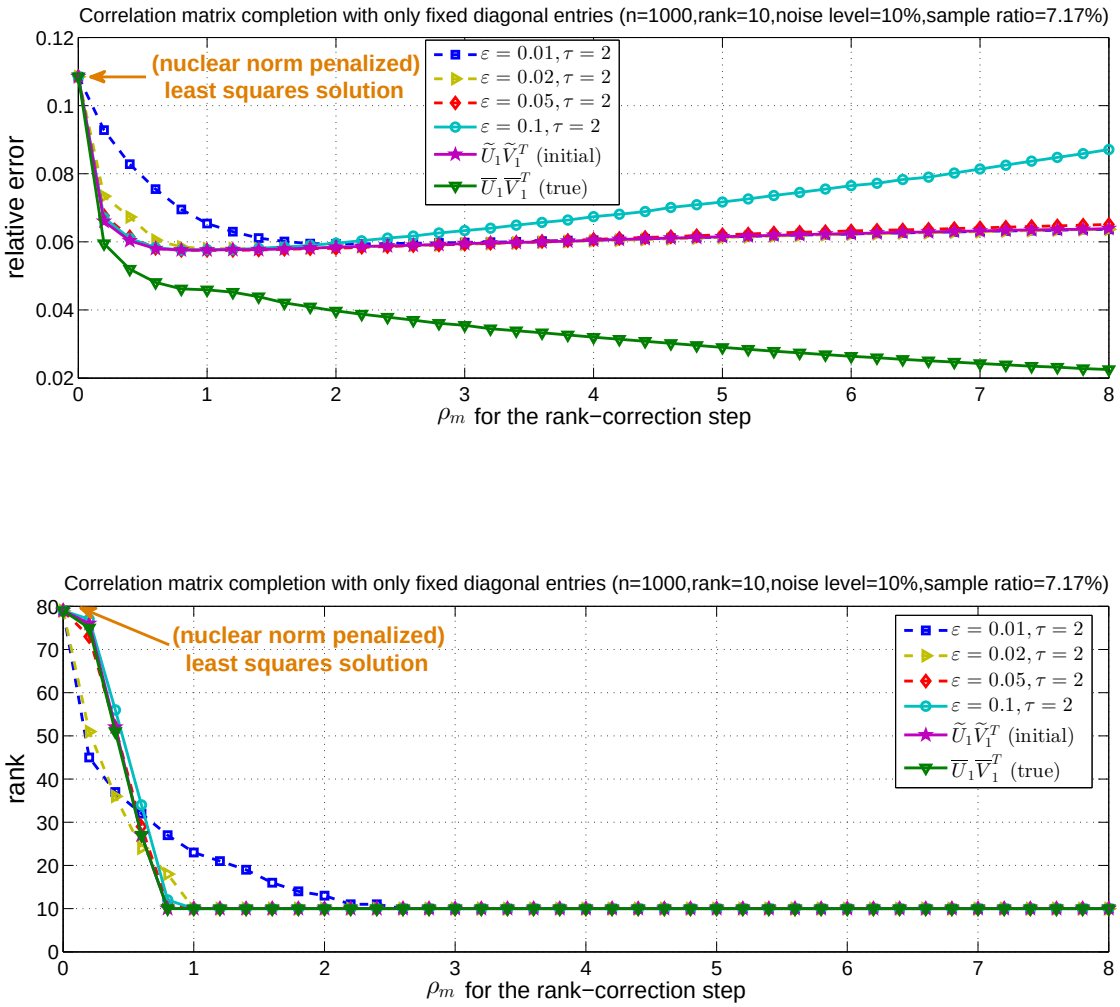


Figure 3.3: Influence of the rank-correction term on the recovery

choose a good parameter ρ_m for the NNPLS estimator according the behavior of the rank. For comparison, if not specified, the parameter ρ_m for the NNPLS estimator is chosen to be the one such that the absolute difference between the relative deviation and the noise level falls attains the minimum. This choice is reasonable and leads to a relatively smaller recover error compared with others in general. (In our experiments, we actually gave priority to choosing ρ_m to be the smallest one such that the rank becomes stable before the relative error is beyond 150% noise level. But this case never happened.)

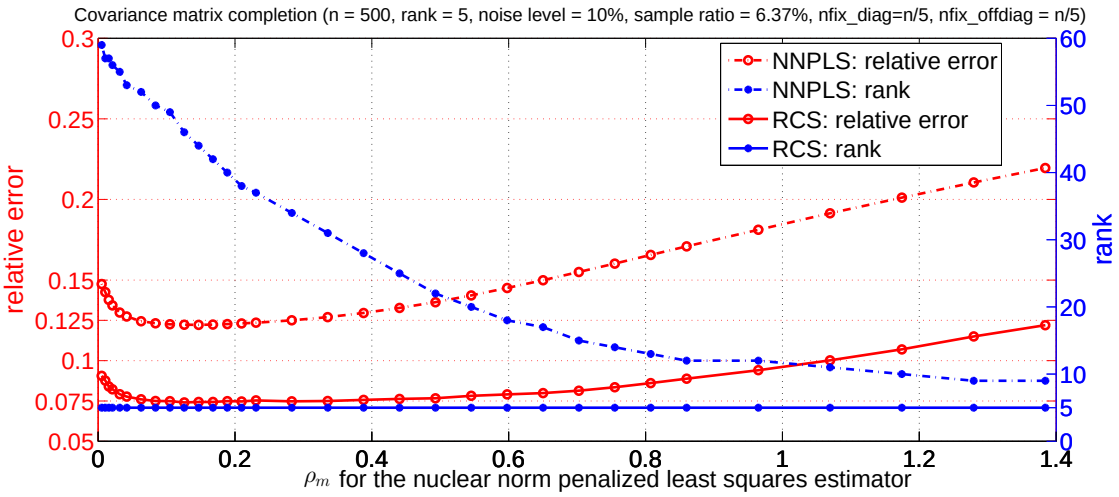


Figure 3.4: Performance of the RCS estimator with different initial $\widetilde{X}_m$

We first take the covariance matrix completion for example to test the performance of the RCS estimator with different initial estimators $\widetilde{X}_m$. The true matrix $\overline{X}$ is generated by the command in Subsection 6.1 with `n` $= 500$, `r` $= 5$, `weight` $= 3$ and `k` $= 1$ except that `D = eye(n)`. We depict the numerical results in Figure 3.4, where the dash curves represent the relative recovery error and the rank of the NNPLS estimator with different ρ_m, and the solid curves represent the relative recovery error and the rank of the RCS estimator with $\widetilde{X}_m$ chosen to be the corresponding NNPLS estimator. As can be seen from Figure 3.4, the RCS estimator

substantially improves the quality of the NNPLS estimator in terms of both the recovery error and the rank. We also observe that when the initial $\widetilde{X}_m$ has a large deviation from the true matrix, the quality of the RCS estimator may still not be satisfied. Thus, it is natural to ask whether further rank-correction steps could improve the quality. The answer can be found from Tables 3.2, 3.3 and 3.4 below, where the numerical results of the covariance matrix completion, density matrix completion and rectangular matrix completion are reported respectively.

For the covariance matrix completion problems, we generated the true matrix $\overline{X}$ by the command in Subsection 3.5.1 with `n` $= 1000$, `weight` $= 3$ and `k` $= 1$ except that `D = eye(n)`. We fixed partial diagonal and upper off-diagonal entries and then sampled partial entries uniformly at random from the rest diagonal and upper off-diagonal entries with i.i.d. Gaussian noise at the noise level 10%. The rank of $\overline{X}$ and the number of fixed diagonal and upper non-diagonal entries of $\overline{X}$ are reported in the first and the second columns of Table 3.2, respectively. The third column reports the sample ratio, which was calculated excluding the number of fixed entries. The first RCS estimator is using the NNPLS estimator as the initial estimator $\widetilde{X}_m$, and the second (third) RCS estimator is using the first (second) RCS estimator as the initial estimator $\widetilde{X}_m$. From Table 3.2, we see that when the sample ratio is reasonable, one rank-correction step is enough to yield a desired result. Meanwhile, when the sample ratio is very low, especially if some off-diagonal entries are further fixed, one or two more rank-correction steps can still improve the quality of estimation. We also remark that the NNPLS estimator with a larger ρ_m could return a matrix of rank lower than what we reported in Table 3.2. However, correspondingly, the recover error will greatly increase.

For the density matrix completion problems, we generated the true density matrix $\overline{X}$ by the following command:

```
M = randn(n,r)+i*randn(n,r);  ML = weight*M(:,1:k);  M(:,1:k) = ML;
```

Table 3.2: Performance for covariance matrix completion problems with $n = 1000$

r	diag/ off–diag	sample ratio	NNPLS	1st RCS	2st RCS	3rd RCS
			relerr (rank)	relerr(rank)	relerr (rank)	relerr (rank)
5	1000/0	2.40%	1.95e-1 (47)	1.27e-1 (5)	1.18e-1 (5)	1.12e-1 (5)
	1000/0	7.99%	6.10e-2 (51)	3.41e-2 (5)	3.37e-2 (5)	3.36e-2 (5)
	500/50	2.39%	2.01e-1 (45)	1.10e-1 (5)	9.47e-2 (5)	8.97e-2 (5)
	500/50	7.98%	7.19e-2 (32)	3.77e-2 (5)	3.59e-2 (5)	3.58e-2 (5)
10	1000/0	5.38%	1.32e-1 (74)	7.68e-2 (10)	7.39e-2 (10)	7.36e-2 (10)
	1000/0	8.96%	9.18e-2 (78)	5.15e-2 (10)	5.08e-2 (10)	5.08e-2 (10)
	500/100	5.37%	1.58e-1 (57)	8.66e-2 (10)	7.74e-2 (10)	7.60e-2 (10)
	500/100	8.96%	1.02e-1 (49)	5.36e-2 (10)	5.24e-2 (10)	5.25e-2 (10)

Table 3.3: Performance for density matrix completion problems with $n = 1024$

noise	r	noise level	sample ratio	NNPLS1			NNPLS2			RCS		
				fidelity	relerr	rank	fidelity	relerr	rank	fidelity	relerr	rank
statistical	3	10.0%	1.5%	0.697	2.59e-1	3	0.955	2.50e-1	3	0.987	1.02e-1	3
		10.0%	4.0%	0.915	8.04e-2	3	0.997	6.84e-2	3	0.998	4.13e-2	3
	5	10.0%	2.0%	0.550	3.71e-1	5	0.908	4.23e-1	5	0.972	1.61e-1	5
		10.0%	5.0%	0.889	1.03e-1	5	0.995	9.18e-2	5	0.997	4.91e-2	5
mixed	3	12.4%	1.5%	0.654	2.93e-1	3	0.957	2.43e-1	3	0.988	1.06e-1	3
		12.4%	4.0%	0.832	1.49e-1	3	0.995	8.14e-2	3	0.997	6.41e-2	3
	5	12.4%	2.0%	0.521	3.95e-1	5	0.912	4.09e-1	5	0.977	1.51e-1	5
		12.5%	5.0%	0.817	1.61e-1	5	0.987	1.01e-1	5	0.996	7.09e-2	5

```
Xtemp = M*M';  X_bar = Xtemp/sum(diag((Xtemp))).
```

During the testing, we set `n` = 1024, `weight` = 2 and `k` = 1, and sampled partial Pauli measurements except the trace of $\overline{X}$ uniformly at random with i.i.d. Gaussian noise at the noise level 10%. Besides the above statistical noise, we further added the depolarizing noise, which frequently appears in quantum systems, with strength 0.01. This case is labeled as the mixed noise in the last four rows of Table 3.3. We remark here that the depolarizing noise differs from our assumption on noise since it does not have randomness. One may refer to [71, 53] for details of the quantum depolarizing channel. In Table 3.3, the (squared) **fidelity** is a measure of the closeness of two quantum states, defined by $\left\|\widehat{X}_m^{1/2}\overline{X}^{1/2}\right\|_*^2$, the NNPLS1 estimator means the NNPLS estimator by dropping the trace one constraint, and the NNPLS2 estimator means the one obtained by normalizing the NNPLS1 estimator to be of trace one. Note that the NNPLS2 estimator was ever used by Flammia et al. [53]. Table 3.3 shows that the RCS estimator is superior to the NNPLS2 estimator in terms of both the fidelity and the relative error.

For the rectangular matrix completion problems, we generated the true matrix $\overline{X}$ by the following command:

```
ML = randn(nr,r);   MR = randn(nc,r);   MW = weight*ML(:,1:k);
                ML(:,1:k) = MW;   X_bar = ML*MR'.
```

We set `weight` = 2, `k` = 1 and took $\overline{X}$ = `X_bar` with different dimensions and ranks. Both the uniform sampling scheme and the non-uniform sampling scheme were tested for comparison. For the non-uniform sampling scheme, the first 1/4 rows and the first 1/4 columns were sampled with probability 3 times than the other rows and columns respectively. In other words, the density of sampled entries in the top-left part is 3 times as much as that in the bottom-left part and the top-right part respectively and 9 times as much as that in the bottom-right part.

We add 10% i.i.d. Gaussian noise to the sampled entries. We also fixed partial entries of $\overline{X}$ uniformly from the rest un-sampled entries. The first RCS estimator is using the NNPLS estimator as the initial estimator $\widetilde{X}_m$, and the second (third) RCS estimator is using the first (second) RCS estimator as the initial estimator $\widetilde{X}_m$. What we observe from Table 3.4 for the rectangular matrix completion is similar to that for the covariance matric completion. Moreover, comparing the performances for the uniform and non-uniform sample schemes, we can see that the non-uniform sampling scheme greatly weakens the recoverability of the NNPLS estimator in terms of both the recovery error and the rank, especially when the sample ratio is low. Meanwhile, the advantage of the RCS estimators in these cases becomes more remarkable.

Table 3.4: Performance for rectangular matrix completion problems

setting	sample	fixed	sample ratio	NNPLS	1st RCS	2st RCS	3rd RCS
				relerr (rank)	relerr (rank)	relerr (rank)	relerr (rank)
$\dim = 1000 \times 1000, \text{rank} = 10$	uniform	0	5.97%	1.98e-1 (73)	7.67e-2 (10)	7.28e-2 (10)	7.27e-2 (10)
		0	11.9%	9.56e-2 (61)	4.47e-2 (10)	4.45e-2 (10)	4.45e-2 (10)
		1000	5.98%	1.98e-1 (75)	7.50e-2 (10)	7.10e-2 (10)	7.08e-2 (10)
		1000	12.0%	8.51e-2 (65)	4.37e-2 (10)	4.35e-2 (10)	4.35e-2 (10)
	non-uniform	0	5.97%	3.27e-1 (94)	1.28e-1 (22)	9.15e-2 (10)	8.66e-2 (10)
		0	11.9%	1.28e-1 (100)	5.19e-2 (10)	5.07e-2 (10)	5.07e-2 (10)
		1000	5.98%	3.21e-1 (90)	1.07e-1 (15)	8.81e-2 (10)	8.35e-2 (10)
		1000	12.0%	1.31e-1 (98)	5.11e-2 (10)	4.95e-2 (10)	4.95e-2 (10)
$\dim = 500 \times 1500, \text{rank} = 5$	uniform	0	4.32%	1.97e-1 (46)	7.86e-2 (5)	7.25e-2 (5)	7.17e-2 (5)
		0	7.98%	9.06e-2 (39)	4.61e-2 (5)	4.57e-2 (5)	4.57e-2 (5)
		1000	4.33%	1.96e-1 (47)	7.63e-2 (5)	6.96e-2 (5)	6.87e-2 (5)
		1000	7.99%	8.87e-2 (48)	4.37e-2 (5)	4.35e-2 (5)	4.35e-2 (5)
	non-uniform	0	4.32%	2.98e-1 (57)	1.97e-1 (5)	1.25e-1 (5)	1.07e-1 (5)
		0	7.98%	1.52e-1 (54)	5.98e-2 (5)	5.48e-2 (5)	5.43e-2 (5)
		1000	4.33%	2.92e-1 (60)	1.51e-1 (6)	1.08e-1 (5)	9.59e-2 (5)
		1000	7.99%	1.46e-1 (60)	6.65e-2 (5)	5.28e-2 (5)	5.16e-2 (5)

Chapter 4

Rank regularized problems with hard constraints

In this chapter, we address the rank regularized problem with hard constraints. The organization of this chapter is as follows: In Section 4.1, we introduce the rank regularized problem with hard constraints and approximate it by a nonconvex but continuous problem. In Section 4.2, we discuss the solution quality of the approximation problem for affine rank minimization problems and general rank regularized problems respectively. In Section 4.3, we propose an adaptive semi-nuclear norm regularization approach to address the rank regularized problem via solving its approximation problem, and also study the convergence of our proposed approach. Discussions of candidate functions used in this approach and comparisons with other existing algorithms can be found in Sections 4.4 and 4.5 respectively. Numerical results of different problems are reported in Section 4.6 to show that the iterative scheme of our proposed approach has advantages of achieving both the low-rank structure preserving ability and the computational efficiency.

4.1 Problem formulation

We consider the general rank regularized minimization problem expressed as

$$\begin{aligned} \min \quad & h(X) + \rho \operatorname{rank}(X) \\ \text{s.t.} \quad & X \in K, \end{aligned} \tag{4.1}$$

where $h : \mathbb{M}^{n_1 \times n_2} \to \mathbb{R}$ is a loss function assumed to be continuously differentiable, $\rho > 0$ is a regularization parameter and K is a nonempty closed convex subset of $\mathbb{M}^{n_1 \times n_2}$. This formulation includes the rank minimization problem — minimizing the rank over a convex set as a special case, provided that the loss function h vanishes. Many practical problems can be cast into the class of rank regularized problems, some of which have already been summarized in Fazel's PhD thesis [49]. For example, the loss function h is used to measure decision objectives such as the accuracy of a model or the cost of a design; while the rank function measures the order, the complexity or the dimensionality. Generally, the goal of the decision maker is a tradeoff between these desired objectives, reflected by the parameter ρ. We take the matrix completion problem discussed in Chapter 3 as an illustrative example. The goal of this problem is to recover a true low-rank matrix from a small number of its linear measurements. For achieving this goal, a tradeoff between two representable objectives — a small deviation from noisy observations and a low rank are considered. These two objectives, together with the structure that the unknown matrix needs to satisfy, make it possible to formulate the matrix completion problem in terms of (3.6), falling into the class of rank regularized problems (4.1).

Another two variants of the tradeoff mentioned above can be formulated as

$$\begin{aligned} \min \quad & \operatorname{rank}(X) \\ \text{s.t.} \quad & h(X) \leq t, \\ & X \in K, \end{aligned} \tag{4.2}$$

and

$$\begin{aligned} \min \quad & h(X) \\ \text{s.t.} \quad & \operatorname{rank}(X) \leq k, \\ & X \in K. \end{aligned} \tag{4.3}$$

Clearly, the formulation (4.2) can be absorbed in the formulation (4.1) if h is convex. Besides, the formulation (4.3) is closely related to (4.2) and thus (4.1). As has been argued in [49, Chapter 2], if the problem (4.2) can be solved, then the problem (4.3) can be solved via bisection as well.

Although the formulations (4.1) and (4.2) are equivalent if the parameters ρ and t are chosen correspondingly, the former one is computationally more favorable than the latter since generally handling one more penalized term in the objective function is easier than handling one more constraint. Penalization is a commonly-used technique to deal with constraints. However, against abuse of this technique, one needs to tell the difference between hard constraints and soft constraints. Hard constraints are those that must be satisfied (with high accuracy), e.g., as described in Chapter 3, the matrix is constrained to be positive semidefinite with diagonal entries being one for the correlation matrix completion and with trace being one for the density matrix completion. Soft constraints are those should be preferably satisfied but violations are allowed will the solution quality being possibly affected, e.g, the deviation from noisy observations in the matrix completion problem. One may prefer to reduce the number of soft constraints by using the penalization, especially for large-scale problems. But, such reduction cannot be applied to hard constraints in general. In this chapter, we are particular interested to address the rank regularized problem (4.1) with hard constraints.

As involving the rank function in the objective, the optimization problem (4.1) is NP-hard in general and computationally difficult to be solved in practice. A recent popular heuristic for solving (4.1) is to replace the rank function with the

nuclear norm as

$$\begin{aligned} \min \quad & h(X) + \rho \|X\|_* \\ \text{s.t.} \quad & X \in K. \end{aligned} \tag{4.4}$$

The problem (4.4) can be regarded as the best convex approximation of the problem (4.1). If one aims to achieve the same tradeoff of objectives as that in (4.1), the parameter $\rho > 0$ in (4.4) needs to be adjusted in general.

The convex relaxation (4.4) has gained great success in many applications. Its success is mainly due to the reason that the matrix having the smallest sum of singular values is very probable to have the smallest rank among all the matrices in the feasible set since the nuclear norm is the convex envelope of the rank function over the unit ball of spectral norm [49]. However, one cannot expect that the nuclear norm always works well as a surrogate of the rank function in any case. In particular, when the feasible set consists of correlation matrices or density matrices as considered in Chapter 3, the nuclear norm reduces to a constant and no longer contains the rank information so that one has to turn to other surrogates of the rank function for help.

The rank function can be alternatively represented in terms of singular values as

$$\operatorname{rank}(X) = \sum_{i=1}^{n} \mathbb{1}\big(\sigma_i(X)\big),$$

where $\mathbb{1}(\cdot)$ is the indicator function define over $\mathbb{R}_+$. Each nonzero singular values contributes equally to the rank function. Differently, each singular values contributes to the nuclear norm proportionally to its magnitude. Due to this difference, the nuclear norm is endowed with the favorable properties in computation — continuity and convexity. However, as the price to pay, the rank-promoting ability of the nuclear norm is somewhat weaker than that of the rank function. Therefore, it is inevitable to face the tradeoff between the computational difficulty and the rank-promoting ability. Nevertheless, one may think about sacrificing the

convexity in change of the improvement of rank-promoting ability by increasing the proportions of contributions of relatively-small singular values in the construction of a surrogate of the rank function, provided that the computational difficulty caused by the non-convexity can be well-handled. This realization motivates us to construct an approximate problem to deal with the tradeoff between minimizing the loss function and the rank function in terms of

$$\begin{aligned} \min \;\; & h(X) + \rho \|F(X)\|_* \\ \text{s.t.} \;\; & X \in K, \end{aligned} \tag{4.5}$$

where $F : \mathbb{M}^{n_1 \times n_2} \to \mathbb{M}^{n_1 \times n_2}$ is a Löwner's operator (see Definition 2.4) associated with a function $f : \mathbb{R}_+ \to \mathbb{R}_+$ which is pre-selected from the set of functions $\mathcal{C}(\mathbb{R}_+)$ defined by

$$\mathcal{C}(\mathbb{R}_+) := \big\{ f : \mathbb{R}_+ \to \mathbb{R}_+ \text{ is non-identical zero, concave with } f(0) = 0 \big\},$$

and $\rho > 0$ is the parameter to control the tradeoff. The function $\|F(X)\|_*$ is a surrogate of the rank function, expressed as

$$\|F(X)\|_* = \sum_{i=1}^{n} f\big(\sigma_i(X)\big) \quad \forall\, X \in \mathbb{M}^{n_1 \times n_2}.$$

In particular, $\|F(X)\|_*$ reduces to $\mathrm{rank}(X)$ if $f(\cdot) = \mathbb{1}(\cdot)$ and reduces to $\|X\|_*$ if $f(\cdot) = \mathrm{id}(\cdot)$. Given any $f \in \mathcal{C}(\mathbb{R}_+)$, the surrogate function $\|F(X)\|_*$ is nonconvex (and also nonconcave) except for the case $\|F(X)\|_* = \|X\|_*$. Moreover, it is easy to check that if $f, g \in \mathcal{C}(\mathbb{R}_+)$, then all their convex combinations belong to $\mathcal{C}(\mathbb{R}_+)$, as well as $g \circ f$ and $\min(f, g)$. Thanks to the concavity of $f \in \mathcal{C}(\mathbb{R}_+)$, the surrogate function $\|F(X)\|_*$ is expected to be endowed with better rank-promoting ability compared with the nuclear norm.

4.2 Approximation quality

In this section, we study the relationship between the solutions to the rank regularized problem (4.1) and its approximation (4.5) and evaluate the quality of the approximation.

4.2.1 Affine rank minimization problems

We first consider a special case of (4.1) — the affine rank minimization problem, taking the form

$$\begin{aligned} \min \quad & \operatorname{rank}(X) \\ \text{s.t.} \quad & \mathcal{A}(X) = b, \end{aligned} \tag{4.6}$$

where the linear map $\mathcal{A} : \mathbb{M}^{n_1 \times n_2} \to \mathbb{R}^l$ and the vector $b \in \mathbb{R}^l$ are given. This problem is of particular interest in many applications such as system control, matrix completion and image reconstruction, to name but a few. Let $f \in \mathcal{C}(\mathbb{R}_+)$ and F be the Löwner's operator associated with f. Then we can construct an approximation of the affine rank minimization problem (4.6), expressed as

$$\begin{aligned} \min \quad & \|F(X)\|_* \\ \text{s.t.} \quad & \mathcal{A}(X) = b. \end{aligned} \tag{4.7}$$

On account of the special structure of (4.7), we are interested in the question: What is the prior guarantee such that the affine rank minimization problem (4.6) and its approximation (4.7) produce the same optimal solution?

Let $f \in \mathcal{C}(\mathbb{R}_+)$ and let r be an integer with $1 \le r \le n$. For any $z \in \mathbb{R}^n_+$, define

$$\theta_{f,r}(z) := \begin{cases} \dfrac{\sum_{i=1}^r f(z_{[i]})}{\sum_{i=1}^n f(z_{[i]})} & \text{if } z \in \mathbb{R}^n_+ \backslash \{0\}, \\ 0, & \text{if } z = 0. \end{cases}$$

and for any linear operator $\mathcal{A} : \mathbb{M}^{n_1 \times n_2} \to \mathbb{R}^l$, we define

$$\Theta_{f,r}(\mathcal{A}) := \sup_{Z \in N(\mathcal{A})} \theta_{f,r}\big(\sigma(Z)\big),$$

where $N(\mathcal{A})$ denotes the nullspace of the linear operator $\mathcal{A}$, i.e.,

$$N(\mathcal{A}) = \{Z \in \mathbb{M}^{n_1 \times n_2} \mid \mathcal{A}(Z) = 0\}.$$

It is easy to see from the definition that

$$0 \leq \Theta_{f,r_1}(\mathcal{A}) \leq \Theta_{f,r_2}(\mathcal{A}) \leq 1 \quad \forall\, 1 \leq r_1 \leq r_2 \leq n.$$

The following result characterizes the uniqueness of the solution to the approximation problem (4.7).

Theorem 4.1. *Any matrix $\overline{X} \in \mathbb{M}^{n_1 \times n_2}$ of rank at most r is the unique solution to the problem (4.7) with $b := \mathcal{A}(\overline{X})$ if and only if*

$$\Theta_{f,r}(\mathcal{A}) < \frac{1}{2}. \tag{4.8}$$

In addition, in this case, $\overline{X}$ is also the unique solution to the problem (4.6).

Proof. Notice that $\Theta_{f,r}(\mathcal{A}) < \frac{1}{2}$ is equivalent to

$$N(\mathcal{A}) = \{0\} \quad \text{or} \quad \sum_{i=1}^{r} f\big(\sigma_i(Z)\big) < \sum_{i=r+1}^{n} f\big(\sigma_i(Z)\big) \quad \forall\, Z \in N(\mathcal{A})\backslash\{0\}.$$

Suppose that (4.8) holds and $N(\mathcal{A}) \neq \{0\}$. From Theorem 2.8, we have that for any matrix $\overline{X}$ of rank at most r and any $N(\mathcal{A}) \ni Z \neq 0$,

$$\begin{aligned}
\|F(\overline{X} + Z)\|_* &= \sum_{i=1}^{n} f\big(\sigma_i(\overline{X} + Z)\big) \\
&\geq \sum_{i=1}^{n} \big| f\big(\sigma_i(\overline{X})\big) - f\big(\sigma_i(Z)\big) \big| \\
&\geq \sum_{i=1}^{r} \Big(f\big(\sigma_i(\overline{X})\big) - f\big(\sigma_i(Z)\big) \Big) + \sum_{i=r+1}^{n} \Big(f\big(\sigma_i(Z)\big) - f\big(\sigma_i(\overline{X})\big) \Big) \\
&= \|F(\overline{X})\|_* - \sum_{i=1}^{r} f\big(\sigma_i(Z)\big) + \sum_{i=r+1}^{n} f\big(\sigma_i(Z)\big) > \|F(\overline{X})\|_*.
\end{aligned}$$

Hence, $\overline{X}$ is the unique optimal solution to the problem (4.7) if (4.8) holds.

Conversely, suppose that any matrix $\overline{X}$ of rank at most r is the unique solution to (4.7) with $b := \mathcal{A}(\overline{X})$. Assume that (4.8) does not hold. Then there exists some $N(\mathcal{A}) \ni \widehat{Z} \neq 0$ such that

$$\sum_{i=1}^{r} f\big(\sigma_i(\widehat{Z})\big) \geq \sum_{i=r+1}^{n} f\big(\sigma_i(\widehat{Z})\big). \tag{4.9}$$

Let $(\widehat{U}, \widehat{V}) \in \mathbb{O}^{n_1,n_2}(\widehat{Z})$ and let

$$\widehat{X} := \widehat{U}\mathrm{Diag}\big(-\sigma_1(\widehat{Z}), \ldots, -\sigma_r(\widehat{Z}), 0, \ldots, 0\big)\widehat{V}^{\mathbb{T}}.$$

Clearly, $\mathrm{rank}(\widehat{X}) \leq r$. Moreover, from (4.9), we obtain

$$\|F(\widehat{X}+\widehat{Z})\|_* = \sum_{i=1}^{n} f\big(\sigma_i(\widehat{X}+\widehat{Z})\big) = \sum_{i=r+1}^{n} f\big(\sigma_i(\widehat{Z})\big) \leq \sum_{i=1}^{r} f\big(\sigma_i(\widehat{Z})\big) = \|F(\widehat{X})\|_*.$$

This means that $\widehat{X}$ is not the unique solution to the problem (4.7) with $b := \mathcal{A}(\widehat{X})$, which leads to a contradiction. Thus, we complete the "if and only if" part.

Furthermore, assume that in this case the problem (4.6) has another optimal solution $\widetilde{X} \neq \overline{X}$. Then from the first part of this theorem, $\widetilde{X}$ is also the unique solution to (4.7), leading to a contradiction. Thus, $\overline{X}$ is also the unique solution to (4.6). □

Theorem 4.1 is an extension of Lemma 6 in [139] for strong recovery of low-rank matrices via the nuclear norm minimization. This result implies that if the nullspace property (4.8) holds, then any matrix of rank at most r can be recovered from its linear measurements via solving the optimization problem (4.7). A similar result has also been obtained in [140, 188] for the Schatten-q quasi-norm. By further applying Theorem 4.1 to the case that $f(\cdot) = \mathbb{1}(\cdot)$, we obtain an implicit recurrence relation, i.e., for any $f \in \mathcal{C}(\mathbb{R}_+)$,

$$\Theta_{f,r} < \frac{1}{2} \quad \Longrightarrow \quad \Theta_{\mathbb{1},r} < \frac{1}{2}, \tag{4.10}$$

where the later one is also equivalent to

$$N(\mathcal{A}) = 0 \quad \text{or} \quad \operatorname{rank}(Z) > 2r \ \ \forall\, Z \in N(\mathcal{A}). \tag{4.11}$$

Thus, the condition (4.11), which also implies $r < n/2$, is necessary for (4.8) to hold true. A stronger version of (4.10) can be achieved as stated below.

Theorem 4.2. *Let $f, g \in \mathcal{C}(\mathbb{R}_+)$. Then for any linear operator $\mathcal{A} : \mathbb{M}^{n_1 \times n_2} \to \mathbb{R}^l$ and any integer r with $1 \le r \le n$,*

$$\Theta_{\mathbb{1},r}(\mathcal{A}) \le \Theta_{g\circ f,r}(\mathcal{A}) \le \Theta_{f,r}(\mathcal{A}) \le \Theta_{\mathrm{id},r}(\mathcal{A}).$$

Proof. It suffices to show that

$$\theta_{\mathbb{1},r}(z) \le \theta_{g\circ f,r}(z) \le \theta_{f,r}(z) \le \theta_{\mathrm{id},r}(z) \quad \forall\, z \in \mathbb{R}^n_+ \backslash \{0\}.$$

We first prove the last inequality. Let $\mathbb{R}^n_+ \ni z \neq 0$ be arbitrary. Notice that $f \in \mathcal{C}(\mathbb{R}_+)$ implies that $f(x)/x$ is nonincreasing on $(0, \infty)$. Then for any integer $1 \le i \le k \le n-1$, we obtain that $z_{[i]} f(z_{[k+1]}) \ge z_{[k+1]} f(z_{[i]})$ since $z_{[i]} \ge z_{[k+1]}$. It follows that

$$\sum_{i=1}^{k} z_{[i]} f(z_{[k+1]}) \ge \sum_{i=1}^{k} z_{[k+1]} f(z_{[i]}) \quad \forall\, 1 \le k \le n-1.$$

From the above inequality, we further obtain that for any $1 \le k \le n-1$,

$$\frac{\sum_{i=1}^{k+1} f(z_{[i]})}{\sum_{i=1}^{k} f(z_{[i]})} = 1 + \frac{f(z_{[k+1]})}{\sum_{i=1}^{k} f(z_{[i]})} \ge 1 + \frac{z_{[k+1]}}{\sum_{i=1}^{k} z_{[i]}} = \frac{\sum_{i=1}^{k+1} z_{[i]}}{\sum_{i=1}^{k} z_{[i]}}.$$

This implies that $\left\{ \frac{\sum_{i=1}^{k} f(z_{[i]})}{\sum_{i=1}^{k} z_{[i]}} \right\}$ is nondecreasing and thus

$$\frac{\sum_{i=1}^{r} f(z_{[i]})}{\sum_{i=1}^{r} z_{[i]}} \le \frac{\sum_{i=1}^{n} f(z_{[i]})}{\sum_{i=1}^{n} z_{[i]}}.$$

Hence, $\theta_{f,r}(z) \le \theta_{\mathrm{id},r}(z)$. This relation further leads to

$$\theta_{g\circ f,r}(z) = \theta_{g,r}\big(f(z)\big) \le \theta_{\mathrm{id},r}\big(f(z)\big) = \theta_{f,r}(z),$$

where $f(z)$ denotes the vector $\big(f(z_1),\ldots,f(z_n)\big)^T$. Then, since $g\circ f\in\mathcal{C}(\mathbb{R}_+)$, we further obtain

$$\theta_{\mathbb{1},r}(z)=\theta_{\mathbb{1}\circ(g\circ f),r}(z)\leq\theta_{g\circ f,r}(z).$$

Thus we complete the proof. □

Theorem 4.2 is an extension of [69, Lemma 7] from the vector case to the matrix case. This result provides us the possibility to compare the preference of functions in $\mathcal{C}(\mathbb{R}_+)$ to construct a surrogate of the rank function. Combing Theorems 4.1 and 4.2 together, we realize that among all the surrogate functions $\|F(X)\|_*$ with $f\in\mathcal{C}(\mathbb{R}_+)$, theoretically, the rank function possesses the best recoverability (or rank-promoting ability), while the nuclear norm possesses the least. In addition, the "more concave" the function f is, the better recoverability the surrogate function $\|F(X)\|_*$ possesses. Here, "more concave" refers to a larger ratio of the increase speeds near zero to that away from zero.

4.2.2 Approximation in epi-convergence

For the general case, it is hard to derive similar results as in Section 4.2.1. Alternatively, we consider the gradual behavior of a sequence of approximation problems that approaches the rank regularized problem (4.1) in terms of their optimal solutions. For this purpose, we use the technique of epi-convergence introduced in Subsection 2.5. The sequential approximation problems take the form

$$\begin{aligned}\min\ \ & h(X)+\rho\,\|F^k(X)\|_*\\ \text{s.t.}\ \ & X\in K,\end{aligned}\tag{4.12}$$

where $F^k(X)$ are Löwner's operators associated with $f^k\in\mathcal{C}(\mathbb{R}_+)$.

Lemma 4.3. *If* $f^k(\cdot)\xrightarrow{e}\mathbb{1}(\cdot)$, *then* $\|F^k(\cdot)\|_*\xrightarrow{e}\mathrm{rank}(\cdot)$.

Proof. Given any matrix $\overline{X} \in \mathbb{M}^{n_1 \times n_2}$ and any sequence $X^k \to \overline{X}$, the (Lipschitiz) continuity of singular values leads to $\sigma_i(X^k) \to \sigma_i(\overline{X})\ \forall\, i = 1, \ldots, n$. Since $f^k(\cdot) \xrightarrow{e} \mathbb{1}(\cdot)$, from the definition of epi-convergence, we obtain that

$$\liminf_{k\to\infty} f^k\big(\sigma_i(X^k)\big) \geq \mathbb{1}\big(\sigma_i(\overline{X})\big) \quad \forall\, i = 1, \ldots, n.$$

This leads to

$$\begin{aligned}\liminf_{k\to\infty} \|F^k(X^k)\|_* = \liminf_{k\to\infty} \sum_{i=1}^n f^k\big(\sigma_i(X^k)\big) &\geq \sum_{i=1}^n \liminf_{k\to\infty} f^k\big(\sigma_i(X^k)\big) \\ &\geq \sum_{i=1}^n \mathbb{1}\big(\sigma_i(\overline{X})\big) = \operatorname{rank}(\overline{X}).\end{aligned}$$

Meanwhile, $f^k(\cdot) \xrightarrow{e} \mathbb{1}(\cdot)$ also implies that there exists a sequence $\mathbb{R}^n_+ \ni x^k \to \sigma(\overline{X})$ such that

$$\lim_{k\to\infty} f^k(x_i^k) = \mathbb{1}\big(\sigma_i(\overline{X})\big) \quad \forall\, i = 1, \ldots, n.$$

Let $(U, V) \in \mathbb{O}^{n_1,n_2}(\overline{X})$ and define $X^k := U\mathrm{Diag}(x^k)V^{\mathbb{T}}$. Then, $X^k \to \overline{X}$ and

$$\begin{aligned}\lim_{k\to\infty} \|F^k(X^k)\|_* = \lim_{k\to\infty} \sum_{i=1}^n f^k(x_i^k) &= \sum_{i=1}^n \lim_{k\to\infty} f^k(x_i^k) \\ &= \sum_{i=1}^n \lim_k \mathbb{1}\big(\sigma_i(\overline{X})\big) = \operatorname{rank}(\overline{X}).\end{aligned}$$

Thus, we complete the proof. □

Corollary 4.4. *If the sequence $\{f^k\}$ is nondecreasing (i.e., $f^{k+1} \geq f^k$) and pointwise converges to $\mathbb{1}(\cdot)$ over $\mathbb{R}_+$, then $\|F^k(\cdot)\|_* \xrightarrow{e} \operatorname{rank}(\cdot)$.*

Proof. This is immediate from Proposition 2.13, Lemma 4.3 and the lower semicontinuity of the functions f^k. □

Now we are in the position to obtain the following result.

Theorem 4.5. *Suppose that $\{f^k\}$ is eventually level-bounded or $K \subset \mathbb{M}^{n_1 \times n_2}$ is compact. For each k, let X^k be an optimal solution and ν^k be the optimal value to*

the approximation problem (4.12). Let $\overline{\nu}$ be the optimal value to the problem (4.1). If $f^k(\cdot) \xrightarrow{e} \mathbb{1}(\cdot)$, then $\nu^k \to \overline{\nu}$. Moreover, any cluster point of $\{X^k\}$ is an optimal solution to (4.1). In particular, if the problem (4.1) has a unique solution $\overline{X}$, then $X^k \to \overline{X}$.

Proof. This is immediate from Proposition 2.13, Theorem 2.14 and Lemma 4.3. □

This result sounds good since it provides us a possible way to solve the rank regularized problem sequentially. However, generally speaking, it is more of theoretical interest rather than for practical implementations. First, since the convergence for each individual nonconvex approximation problem is not guaranteed, one cannot expect to numerically obtain a sequence of solutions that converges to the optimal solution to the rank regularized problem in practice. Secondly, as we know, the closer two functions are, the more similarities they share with each other. Therefore, if the surrogate function approximates the rank function too aggressively, computational difficulties will become apparent and hard to be well-handled. As a result, a more practical way is to solve only one approximation problem that is suitably pre-determined on account of the tradeoff between theoretical advantage and numerical convenience, or to solve at most several approximation problems and choose the best solution among them. Despite the above discussion, one may still gain a further understanding of the quality of solution to the approximation problem (4.5) from the obtained results in this subsection.

4.3 An adaptive semi-nuclear norm regularization approach

In this section, we propose an adaptive semi-nuclear norm regularization approach to address the rank regularized problem via solving its nonconvex approximation

problem (4.5). In general, a basic approach for solving a nonconvex optimization problem is to sequentially solve a sequence of convex approximation optimization problems instead. This motivates us to apply the majorized proximal gradient method, i.e., Algorithm 2.1, to the approximation problem (4.5) by setting

$$\mathbb{X} := \mathbb{M}^{n_1\times n_2}, \quad h(x) = h(X), \quad g(x) = \delta_K(X), \quad p(x) = \|F(X)\|_*.$$

4.3.1 Algorithm description

As can be seen from Algorithm 2.1, a crucial step is to construct a sequence of suitable convex functions to majorize the nonconvex function $\|F(X)\|_*$. For this purpose, the Lipschitz continuity of the function $\|F(X)\|_*$ is necessary. Therefore, we restrict the function f to be selected from the set $\overline{\mathcal{C}}(\mathbb{R}_+) \subset \mathcal{C}(\mathbb{R}_+)$ defined as

$$\overline{\mathcal{C}}(\mathbb{R}_+) := \left\{ \begin{array}{r} f : \mathbb{R}_+ \to \mathbb{R}_+ \text{ is non-identical zero, concave} \\ \text{with } f(0) = 0 \text{ and } f'_+(0) < \infty \end{array} \right\}.$$

Then we easily have the following result.

Lemma 4.6. *Let F be the Löwner's operator associated with $f \in \overline{\mathcal{C}}(\mathbb{R}_+)$. Then the function $\|F(X)\|_*$ is Lipschitz continuous with constant at most $\sqrt{n}f'_+(0)$.*

Proof. Note that for any $f \in \overline{\mathcal{C}}(\mathbb{R}_+)$, we have $f(t) \le f'_+(0)\,t$ for any $t \ge 0$. Then for any $X, Y \in \mathbb{M}^{n_1\times n_2}$, according to the inequality (2.13) and the fact $\|X\|_* \le \sqrt{\mathrm{rank}(X)}\|X\|_F$, we obtain

$$\big|\|F(X)\|_* - \|F(Y)\|_*\big| \le \|F(X-Y)\|_* \le f'_+(0)\|X-Y\|_* \le \sqrt{n}f'_+(0)\|X-Y\|_F.$$

Thus, we complete the proof. □

Notice that the surrogate function $\|F(X)\|_*$ with $f \in \overline{\mathcal{C}}(\mathbb{R}_+)$ is actually concave over the positive semidefinite cone $\mathbb{S}^n_+$. Therefore, for the positive semidefinite

case $K \subset \mathbb{S}^n_+$, one can easily construct a majorization function of $\|F(X)\|_*$ by using its linearization. This fact was observed by Fazel, Hindi and Boyd in [51] (see also [49]), leading to the proposal of the reweighted trace minimization for positive semidefinite matrix rank minimization problems. However, for the general rectangular case, one has to make more efforts to construct a suitable convex majorization function of $\|F(X)\|_*$. In the following, we provide an efficient strategy for such consideration, extending the linear majorization for the positive semidefinte case to its variant for the rectangular case.

Clearly, the concavity of f implies that

$$0 \le f'_-(t_1) \le f'_+(t) \le f'_-(t) \le f'_+(t_2) \quad \forall\, 0 \le t_2 \le t \le t_1,$$

where f'_- and f'_+ denote the left derivative and the right derivative of f respectively. Moreover, for any $\bar{t} \in \mathbb{R}_+$, we have a global linear overestimate of f over $\mathbb{R}_+$ as

$$f(t) \le f(\bar{t}) + s(t - \bar{t}) \quad \forall\, t \in \mathbb{R}_+,$$

where $s \in [f'_+(\bar{t}), f'_-(\bar{t})]$. Hereafter, we set $f'_-(0) := f'_+(0)$. Then given any matrix $Y \in \mathbb{M}^{n_1 \times n_2}$, one may construct a vector $w \in \mathbb{R}^n_+$ such that

$$w = \big(1 - s_1/\mu, \ldots, 1 - s_n/\mu\big)^T$$

with $s_i \in \big[f'_+\big(\sigma_i(Y)\big), f'_-\big(\sigma_i(Y)\big)\big]$ and $\mu \ge s_n$. Basically, there are two simple options: $\mu := s_n$ and $\mu := f'_+(0)$. (Indeed, if Y is not of full rank, these two options lead to the same value.) In addition,

$$0 \le s_1 \le \cdots \le s_n \le f'_+(0) \quad \Longrightarrow \quad 1 \ge w_1 \ge \cdots \ge w_n \ge 0.$$

Then, we have the first step majorization as

$$\begin{aligned} \|F(X)\|_* &\le \sum_{i=1}^n \Big(\big(f(\sigma_i(Y)\big) + s_i\big(\sigma_i(X) - \sigma_i(Y)\big)\Big) \\ &= \sum_{i=1}^n f\big(\sigma_i(Y)\big) + \mu \sum_{i=1}^n (1 - w_i)\sigma_i(X) - \mu \sum_{i=1}^n (1 - w_i)\sigma_i(Y) \\ &= \mu\big(\|X\|_* - \|X\|_w\big) + \|F(Y)\|_* - \mu\big(\|Y\|_* - \|Y\|_w\big) \end{aligned} \tag{4.13}$$

where $\|\cdot\|_w$ denotes the w-weighted norm, i.e., $\|X\|_w = \sum_{i=1}^n w_i\sigma_i(X)\ \forall X \in \mathbb{M}^{n_1\times n_2}$ (see the definition in Section 2.3). Notice that the right-hand side of the above inequality (4.13) is a difference of two convex functions. Therefore, the second step majorization is to further linearize the second convex function $\|\cdot\|_w$ as

$$\|X\|_w \geq \|Y\|_w + \langle G, X - Y\rangle \quad \forall\, X \in \mathbb{M}^{n_1\times n_2}, \tag{4.14}$$

where G is any element in the subdifferential of $\|\cdot\|_w$ at Y, whose characterization can be found in Theorem 2.5. A particular choice of $G \in \partial\|Y\|_w$ is

$$G := U\mathrm{Diag}(w)V^{\mathbb{T}}, \tag{4.15}$$

where $(U, V) \in \mathbb{O}^{n_1,n_2}(Y)$. For notational simplicity, we define

$$\widehat{p}(X, Y) := \mu\big(\|X\|_* - \langle G, X - Y\rangle\big) + \|F(Y)\|_* - \mu\|Y\|_*, \tag{4.16}$$

though abuse of notation may occur for $\widehat{p}(X, Y)$ due to the possible freedom for choosing μ and G. By substituting (4.14) into (4.13), we obtain a convex majorization of $\|F(X)\|_*$ as

$$\|F(X)\|_* = \widehat{p}(X, X) \quad \text{and} \quad \|F(X)\|_* \leq \widehat{p}(X, Y) \quad \forall\, X, Y \in \mathbb{M}^{n_1\times n_2}.$$

One may notice that for the positive semidefinite case, this majorization is nothing but the linearization of $\|F(X)\|_*$ at Y.

Based on the above construction of the majorization function, now we describe the basic framework of the adaptive semi-nuclear norm regularization approach.

Algorithm 4.1. (Adaptive semi-nuclear norm regularization approach)

Step 0. *Choose* $f \in \overline{\mathcal{C}}(\mathbb{R}_+)$. *Input* $X^0 \in \mathbb{M}^{n_1\times n_2}$. *Set* $k := 0$.

Step 1. *Compute* $\sigma(X^k)$. *Construct* $\mathbb{R}^n_+ \ni w^k = (1 - s^k_1/\mu^k, \ldots 1 - s^k_n/\mu^k)^T$ *with* $s^k_i \in [f'_+(\sigma_i(X^k), f'_-(\sigma_i(X^k))]$ *and* $\mu^k \geq s^k_n$. *Choose* G^k *to be an element in the subdifferential of* $\|\cdot\|_{w^k}$ *at* X^k. *Construct a convex function* h^k *that majorizes* h *at* X^k *over* K.

Step 2. *Compute the optimal solution* X^{k+1} *to the convex problem:*

$$\begin{array}{ll} \min & h^k(X) + \rho\mu^k\big(\|X\|_* - \langle G^k, X\rangle\big) \\ \text{s.t.} & X \in K. \end{array} \tag{4.17}$$

Step 3. *If converged, stop; otherwise, set* $k := k + 1$ *and go to* **Step 1**.

Notice that for each $k \geq 0$, $w^k_i \in [0, 1]$, $i = 1, \ldots, n$. Then, from the characterization of G^k in Theorem 2.5, we have $\sigma_i(G^k) \in [0, 1]$, $i = 1, \ldots, n$. This leads to

$$\|X\|_* - \langle G^k, X\rangle \geq \|X\|_* - \sum_{i=1}^{n} \sigma_i(G^k)\sigma_i(X) \geq 0 \quad \forall\, X \in \mathbb{M}^{n_1\times n_2}, \tag{4.18}$$

where the first inequality follows from von Neumann's inequality [174], i.e.,

$$\langle X, Y\rangle \leq \sum_{i=1}^{n} \sigma_i(X)\sigma_i(Y) \quad \forall\, X, Y \in \mathbb{M}^{n_1\times n_2}.$$

The nonnegative property (4.18) implies that for each $k \geq 0$, $\|X\|_* - \langle G^k, X\rangle$ actually defines a semi-norm, called a semi-nuclear norm (associated with G^k) hereafter, with its form depending on the current iterate X^k. This is the reason why we call Algorithm 4.2 the adaptive semi-nuclear norm regularization approach. Thanks to the regularization of a sequence of semi-nuclear norms, the iterative scheme preserves the low-rank structure of the solution in each iteration.

It is recommended to choose the initial point $X^0 = 0$ in the implementation of Algorithm 4.1. Under this setting, $G^0 = 0$ and thus the first iteration solves a nuclear norm regularized problem. It is well-known that the nuclear norm regularization is able to provide a reasonable low-rank solution in many cases. Therefore, the initial input $X^0 = 0$ allows the algorithm to have a probable good start. When the loss function h is convex, one may simply choose $h^k \equiv h$ in the implementation of Algorithm 4.1, provided that the subproblem can be efficiently solved by certain methodologies. In fact, the adaptive semi-nuclear norm regularization approach can be regarded as a sequential corrections of the nuclear norm regularization for achieving better performance.

As we discussed in Subsection 2.6, the efficiency of a majorization method depends on the quality of the constructed majorization functions. For general cases, majorizing the function h by using a quadratic function could be a straightforward choice. However, the construction of a suitable majorization function may not be easy or even possible in practice. Instead, line search could help to ensure the decrease of objective values. Even when majorization functions are available, line search can also help to reduce the deviation from the original function. Hence, we recommend a variant of Algorithm 4.1 with linear search as follows.

Algorithm 4.2. (Adaptive semi-nuclear norm regularization approach with line search)

Step 0. *Choose* $f \in \overline{\mathcal{C}}(\mathbb{R}_+)$. *Input* $X^0 \in \mathbb{M}^{n_1\times n_2}$. *Choose* $\tau \in (0,1)$ *and* $\delta \in (0,1)$. *Set* $k := 0$.

Step 1. *Compute* $\sigma(X^k)$. *Construct* $\mathbb{R}^n_+ \ni w^k = (1 - s^k_1/\mu^k, \ldots 1 - s^k_n/\mu^k)^T$ *with* $s^k_i \in [f'_+(\sigma_i(X^k), f'_-(\sigma_i(X^k))]$ *and* $\mu^k \geq s^k_n$. *Choose* G^k *to be an element in the subdifferential of* $\|\cdot\|_{w^k}$ *at* X^k. *Choose* $\gamma^k > 0$.

Step 2. *Compute the optimal solution* $\widetilde{X}^{k+1}$ *to the (strongly) convex problem:*

$$\begin{array}{ll} \min & \langle \nabla h(X^k), X\rangle + \dfrac{\gamma^k}{2}\|X - X^k\|^2_F + \rho\mu^k\big(\|X\|_* - \langle G^k, X\rangle\big) \\ \text{s.t.} & X \in K. \end{array} \tag{4.19}$$

Step 3. *Choose* $\overline{\alpha}^k > 0$ *and let* l_k *be the smallest nonnegative integer satisfying*

$$\phi\big(X^k + \overline{\alpha}^k\tau^l(\widetilde{X}^{k+1} - X^k)\big) \leq \phi(X^k) + \delta\overline{\alpha}^k\tau^l\Delta^k, \tag{4.20}$$

where $\phi(X) := h(X) + \rho\,\|F(X)\|_*$ *and*

$$\Delta^k := \langle \nabla h(X^k), \widetilde{X}^{k+1} - X^k\rangle + \rho\mu^k\big(\|\widetilde{X}^{k+1}\|_* - \|X^k\|_* - \langle G^k, \widetilde{X}^{k+1} - X^k\rangle\big).$$

Set $\alpha^k := \overline{\alpha}^k\tau^{l_k}$ *and* $X^{k+1} := X^k + \alpha^k(\widetilde{X}^{k+1} - X^k)$.

Step 4. *If converged, stop; otherwise, set* $k := k + 1$ *and go to* **Step 1**.

Algorithm 4.2 can be regarded as a direct application of the majorized proximal gradient method discussed in Section 2.6 to solve the approximation problem (4.5) with $f \in \overline{\mathcal{C}}(\mathbb{R}_+)$. In particular, when the loss function f is convex and ∇h is Lipschitz continuous with constant κ, Algorithm 4.2 can be simplified without line search by setting $\gamma^k := \kappa$ for all $k \geq 0$ since the objective function of (4.19)

reduces to a regular majorization.

4.3.2 Convergence results

In this subsections, we aim to discuss the convergence of the adaptive semi-nuclear norm regularization approach with line search. For convergence analysis, we need the following concepts. For any $\overline{X} \in K$, let $N_K(\overline{X})$ denote the normal cone of K at $\overline{X}$, i.e.,

$$N_K(\overline{X}) := \{Z \in \mathbb{M}^{n_1 \times n_2} \mid \langle Z, X - \overline{X} \rangle \leq 0 \ \forall\, X \in K\}.$$

We denote by $\partial \|F(\overline{X})\|_*$ the Clarke's generalized gradient of $\|F(X)\|_*$ at $\overline{X}$.

Definition 4.1. *Let $f \in \overline{\mathcal{C}}(\mathbb{R}_+)$. We say that $\overline{X} \in K$ is a stationary point of the problem (4.5) if*

$$0 \in \nabla h(\overline{X}) + \rho\, \partial \|F(\overline{X})\|_* + N_K(\overline{X}). \tag{4.21}$$

Recall that for any $f \in \overline{\mathcal{C}}(\mathbb{R}_+)$, $\|F(X)\|_*$ is Lipschitz continuous and thus $h(X) + \rho\, \|F(X)\|_*$ is locally Lipschitz continuous. Then, it is known from [27, Corollary 2.4.3] that (4.21) is a necessary condition for $\overline{X}$ to be an optimal solution to the problem (4.5).

The general convergence result in Theorem 2.19 can be applied to the adaptive semi-nuclear norm regularization approach. The following result will be helpful in this regard.

Lemma 4.7. *Suppose that $f \in \overline{\mathcal{C}}(\mathbb{R}_+)$ is continuously differentiable. If $\mu := s_n$ or $\mu := f'_+(0)$, and G is chosen as (4.15), then $\widehat{p}(X, Y)$ defined by (4.16) is a continuous function.*

Proof. The continuity of $\mu := s_n$ with respect to Y comes from the continuous differentiability of f. In this case, we can write G as a spectral operator

$$G = U \mathrm{Diag}\big(\psi(\sigma(Y))\big) V^{\mathbb{T}},$$

associated with a symmetric function $\psi : \mathbb{R}^n \to \mathbb{R}^n$ defined by

$$\psi_i(x) = \begin{cases} \operatorname{sgn}(x_i)\left(1 - \dfrac{f'(|x_i|)}{f'(|x|_{\min})}\right) & \text{if } x \in \mathbb{R}^n \backslash \{0\}, \\ 0 & \text{if } x = 0, \end{cases} \tag{4.22}$$

where $(U, V) \in \mathbb{O}^{n_1 \times n_2}(Y)$ and $|x|_{\min}$ denotes the smallest component of $|x|$. It is easy to see that the function ψ is continuous. It then follows from [33, Chapter 3] that G is also continuous with respect to Y. Hence, $\widehat{p}(X, Y)$ defined by (4.16) is a continuous function. The argument for the case that $\mu = f'_+(0)$ is the same expect replacing $f'(|x|_{\min})$ with $f'_+(0)$ in (4.22). □

Theorem 4.8. *Let $\{X^k\}$ be a sequence generated from Algorithm 4.2. Suppose that $f \in \overline{\mathcal{C}}(\mathbb{R}_+)$ and $0 < \underline{\gamma} \le \gamma^k \le \overline{\gamma} < \infty \ \forall\, k \ge 0$. Then, the sequence $\{h(X^k) + \rho\|F(X^k)\|_*\}$ is monotonically decreasing. In addition, suppose that f is further continuously differentiable and $\inf_k \alpha^k > 0$. If $\mu^k := s_n^k$ or $\mu^k := f'_+(0)$, and G^k is chosen as (4.15), then any limit point of $\{X^k\}$ is a stationary point of the problem (4.5).*

Proof. The problem (4.5) can be equivalently written as

$$\min_{X \in \mathbb{M}^{n_1 \times n_2}} \quad h(X) + \rho\,\|F(X)\|_* + \delta_K(X). \tag{4.23}$$

According to Lemma 4.7, the function $\widehat{p}(X, Y)$ is continuous. Notice that $\partial \delta_K(X) = N_K(X)$ for any $X \in K$. Then after applying Theorem 2.19 to the problem (4.23), we only need to show that for any limit point $\overline{X}$ of $\{X^k\}$,

$$\partial \widehat{p}_{\overline{X}}(\overline{X}) = f'\big(\sigma_n(\overline{X})\big)\big(\partial\|\overline{X}\|_* - \overline{G}\big) \subseteq \partial\|F(\overline{X})\|_*,$$

where $\widehat{p}_{\overline{X}}(X) := \widehat{p}(X, \overline{X})$ with the function $\widehat{p}(X, Y)$ being defined by (4.16), and $\overline{G} = \overline{U}\mathrm{Diag}\big(\psi(\sigma(\overline{X}))\big)\overline{V}^{\mathbb{T}}$ with ψ being defined by (4.22) and $(\overline{U}, \overline{V}) \in \mathbb{O}^{n_1, n_2}(\overline{X})$. Let $r = \operatorname{rank}(\overline{X})$ and write $\overline{U} = [\overline{U}_1, \overline{U}_2]$ with $\overline{U}_1 \in \mathbb{O}^{n_1 \times r}$, $\overline{U}_2 \in \mathbb{O}^{n_1 \times (n_1 - r)}$ and

$\overline{V} = [\overline{V}_1, \overline{V}_2]$ with $\overline{V}_1 \in \mathbb{O}^{n_2 \times r}$, $\overline{V}_2 \in \mathbb{O}^{n_2 \times (n_2 - r)}$. Since $\psi_i(x) = 0$ if $x_i = 0$, we can rewrite $\overline{G}$ as

$$\overline{G} = \overline{U}_1 \mathrm{Diag}\left(1 - \frac{f'(\sigma_1(\overline{X}))}{f'(\sigma_n(\overline{X}))}, \ldots, 1 - \frac{f'(\sigma_r(\overline{X}))}{f'(\sigma_n(\overline{X}))}\right)\overline{V}_1^{\mathbb{T}}.$$

Moreover, it follows from [176, 177] that the subdifferential of the nuclear norm at $\overline{X}$ takes the form

$$\partial\|\overline{X}\|_* = \left\{\overline{U}_1\overline{V}_1^{\mathbb{T}} + \overline{U}_2 W \overline{V}_2^{\mathbb{T}} \mid \ W \in \mathbb{M}^{(n_1 - r)\times(n_2 - r)} \text{ with } \|W\| \leq 1\right\}.$$

Then from Theorem 2.4, we obtain that

$$\begin{aligned} & f'\big(\sigma_n(\overline{X})\big)\big(\partial\|\overline{X}\|_* - \overline{G}\,\big) \\ = & \left\{\overline{U}_1 \mathrm{Diag}\big(f'(\sigma_1(\overline{X})), \ldots, f'(\sigma_r(\overline{X}))\big)\overline{V}_1^{\mathbb{T}} + \overline{U}_2 W \overline{V}_2^{\mathbb{T}} \;\middle|\; \|W\| \leq 1\right\} \\ \subseteq & \ \partial\|F(\overline{X})\|_*. \end{aligned}$$

Thus, we complete the proof. □

The two options $\mu^k := s_n^k$ and $\mu^k := f'_+(0)$ differ very little in practice since each iteration produces a low-rank solution due to a semi-nuclear norm and thus these two options lead to the same value of μ^k. Moreover, as can be seen from the proof of Theorem 2.19, the results in Theorem 4.8 also hold if the global continuous differentiability of the function f is slightly relaxed to the local continuous differentiability at all $\sigma_i(\overline{X}), i = 1, \ldots, n$.

We further remark here that according to Theorem 4.8, it is recommended to choose a strictly increasing continuously differentiable function $f \in \overline{\mathcal{C}}(\mathbb{R}_+)$ to construct the approximation problem (4.5). The continuous differentiability is for the theoretical convergence guarantee. Meanwhile, the strict monotonicity is to reduce the number of undesired stationary points of the problem (4.5) at which Algorithm 4.2 may terminate unexpectedly, especially for the case that $h \equiv 0$. The strict monotonicity also helps to avoid the possibility of $\mu^k = 0$ in the subproblem (4.19) if we choose $\mu^k = s_n^k$.

4.3.3 Related discussions

The proposal of using the adaptive semi-nuclear norm regularization approach is inspired by the rank-corrected procedure for matrix completion problems discussed in Chapter 3. Recall that in the rank-correction step (3.8), combing the nuclear norm with either the spectral operator defined by (3.70) when the rank is known or the spectral operator defined by (3.71), (3.72), (3.73) yields a semi-nuclear norm. As a result, theoretical results and numerical experiments in Chapter 3 already provide substantial evidences to support the efficiency of the semi-nuclear technique for addressing the low-rank structure.

Let us take a look at the computational cost of our proposed adaptive semi-nuclear norm regularization approach — Algorithms 4.1 and 4.2, which mainly lies in solving the subproblems (4.17) and (4.19) respectively. Notice that the subproblem (4.19) in Algorithm 4.2 can be equivalently (up to a constant term in the objective function) written as

$$\begin{array}{ll} \min & \dfrac{\gamma^k}{2}\|X - Y^k\|_F^2 + \rho\mu^k\|X\|_* \\ \text{s.t.} & X \in K. \end{array} \tag{4.24}$$

where

$$Y^k := X^k - \frac{1}{\gamma^k}\nabla h(X^k) + \frac{\rho\mu^k}{\gamma^k}G^k.$$

For the unconstrained case, i.e., $K = \mathbb{M}^{n_1\times n_2}$, by using the singular value soft-thresholding operator defined by (2.4), the problem (4.24) has the unique solution $\widetilde{X}^{k+1}$ taking the form

$$\widetilde{X}^{k+1} = \mathcal{P}_{\tau^k}^{\rm soft}(Y^k) \quad \text{with} \quad \tau^k := \rho\mu^k/\gamma^k.$$

For more general constraints, due to a semi-nuclear norm regularization, the favorable singular value soft-thresholding operator can be fully used to design efficient algorithms to solve the subproblems (4.17) and (4.19). This is another aspect what

we can benefit from a semi-nuclear norm besides its low-rank structure-preserving ability.

Many existing algorithms designed for the nuclear norm regularization problems can be directly applied or slightly modified to solve the subproblems (4.17) and (4.19), e.g., [17, 109, 115, 81, 80], to name but only a few. For example, when K is the set of linear equality and/or inequality constraints, the proximal alternating direction method of multipliers (proximal ADMM) could be a satisfactory choice for solving (4.17) in many situations, especially when the loss function h is a quadratic function. One may refer to Appendix B of [52] for detailed discussions on the convergence of this algorithm. The advantage of the proximal ADMM is the fast reduction of the loss and the rank. However, in some circumstances, it also has difficulty in achieving high feasibility for hard problems, which is a common shortcoming of first-order methods. To meet the demand of high feasibility of certain hard constraints, the semismooth/smoothing Newton-CG method could be a suitable choice. Recent developed methodologies for the nuclear norm regularized problems in [79, 81, 80] had demonstrate the efficiency of using the semismooth Newton-CG method and the smoothing Newton-BiCG method for hard constraints. It is worthy of noticing that the main aim of the first few iterations is to gain a proper semi-nuclear norm. Therefore, for improving the computational efficiency, moderate accuracy of solutions is already enough in the first few iterations. In view of this, a potentially good idea is taking the full use of the advantages of both the proximal ADMM and the semismooth/smoothing Newton-CG method — first using the proximal ADMM for finding a proper semi-nuclear norm and then switching to the semismooth/smoothing Newton-CG method for the required high feasibility. However, so far, large-scale problems with hard constraints remain to be further explored since the semismooth Newton-CG method and the smoothing Newton-BiCG method also have difficulty in dealing with a

large number of constraints.

In particular, for the rank minimization problem

$$\begin{aligned} \min \ & \operatorname{rank}(X) \\ \text{s.t.} \ & X \in K, \end{aligned} \tag{4.25}$$

we propose the adaptive semi-nuclear norm minimization, which is a specialized version of Algorithm 4.1 when the loss function h vanishes.

Algorithm 4.3. (Adaptive semi-nuclear norm minimization)

Step 0. *Choose* $f \in \overline{\mathcal{C}}(\mathbb{R}_+)$. *Input* $X^0 \in \mathbb{M}^{n_1 \times n_2}$. *Set* $k := 0$.

Step 1. *Compute* $\sigma(X^k)$. *Construct* $\mathbb{R}^n_+ \ni w^k = (1 - s^k_1/\mu^k, \ldots 1 - s^k_n/\mu^k)^T$ *with* $s^k_i \in [f'_+(\sigma_i(X^k), f'_-(\sigma_i(X^k))]$ *and* $\mu^k \geq s^k_n$. *Choose* G^k *to be an element in the subdifferential of* $\|\cdot\|_{w^k}$ *at* X^k. *Choose* $\gamma^k \downarrow \overline{\gamma} \geq 0$.

Step 2. *Compute the optimal solution* X^{k+1} *to the (strongly) convex problem:*

$$\begin{aligned} \min \ & \|X\|_* - \langle G^k, X\rangle + \frac{\gamma^k}{2}\|X - X^k\|^2_F \\ \text{s.t.} \ & X \in K. \end{aligned} \tag{4.26}$$

Step 3. *If converged, stop; otherwise, set* $k := k + 1$ *and go to* **Step 1**.

Algorithm 4.3 is specialized version of Algorithm 4.1 when the loss function h vanishes. For the special positive semidefinite case $K \subset \mathbb{S}^n_+$, the adaptive semi-nuclear norm minimization reduces to the reweighted trace minimization of Fazel, Hindi and Boyd in [51], except for the proximal term $\frac{\gamma^k}{2}\|X - X^k\|^2_F$ added in the objective in each subproblem of the adaptive semi-nuclear norm minimization. The proximal term plays an important role in two aspects. One aspect is to stabilize the solution to the subproblem (4.26) to make it unique and bounded. The other aspect is to save the computational cost for solving the subproblem. In

particular, if the surrogate of the rank function is the nuclear norm, i.e., $\|F(X)\|_* = \|X\|_*$, then $G^k \equiv 0$ for all $k \geq 0$ and thus Algorithm 4.3 reduces to the (primal) proximal point algorithm for solving the nuclear norm minimization problem, e.g., see [109]. Roughly speaking, the computational cost of the adaptive semi-nuclear norm minimization is of the same order as that of the nuclear norm minimization in general. We also remark here that in the practical implementation, one may simply set the parameter $\gamma^k \equiv 0$ provided that the subproblem (4.26) can be efficiently solved by a certain method.

Finally, for the special case $K = \mathbb{M}^{n_1 \times n_2}$, we also provide another simple iterative method to solve the unconstrained version of (4.5). For simplicity of illustration, we assume that ∇h is Lipschitz continuous with constant $\kappa > 0$. Then we can easily construct a majorization of the objective function such that for any $X, Y \in \mathbb{M}^{n_1 \times n_2}$,

$$h(X) + \rho\|F(X)\|_* \leq h(Y) + \langle \nabla h(Y), X - Y\rangle + \frac{\kappa}{2}\|X - Y\|_F^2 + \rho\|F(X)\|_*.$$

Compared with the adaptive semi-nuclear norm regularization approach, here, we only majorize the function $h(X)$ but not $\|F(X)\|_*$. Thus, we do not need the restriction $f'_+(0) < \infty$. According to the framework of majorization methods discussed in Section 2.6, with an initial input X^0, we generate $\{X^{k+1}\}$ by solving a sequence of nonconvex optimization problems as

$$X^{k+1} \in \underset{X \in \mathbb{M}^{n_1 \times n_2}}{\arg\min} \left\{ \frac{\kappa}{2}\|X - Z^k\|_F^2 + \rho\|F(X)\|_* \right\}, \tag{4.27}$$

where $Z^k = X^k - \frac{1}{\kappa}\nabla h(X^k)$. Although being nonconvex, due to the simplicity of no constraint, each problem (4.27) has an optimal solution (may not unique) taking the form

$$X^{k+1} = U_k \mathrm{Diag}(x^{k+1}) V_k^{\mathbb{T}},$$

where $(U_k, V_k) \in \mathbb{O}^{n_1,n_2}(Z^k)$ and $x^k \in \mathbb{R}^n_+$ such that

$$x_i^{k+1} \in \arg\min_{t\in\mathbb{R}_+} \left\{ \frac{\kappa}{2}\big(t - \sigma_i(Z^k)\big)^2 + \rho f(t) \right\} \quad \forall\, i = 1, \ldots, n. \tag{4.28}$$

Thus, in each iteration, one only needs to solve n numbers of one-dimensional optimization problems. For some special functions, say $f(t) = t^{1/2}$ or $f(t) = \log(t)$, each problem (4.28) has a closed-form solution. Even if not for general cases, solving such one-dimensional problems is essentially not a time-consuming work in matrix optimization problems. In this iterative method, line search may also be included as what we did in the adaptive semi-nuclear norm regularization approach, especially when ∇f is not Lipschitz continuous. However, as majorization functions are nonconvex, the general convergence result of this iterative method described above has not been explored yet. This will be left to the further work.

4.4 Candidate functions

Now we list several families of candidate functions from $\overline{\mathcal{C}}(\mathbb{R}_+)$ that are available for our proposed adaptive semi-nuclear norm regularization approach in Table 4.4. We also plot a representable function in each family under the same standard in Figure 4.4 for comparison.

Let us take a look at the functions listed in Table 4.4. For a fixed $\varepsilon > 0$, the functions $f_\varepsilon^{[i]}$, $i = 1, \ldots, 6$, are continuously differentiable over $\mathbb{R}_+$; while the function $f_\varepsilon^{[7]}$ is continuously differentiable over $\mathbb{R}_+$ if $p \geq 2$. The functions $f_\varepsilon^{[i]}$, $i = 1, \ldots, 5$, are strictly increasing over $\mathbb{R}_+$. When $\varepsilon \downarrow 0$, the sequences of functions $\{f_\varepsilon^{[i]}\}$, $i = 3, \ldots, 7$, are nondecreasing and pointwise converge to the indicator function $\mathbb{1}(\cdot)$ over $\mathbb{R}_+$ respectively. Thus, it follows from Corollary 4.4 that $f_\varepsilon^{[i]}(\cdot) \xrightarrow{e} \mathbb{1}(\cdot)$ and $\|F_\varepsilon^{[i]}(\cdot)\|_* \xrightarrow{e} \mathrm{rank}(\cdot)$ as $\varepsilon \downarrow 0$ for $i = 3, \ldots, 7$.

All the functions listed in Table 4.4 can be found in the literature to be used to

Table 4.1: Several families of candidate functions defined over $\mathbb{R}_+$ with $\varepsilon > 0$

Log function	$f_\varepsilon^{[1]}(t) := \log(t+\varepsilon) - \log(\varepsilon)$.
L_q function $(0 < q < 1)$	$f_\varepsilon^{[2]}(t) := (t+\varepsilon)^q - \varepsilon^q$.
Fractional function	$f_\varepsilon^{[3]}(t) := t/(t+\varepsilon)$
Exponential function	$f_\varepsilon^{[4]}(t) := 1 - e^{-t/\varepsilon}$
Arctan function	$f_\varepsilon^{[5]}(t) := \frac{2}{\pi}\arctan(t/\varepsilon)$
Smoothly clipped absolute deviation (SCAD) function $(a > 2)$	$f_\varepsilon^{[6]}(t) := \begin{cases} \frac{2t}{(a+1)\varepsilon} & \text{if } 0 \le t < \varepsilon \\ -\frac{(t^2-2a\varepsilon t+\varepsilon^2)}{(a^2-1)\varepsilon^2} & \text{if } \varepsilon \le t < a\varepsilon \\ 1 & \text{if } t \ge a\varepsilon \end{cases}$
Generalized hard-thresholding (GHT) function $(p \ge 1)$	$f_\varepsilon^{[7]}(t) := 1 - \left((1 - t/\varepsilon)_+\right)^p$

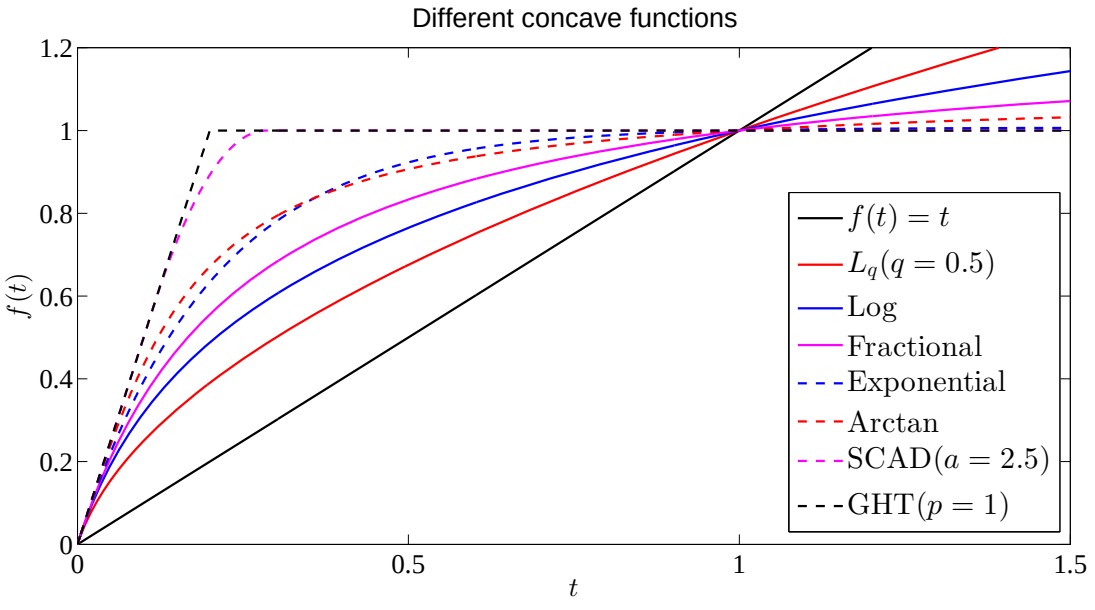


Figure 4.1: For comparison, each function f is scaled with a suitable chosen parameter such that $f(0) = 0$, $f(1) = 1$ and $f'_+(0) = 5$.

construct sparsity-promoting functions, especially in the field of variable selection and signal/image restoration. Some of them can be tracked back to the early 90's, e.g., see [63, 58]. These functions are designed towards "bridging" the gap between the cardinality and the l_1 norm via using concave functions to achieve a sparse vector with certain desired properties. In particular, the function $f_\varepsilon^{[6]}$ is the normalized version of the SCAD penalty function proposed in [42] (see also [43]), and the function $f_\varepsilon^{[7]}$ with $p = 2$ is the normalized version of the hard-thresholding (HT) penalty function (see [36]), which is also a special case of the minimax concave penalty (MCP) function proposed in [190]. (The SCAD penalty function and the MCP function are used as Type 2 penalties in [43] and [190] respectively, where "Type 2" means that the regularization parameter cannot be separated from the penalty function. Differently, here, we use these two kinds of functions as Type 1 penalties in the approximation problem (4.5), allowing for the separation of the regularization parameter and the penalty function.)

We also remark that the functions $f(t) = t^q$, $0 < q < 1$, are ruled out to be candidate functions for our proposed adaptive semi-nuclear norm regularization approach, since the requirement $f'_+(0) < +\infty$ is not satisfied. In compressed sensing, the l_q quasi-norm with $0 < q < 1$, given by $\|x\|_q^q := \sum_{i=1}^n |x_i|^q$, has been shown to have better theoretical recoverability for a sparse vector from a number of linear measurements, compared with the l_1 norm. However, the resulting l_q regularized problem is computationally intractable, and therefore in many works of literature, the l_q quasi-norm is approximated by $\sum_{i=1}^n (|x_i|+\varepsilon)^q$ or $\sum_{i=1}^n (x_i^2+\varepsilon)^{q/2}$ for some small $\varepsilon > 0$, e.g., see [24, 25, 55, 95]. A similar situation also occurs for the matrix case. The Schatten-q quasi-norm with $0 < q < 1$, given by $\|X\|_q^q := \sum_{i=1}^n \sigma_i^q(X)$, has been shown to be a better surrogate of the rank function in term of recoverability, e.g., see [140, 94, 188]. Some specialized algorithm have also been designed for special cases, e.g., see the one we discussed at the end of Section

4.3.3 and see also [116]. However, as our proposed adaptive semi-nuclear norm regularization approach is a general-purpose method including dealing with hard constraints, the Schatten-q quasi-norm is ruled out to be a surrogate of the rank function due to its computational difficultly. But, this sacrifice does not matter because other candidate functions are capable of taking its place without any loss of effectiveness.

Other surrogates of the rank function have also been discussed in the literature. For example, Mohan and Fazel [130], as well as Fornasier, Rauhut and Ward [54], considered the surrogate function $\sum_{i=1}^{n}\left(\sigma_i(X)^2+\varepsilon\right)^{q/2}$, $0<q<1$ and proposed the iterative reweighted least squares minimization, which will be discussed in details in Subsection 4.5.1. Another similar function $\sum_{i=1}^{n}\log\left(\sigma_i(X)^2+\varepsilon\right)$ was also considered in [130]. Zhao [192] considered the surrogate function $\sum_{i=1}^{n}\sigma_i^2(X)/\left(\sigma_i^2(X)+\varepsilon\right)$ and reformulated the resulting problem to be a linear bilevel semidefinite programming (SDP) which was further solved approximately by an SDP. Moreover, Ghasemi et al. [66] considered the surrogate function $n-\sum_{i=1}^{n}e^{-\sigma_i^2(X)/\varepsilon}$ for the matrix completion problem and applied the gradient projection method to solve the resulting problem without convergence analysis. To take a closer look, we can find that all the above four surrogate functions are of the form $\|F(X^{\mathbb{T}}X)\|_*$ with F being the Löwner's operator associated with some function $f\in\overline{\mathcal{C}}(\mathbb{R}_+)$, more precisely, the functions $f_\varepsilon^{[i]}$, $i=1,\ldots,4$, respectively. As $\mathrm{rank}(X)=\mathrm{rank}(X^{\mathbb{T}}X)$, this kind of surrogate functions can also be interpreted as being proposed according to the same principle of approximation which we follow. Differently, each function $\|F(X^{\mathbb{T}}X)\|_*$ with $f\in\overline{\mathcal{C}}(\mathbb{R}_+)$ is smooth. (This property is not needed in our proposed adaptive semi-nuclear norm regularization approach.) However, in exchange, the rank-promoting ability of $\|F(X^{\mathbb{T}}X)\|_*$ is somewhat weaker than that of the corresponding surrogate function $\|F(X)\|_*$. This is because the function $f(t^2)$ is no longer concave but indeed convex in $[\,0,\,\bar{t}\,]$ for some $\bar{t}>0$. A special

case of the log function $f_\varepsilon^{[2]}$ can be seen in Figure 4.4. This phenomenon can be seen much more clearly in the extreme case $f(t) = t\ \forall\, t \geq 0$, i.e., $\|F(X)\|_* = \|X\|_*$ and $\|F(X^{\mathbb{T}}X)\|_* = \|X\|_F^2$, as we know that the Frobenius norm does not encourage the low-rank in general. This inherent weaknesses may limit the efficiency of using the surrogate function $\|F(X^{\mathbb{T}}X)\|_*$ for finding a low-rank solution. Moreover, the S-shape of such function may also bring difficulty for designing an efficient algorithm.

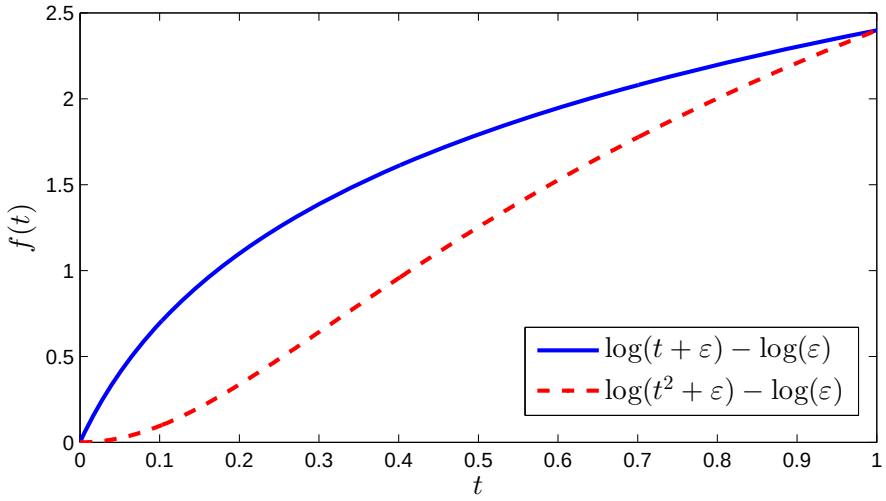


Figure 4.2: Comparison of $\log(t+\varepsilon)-\log(\varepsilon)$ and $\log(t^2+\varepsilon)-\log(\varepsilon)$ with $\varepsilon = 0.1$.

We also attempt to discuss the difference of all these candidate functions — which one we prefer to choose in general to form a surrogate of the rank function? Without the support of numerical experiments, this seems to be hard. However, we can still find some inklings of this matter right now. As we discussed at the end of Section 4.2.1, for better rank-promoting ability, the function $f \in \overline{\mathcal{C}}(\mathbb{R}_+)$ needs to be "more concave" to imitate the behavior of indicator function $\mathbb{1}(\cdot)$ more aggressively. But the discussion there does not take into account of the computational difficulty for a nonconvex optimization problem. Indeed, after applying the adaptive semi-nuclear norm regularization approach, the decrease of the rank could be slow for

a too aggressive choice. To gain insights into this, we may consider a simple one-dimensional example as

$$\min_{t\geq 0} \quad \frac{1}{2}(t-1)^2 + \mathbb{1}(t). \tag{4.29}$$

It is easy to see that the optimal solution to (4.29) is $t^* = 0$. Suppose that we replace the indicator function $\mathbb{1}(\cdot)$ with a differentiable function $f \in \overline{\mathcal{C}}(\mathbb{R}_+)$ as

$$\min_{t\geq 0} \quad \frac{1}{2}(t-1)^2 + f(t). \tag{4.30}$$

At each iterate $t^k \neq 0$, the adaptive semi-nuclear norm regularization approach yields the next iterate t^{k+1} as

$$t^{k+1} := \arg\min_{t\geq 0} \left\{ \frac{1}{2}\big(t - (1 - f'(t^k)\big)^2 \right\} = \begin{cases} 1 - f'(t^k) & \text{if } f'(t^k) < 1, \\ 0 & \text{if } f'(t^k) \geq 1. \end{cases}$$

Therefore, if the function f is "too concave", then $f'(t^k)$ will be small when t^k is away from zero. Thus, if the initial point is not near zero, many iterations will be needed for achieving the optimal solution $t^* = 0$ to the problem (4.29). We also see that $f'(0) \geq 1$ is necessary in this example to make sure that the iterative scheme capable to achieve $t^* = 0$ eventually. This refers to the zero-promoting ability of the surrogate function f. (In this example, one may think of choosing $f(t) = \rho\, t \;\forall\, t \geq 0$ with sufficiently large $\rho > 0$ for achieving $t^* = 0$ in one-step, rather than strengthening the concavity of f. However, this choice distorts the original target of the problem (4.29) since the problem (4.30) with this choice is more like checking whether $t^* = 0$ is a feasible solution, rather than minimizing the objective function of (4.29).) As a result, in order to construct a suitable surrogate of the rank function, one needs to balance the rank-promoting ability and the computational efficiency. Taking a look at the shape of the function f may be helpful to make a prediction. In general, a function $f \in \overline{\mathcal{C}}(\mathbb{R}_+)$ with a steady decrease of derivatives (or subderivatives) to zero as $t \to \infty$ may be preferred for computational robustness. If one aims to be more aggressive, much more attention should be paid for choosing the initial point and the parameters.

4.5 Comparison with other works

In this section, we compare the adaptive semi-nuclear norm regularization approach with some existing algorithms for solving the rank regularized problem (4.1).

4.5.1 Comparison with the reweighted minimizations

For simplicity, the comparison in this subsection focuses on solving the rank minimization problem (4.25). As the rank of a matrix is the cardinality of its singular values, the rank minimization problem is an extension of the cardinality minimization problem taking the form

$$\begin{aligned} \min \quad & \|x\|_0 \\ \text{s.t.} \quad & x \in C, \end{aligned} \tag{4.31}$$

where C is a closed convex subset of $\mathbb{R}^n$. Two classes of iterative algorithms, called iterative reweighted l_1 and l_2 minimizations respectively, have been shown to outperform the l_1 minimization for finding a sparse solution in terms of solution quality, though some results related to the convergence and the sparse estimation have not been fully exploited yet. Let $f \in \overline{\mathcal{C}}(\mathbb{R}_+)$ be chosen. At each iterate x^k, the iterative reweighted l_1 minimization yields the next iterate x^{k+1} by solving the weighted l_1 minimization as

$$x^{k+1} := \arg\min_{x \in C} \ \sum_{i=1}^{n} w_i^k |x_i|, \tag{4.32}$$

where $w_i^k \in \left[f'_+(|x_i^k|), f'_-(|x_i^k|)\right]$, $i = 1, \ldots, n$; while the iterative reweighted l_2 minimization yields the next iterate x^{k+1} by solving the weighted l_2 minimization as

$$x^{k+1} := \arg\min_{x \in C} \ \sum_{i=1}^{n} w_i^k x_i^2, \tag{4.33}$$

where $w_i^k \in \left[f'_+\big((x_i^k)^2\big), f'_-\big((x_i^k)^2\big)\right]$, $i = 1, \ldots, n$. Among all the functions $f \in \overline{\mathcal{C}}(\mathbb{R}_+)$, two most common choices are the l_q function and the log functions listed in

Table 4.4. One may refer to various detailed discussions related to the l_1 reweighting scheme in [58, 42, 194, 195, 190, 111, 23] and the l_2 reweighting scheme in [148, 26, 29], to name but a few. A survey of these two methods can be found in [181].

Both the iterative reweighted l_1 and l_2 minimizations fall into the category of majorization methods. Since $f \in \overline{\mathcal{C}}(\mathbb{R}_+)$, the function $\sum_{i=1}^n f(x_i)$ is concave over $\mathbb{R}^n_+$ and thus has a linear majorization over $\mathbb{R}^n_+$ at any $y \in \mathbb{R}^n_+$ as

$$\sum_{i=1}^n f(x_i) \leq \sum_{i=1}^n \big(f(y_i) + w_i(x_i - y_i)\big) \quad \forall\, x \in \mathbb{R}^n_+, \tag{4.34}$$

where $w_i \in \big[f'_+(y_i), f'_-(y_i)\big]$, $i = 1, \ldots, n$. The constructions of the majorization functions in these two iterative methods both are based on the property (4.34). The difference between them is how to make use of this property. The design of the iterative reweighted l_1 minimization is based on the equivalence of (4.31) and the problem

$$\begin{aligned} \min \quad & \|z\|_0 \\ \text{s.t.} \quad & |x_i| \leq z_i, \quad i = 1, \ldots, n, \\ & x \in C, \end{aligned} \tag{4.35}$$

where the equivalence means that if x^* is an optimal solution to (4.31), then $(x^*, |x^*|)$ is an optimal solution to (4.35); and conversely if (x^*, z^*) is an optimal solution to (4.35), then x^* is an optimal solution to (4.31). The iterative scheme can be derived from solving the approximation problem

$$\begin{aligned} \min \quad & \sum_{i=1}^n f(z_i) \\ \text{s.t.} \quad & |x_i| \leq z_i, \quad i = 1, \ldots, n, \\ & x \in C, \end{aligned} \tag{4.36}$$

Introducing the auxiliary variable $z \in \mathbb{R}^n_+$ makes the property (4.34) be applicable to majorizing the objective function of (4.36). This gives the iterative scheme for

solving (4.36) as

$$(x^{k+1}, z^{k+1}) := \underset{x,z\in\mathbb{R}^n}{\arg\min} \left\{ \sum_{i=1}^{n} w_i^k z_i \;\middle|\; |x_i| \le z_i, i = 1, \ldots, n,\ x \in C \right\},$$

where $w_i^k \in \left[f'_+(z_i^k), f'_-(z_i^k)\right]$, $i = 1, \ldots, n$. Then, one can recognize the iterative scheme (4.32) by noting that $z^k = x^k$ in each iteration. Differently, the design of the iterative reweighted l_2 minimization is based on the simple fact that $\|x\|_0 = \|x^2\|_0$. The iterative scheme can be derived from solving the approximation problem

$$\begin{aligned} \min \quad & \sum_{i=1}^{n} f(x_i^2) \\ \text{s.t.} \quad & x \in C. \end{aligned}$$

By introducing $z = x^2$ and applying the same argument as above, one can recognize the iterative scheme (4.33). We also remark here that the above discussions only focus on the basic frameworks of the l_1 and l_2 reweighting schemes. For the detailed implementation, some variants of the above reweighting schemes also exist and may further improve the performance.

As an extension of the iterative reweighted l_1 minimization from the vector case to the matrix case, Fazel, Hindi and Boyd [51] (see also [49]) proposed the reweighted trace minimization for the positive semidefinite matrix rank minimization problem

$$\begin{aligned} \min \quad & \operatorname{rank}(X) \\ \text{s.t.} \quad & X \in K \subseteq \mathbb{S}_+^n, \end{aligned}$$

where K is a closed convex subset of $\mathbb{S}_+^n$. Among different implementations, the most representable one is called the log-det heuristic, where the rank function over $\mathbb{S}_+^n$ is surrogated by $\log\det(X + \varepsilon I_n)$, $X \in \mathbb{S}_+^n$ for some $\varepsilon > 0$. This surrogate function is equal to $\sum_{i=1}^n \log\left(\lambda_i(X)+\varepsilon\right)$, $X \in \mathbb{S}_+^n$ and thus corresponds to $\|F(X)\|_*$ associated with the log function listed in Table 4.4. It is not hard to see that the

function $\log\det(X+\varepsilon I_n)$ is concave over $\mathbb{S}^n_+$ so that it has a linear majorization over $\mathbb{S}^n_+$ at any $Y \in \mathbb{S}^n_+$ as

$$\log\det(X+\varepsilon I_n) \leq \log\det(Y+\varepsilon I_n) + \langle (Y+\varepsilon I_n)^{-1}, X-Y\rangle \quad \forall\, X \in \mathbb{S}^n_+. \tag{4.37}$$

Here, $(Y+\varepsilon I_n)^{-1}$ is the derivative of the function $\log\det(X+\varepsilon I_n)$ at $Y \in \mathbb{S}^n_+$. Thus, for the case $K \subseteq \mathbb{S}^n_+$, at each iterate X^k, the log-det heuristic yields the next iterate X^{k+1} by solving the following convex optimization problem:

$$X^{k+1} := \underset{X\in K}{\arg\min}\ \langle (X^k+\varepsilon I_n)^{-1}, X\rangle.$$

For the general case $K \subseteq \mathbb{M}^{n_1\times n_2}$, the rank minimization problem (4.25) can be equivalently written as

$$\begin{aligned} \min\ & \frac{1}{2}\big(\operatorname{rank}(Y)+\operatorname{rank}(Z)\big) \\ \text{s.t.}\ & \begin{pmatrix} Y & X \\ X^{\mathbb{T}} & Z \end{pmatrix} \in \mathbb{S}^{n_1+n_2}_+, \quad X \in K. \end{aligned} \tag{4.38}$$

The two auxiliary variables $Y \in \mathbb{S}^{n_1}_+$ and $Z \in \mathbb{S}^{n_2}_+$ make the property (4.37) applicable to the log-det approximation of (4.38). This gives an iterative scheme as

$$\begin{aligned} (X^{k+1}, Y^{k+1}, Z^{k+1}) := \arg\min\ & \frac{1}{2}\big(\langle (Y^k+\varepsilon I_{n_1})^{-1}, Y\rangle + \langle (Z^k+\varepsilon I_{n_2})^{-1}, Z\rangle\big) \\ \text{s.t.}\ & \begin{pmatrix} Y & X \\ X^{\mathbb{T}} & Z \end{pmatrix} \in \mathbb{S}^{n_1+n_2}_+, \quad X \in K. \end{aligned} \tag{4.39}$$

However, the positive semidefinite constraint increases the problem size and thus leads to more computational difficulty. In view of this, later, Mohan and Fazel [132] further simplified the iterative scheme (4.39) by eliminating the two auxiliary variables Y and Z as

$$X^{k+1} := \underset{X\in K}{\arg\min}\ \|W_1^k X W_2^k\|_*,$$

where for each $k \geq 0$, the weights are updated as

$$W_1^{k+1} = (Y^{k+1} + \varepsilon I_{n_1})^{-1/2} \quad \text{and} \quad W_2^{k+1} = (Z^{k+1} + \varepsilon I_{n_2})^{-1/2}. \tag{4.40}$$

Here, Y^{k+1} and Z^{k+1} are the optimal solution to (4.39), taking the form

$$\begin{cases} Y^{k+1} = (W_1^k)^{-1}\widetilde{U}_{k+1}\mathrm{Diag}\big(\sigma(\widetilde{X}^{k+1})\big)\widetilde{U}_{k+1}^{\mathbb{T}}(W_1^k)^{-1}, \\ Z^{k+1} = (W_2^k)^{-1}\widetilde{V}_{k+1}\mathrm{Diag}\big(\sigma(\widetilde{X}^{k+1})\big)\widetilde{V}_{k+1}^{\mathbb{T}}(W_2^k)^{-1}. \end{cases} \tag{4.41}$$

where $\widetilde{X}^{k+1} = W_1^k X^{k+1} W_2^k$ and $(\widetilde{U}_{k+1}, \widetilde{V}_{k+1}) \in \mathbb{O}^{n_1,n_2}(\widetilde{X}^{k+1})$. This procedure is called the reweighted nuclear norm minimization in [132]. Other concave surrogate functions are also applicable to the reweighted nuclear norm minimization.

As can be seen from the above, the way to majorize the surrogate of the rank function in the reweighed nuclear norm minimization in [132] looks totally different from that of our proposed adaptive semi-nuclear norm minimization. Nevertheless, by a simple deduction, it is interesting to see that for the positive semidefinite case $K \subseteq \mathbb{S}_+^n$, the two different ways result in the same linear majorization, in particular whose form for the case of the log function has been shown in (4.37). Theoretically, it is difficult to say which majorization is better. However, computationally, majorizing by a semi-nuclear norm is superior to majorizing by a weighted nuclear norm in general because the former one could make full use of the advantage of the soft-thresholding operator (2.4) but the latter one cannot. Moreover, the construction of the weights W_1^k and W_2^k for a weighted nuclear norm needs three SVDs in each iteration, while the construction of a semi-nuclear norm needs only one SVD. This could make a considerable difference when the problem size is not small. Even if for the positive semidefinite case $K \subseteq \mathbb{S}_+^n$, the adaptive semi-nuclear norm minimization is also distinguished from the reweighted nuclear norm minimization due to the allowance of the proximal term $\frac{\gamma^k}{2}\|X - X^k\|_F^2$ for reducing the computational cost.

In an alternative line of work, the iterative reweighted l_2 minimization for minimizing the vector cardinality was extended by Mohan and Fazel [130] and

also Fornasier, Rauhut and Ward [54] to the iterative reweighted least squares minimization for minimizing the matrix rank. A similar work can also be found in the paper of Lai, Xu and Yin [96]. The design of this method is based on the simple fact $\text{rank}(X) = \text{rank}(X^{\mathbb{T}}X)$. We also take the log-det surrogate as an example, i.e., the approximation problem takes the form

$$\begin{aligned} \min \quad & \log\det(X^{\mathbb{T}}X + \varepsilon I_{n_2}) \\ \text{s.t.} \quad & X \in K. \end{aligned} \tag{4.42}$$

The positive semidefiniteness of $X^{\mathbb{T}}X$ makes the property (4.37) be applicable to majorizing this surrogate function. Thus, at each iterate X^k, the iterative reweighted least squares minimization yields the next iterate X^{k+1} by solving a weighted Frobenius norm minimization as

$$X^{k+1} := \arg\min_{X\in K} \ \left\langle \left((X^k)^{\mathbb{T}}X^k + \varepsilon I_{n_2}\right)^{-1}, X^{\mathbb{T}}X \right\rangle. \tag{4.43}$$

The smoothness of the problem (4.43) in each iteration provides possibility for achieving computational efficiency in each iteration. For the case of exact matrix completion, the subproblem (4.43) was solved in column wise by using a number of inversions in [54], (which can be expensive when the matrix size is large), and was solved by using the gradient projection method in [130]. (Another iterative algorithm — a gradient projection algorithm directly applied to (4.42), was also proposed in [130] for exact matrix completion, which is nothing but the iterative reweighted least squares minimization with the subproblem (4.43) being solved by the gradient projection method but terminated after only one iteration.) However, such advantage disappears for rank minimization problems with general hard constraints rather than the matrix completion problem.

Compared with the adaptive semi-nuclear norm minimization, as well as the reweighted nuclear norm minimization, there are two main disadvantages of the iterative reweighted least squares minimization. The first disadvantage is that each

single iteration does not encourage a low-rank solution because the objective function is a weighted Frobenius norm rather than the nuclear norm or a semi-nuclear norm. This may lead to more computational cost for updating the weight in each iteration since full singular value decompositions may be needed. The second disadvantage is that more iterations are needed in general, which is a bottleneck especially if solving the subproblem is time consuming. These disadvantages are crucial when hard constraints are involved. Furthermore, the rank-promoting ability of the surrogate function in the iterative least squares minimization is somewhat slightly weaker than that of the corresponding one in the adaptive semi-nuclear norm minimization and the reweighted nuclear norm minimization, as having been discussed in Section 4.4.

4.5.2 Comparison with the penalty decomposition method

Lu and Zhang [113] proposed a penalty decomposition method for solving the rank regularized problem (4.1). The essential idea is to introduce an auxiliary variable $Y \in \mathbb{M}^{n_1 \times n_2}$ such that $Y = X$ and then to solve a sequence of the penalized problems with increasing penalty parameters. Each penalized problem takes the form

$$\begin{aligned} \min \quad & h(X) + \rho \operatorname{rank}(Y) + \rho\mu \|X - Y\|_F^2 \\ \text{s.t.} \quad & X \in K, \end{aligned} \tag{4.44}$$

where $\mu > 0$. In fact, other equality and inequality constraints are also penalized into the objective function in the proposed penalty decomposition method in [113]. Here, we only focus on the the core idea of this method for comparison. Applying the block coordinate descent method (also known as the alternating minimization

method) to the problem (4.44) yields an iterative scheme as

$$\begin{cases} Y^{k+1} \in \underset{Y\in\mathbb{M}^{n_1\times n_2}}{\arg\min} \left\{\mathrm{rank}(Y) + \mu\|Y - X^k\|_F^2\right\}, \\ X^{k+1} = \underset{X\in K}{\arg\min} \left\{h(X) + \rho\mu\|X - Y^{k+1}\|_F^2\right\}. \end{cases} \tag{4.45}$$

At the first glance of (4.45), it seems that the penalty decomposition method is irrelative to the adaptive semi-nuclear norm regularization approach. But, this is not true. In the following, we disclose the connection between them with details.

By using the singular value hard-thresholding operator defined by (2.5), the iterative scheme (4.45) could be simplified as

$$X^{k+1} = \underset{X\in K}{\arg\min} \left\{h(X) + \rho\mu\left\|X - \mathcal{P}^{\mathrm{hard}}_{1/\mu^{1/2}}(X^k)\right\|_F^2\right\}. \tag{4.46}$$

Let F and $\widehat{F}$ be the Löwner's operators, respectively, associated with

$$f(t) := 1 - (1-\mu t)_+ \quad \text{and} \quad \widehat{f}(t) := f(t^2) = 1 - (1-\mu t^2)_+ \quad \forall\, t \geq 0,$$

where the former one is the GHT function with $p = 1$ listed in Table 4.4. Now, we alternatively consider the following optimization problem

$$\begin{aligned} \min \;\; & h(X) + \rho\,\|\widehat{F}(X)\|_* \\ \text{s.t.} \;\; & X \in K. \end{aligned} \tag{4.47}$$

This problem is nonconvex and nonconcave. Without loss of generality, we assume $n_1 \geq n_2$. (Recall that $n = \min\{n_1, n_2\}$.) Note that $\|\widehat{F}(X)\|_* = \|F(X^{\mathbb{T}}X)\|_*$. Then by regarding $X^{\mathbb{T}}X$ as a whole, according to the majorization scheme in our proposed adaptive semi-nuclear norm regularization approach, we obtain that for any $X, Y \in \mathbb{M}^{n_1\times n_2}$,

$$\begin{aligned} \|\widehat{F}(X)\|_* &\leq \mu\big(\|X^{\mathbb{T}}X\|_* - \|X^{\mathbb{T}}X\|_w\big) + \|F(Y^{\mathbb{T}}Y)\|_* - \mu\big(\|Y^{\mathbb{T}}Y\|_* - \|Y^{\mathbb{T}}Y\|_w\big) \\ &\leq \mu\big(\|X^{\mathbb{T}}X\|_* - \langle G, X^{\mathbb{T}}X - Y^{\mathbb{T}}Y\rangle\big) + \|F(Y^{\mathbb{T}}Y)\|_* - \mu\|Y^{\mathbb{T}}Y\|_*, \end{aligned} \tag{4.48}$$

where the first inequality follows from (4.13), the second inequality follows from (4.14), $w \in \mathbb{R}^n_+$ is given by

$$w_i = \begin{cases} 1 & \text{if } \sigma_i(Y) > \sqrt{1/\mu}, \\ 0 & \text{if } \sigma_i(Y) \le \sqrt{1/\mu}, \end{cases} \qquad i = 1, \ldots, n,$$

and G is an element of the subdifferential of $\|\cdot\|_w$ at $Y^{\mathbb{T}}Y$ taking the form

$$G = V\mathrm{Diag}(w)V^{\mathbb{T}},$$

where $(U, V) \in \mathbb{O}^{n_1,n_2}(Y)$. We further notice that $\langle G, X^{\mathbb{T}}X\rangle$ is a convex function because G is positive semidefinite. Thus,

$$\langle G, X^{\mathbb{T}}X\rangle \le \langle G, Y^{\mathbb{T}}Y\rangle + 2\langle YG, X - Y\rangle \quad \forall\, X, Y \in \mathbb{M}^{n_1\times n_2}. \tag{4.49}$$

It is easy to check that $YG = \mathcal{P}^{\mathrm{hard}}_{1/\mu^{1/2}}(Y)$. Then, combining (4.48) and (4.49) leads to a majorization function of $\|\widehat{F}(X)\|_*$ at Y, i.e., for any $X, Y \in \mathbb{M}^{n_1\times n_2}$,

$$\begin{aligned} \|\widehat{F}(X)\|_* &\le \mu\big(\|X\|_F^2 - 2\big\langle \mathcal{P}^{\mathrm{hard}}_{1/\mu^{1/2}}(Y), X - Y\big\rangle\big) + \|\widehat{F}(Y)\|_* - \mu\|Y\|_F^2 \\ &= \mu\big(\big\|X - \mathcal{P}^{\mathrm{hard}}_{1/\mu^{1/2}}(Y)\big\|_F^2 + \big\|\mathcal{P}^{\mathrm{hard}}_{1/\mu^{1/2}}(Y)\big\|_F^2\big) + \|\widehat{F}(Y)\|_* - \mu\|Y\|_F^2 \\ &= \mu\big\|X - \mathcal{P}^{\mathrm{hard}}_{1/\mu^{1/2}}(Y)\big\|_F^2 + \|\widehat{F}(Y)\|_* - \mu\big\|Y - \mathcal{P}^{\mathrm{hard}}_{1/\mu^{1/2}}(Y)\big\|_F^2. \end{aligned} \tag{4.50}$$

Thus, applying the majorization method gives an iterative scheme exactly the same as (4.46). (The above deduction can also be simplified by using the fact that $\|X^{\mathbb{T}}X\|_w$ is a convex function in X and $\mathcal{P}^{\mathrm{hard}}_{1/\mu^{1/2}}(Y)$ is a subgradient of $\|X^{\mathbb{T}}X\|_w$ at Y.)

Consequently, the penalty decomposition method for the rank regularized problem (4.1) can be interpreted as solving a sequence of the approximation problems (4.47) with $\mu > 0$ increasing by applying the majorization method described above. If we further look closer at the above majorization procedure, we can find that the intermediate step (4.48) can be rewritten as

$$\|\widehat{F}(X)\|_* \le \mu\langle I_n - G, X^{\mathbb{T}}X\rangle + \mu\langle G, Y^{\mathbb{T}}Y\rangle + \|F(Y^{\mathbb{T}}Y)\|_* - \mu\|Y^{\mathbb{T}}Y\|_*. \tag{4.51}$$

This indeed already provides a convex majorization of $\|\widehat{F}(X)\|_*$, which is used in the corresponding reweighted least squares minimization described in Section 4.5.1. We can see that compared with (4.51), the majorization (4.50) sacrifices the tightness of approximation in change of the simplicity of the obtained function. Therefore, compared with the adaptive semi-nuclear norm regularization approach, the advantage and disadvantage of the penalized decomposition method (for a single μ) are similar to that of the iterative reweighted least squares minimization discussed in Section 4.5.1, only differing in more possible computational convenience in each iteration but accompanied by more iterations as the price.

4.5.3 Related to the MPEC formulation

The rank function over the positive semidefinite cone $\mathbb{S}^n_+$ has a positive semidefinite representation as follows:

Lemma 4.9. *For any matrix $X \in \mathbb{S}^n_+$, one has*

$$\operatorname{rank}(X) = \min\big\{\langle I_n, W\rangle \;\big|\; \langle I_n - W, X\rangle = 0,\ W \in \mathbb{S}^n_+,\ I_n - W \in \mathbb{S}^n_+\big\}.$$

Proof. Suppose that $\operatorname{rank}(X) = r$. For any feasible solution W, the three constraints together implies that X and $I_n - W$ have a simultaneous eigenvalue decomposition and in addition, $\operatorname{rank}(I_n - W) + \operatorname{rank}(X) = n$. This further implies that $\lambda_i(W) = 1$, $i = 1, \ldots, r$. Thus, $\langle I_n, W\rangle = \sum_{i=1}^n \lambda_i(W)$ achieves the minimum value r when $\operatorname{rank}(W) = r$. □

This result is an extension of the well-known representation of the cardinality of a vector, i.e., for any vector $x \in \mathbb{R}^n$,

$$\|x\|_0 = \min\big\{\|w\|_1 \;\big|\; (1 - w_i)x_i = 0,\ w_i \geq 0,\ i = 1, \ldots, n\big\}.$$

Therefore, for the positive semidefinite case $K \subseteq \mathbb{S}^n_+$, according to Lemma 4.9, the rank regularized problem (4.1) can be equivalently written to be a mathematical

programming with equilibrium constraints (MPEC) problem as

$$\begin{aligned} \min \;\; & h(X) + \rho \langle I_n, W \rangle \\ \text{s.t.} \;\; & \langle I_n - W, X \rangle = 0, \\ & W \in \mathbb{S}^n_+, \; I_n - W \in \mathbb{S}^n_+, \\ & X \in K \subseteq \mathbb{S}^n_+. \end{aligned} \tag{4.52}$$

A bilinear constraint $\langle I_n - W, X \rangle = 0$ occurs in this conversion in exchange for continuity. By penalizing this constraint into the objective function with a sufficient large $\mu > 0$, we obtain an approximation problem written as

$$\begin{aligned} \min \;\; & h(X) + \rho \langle I_n, W \rangle + \rho\mu \langle I_n - W, X \rangle \\ \text{s.t.} \;\; & W \in \mathbb{S}^n_+, \; I_n - W \in \mathbb{S}^n_+, \\ & X \in K \subseteq \mathbb{S}^n_+. \end{aligned} \tag{4.53}$$

Applying the block coordinate descent method (or alternating minimization method) to the problem (4.53) yields the iterative scheme as

$$\begin{cases} W^{k+1} = \arg\min_W \left\{ \langle I_n - \mu X^k, W \rangle \mid W \in \mathbb{S}^n_+, \; I_n - W \in \mathbb{S}^n_+ \right\}, \\ X^{k+1} = \arg\min_X \left\{ h(X) + \rho\mu \langle I_n - W^{k+1}, X \rangle \mid X \in K \subseteq \mathbb{S}^n_+ \right\}. \end{cases}$$

A simple calculation yields the closed-form solution of W^{k+1} as

$$W^{k+1} = P_k \mathrm{Diag}(w^{k+1}) P_k^T,$$

where $P_k \in \mathbb{O}^n(X^k)$ and $w^k \in \mathbb{R}^n_+$ is given by

$$w_i^{k+1} = \begin{cases} 1, & \text{if } \lambda_i(X^k) > 1/\mu, \\ 0, & \text{if } \lambda_i(X^k) \leq 1/\mu, \end{cases} \qquad i = 1, \ldots, n.$$

One may notice that the above iterative scheme coincides the one in our proposed adaptive semi-nuclear norm regularization approach if the GHT function $f(t) = 1 - (1 - \mu t)_+ \; \forall\, t \geq 0$ is chosen and the function h is not majorized. This interpretation

provides another interesting perspective to understand the adaptive semi-nuclear norm regularization approach, as well as the reweighted trace norm, for the positive semidefinite case.

4.6 Numerical experiments

In this section, we validate the efficiency of our proposed adaptive semi-nuclear norm technique for different problems involving minimizing the rank. All tests were run in MATLAB under Mac OS X system on an Intel Core i7 2.3 GHz processor with 8.00GB RAM.

Before we go into the details of each test problem, we first describe the method we used to solve each subproblem involved in our proposed adaptive semi-nuclear regularization approach in our experiments. All the test problems below in this section can be unified with the formulation as

$$\begin{array}{ll} \min & \dfrac{1}{2}\|\mathcal{A}(X)-b\|_2^2+\rho\,\|F(X)\|_* \\ \text{s.t.} & \mathcal{B}(X)-d\in\mathcal{Q}, \end{array} \tag{4.54}$$

where $\mathcal{A}:\mathbb{R}^{n_1\times n_2}\to\mathbb{R}^m$ and $\mathcal{B}:\mathbb{R}^{n_1\times n_2}\to\mathbb{R}^l$ are linear maps, $b\in\mathbb{R}^m$ and $d\in\mathbb{R}^l$ are vectors, $\mathcal{Q}=\mathbf{0}^{l_1}\times\mathbb{R}^{l_2}_+$ with $l_1+l_2=l$, and F is the Löwner's operator associated with some $f\in\overline{\mathcal{C}}(\mathbb{R}^n_+)$. (For simplicity, we omit such description of F in this sequel of this section.) Then in the k-th iteration, the subproblem (4.17) involved in Algorithm 4.1 with $h^k\equiv h$ and $\mu^k\equiv\mu=f'_+(0)$ can be specified to be

$$\begin{array}{ll} \min & \dfrac{1}{2}\|\mathcal{A}(X)-b\|_2^2+\rho\mu\big(\|X\|_*-\langle G^k,X\rangle\big) \\ \text{s.t.} & \mathcal{B}(X)-d\in\mathcal{Q}. \end{array} \tag{4.55}$$

For notational simplicity, from now on, we omit the superscript "k". In what follows, we introduce the proximal alternating direction multiplier method (proximal ADMM) for solving the problem (4.55), and to be more exact, for its equivalent

reformulation

$$\begin{aligned} \min_{X,y,z} \ & \frac{1}{2}\|y\|_2^2 + \rho\mu\big(\|X\|_* - \langle G, X\rangle\big) \\ \text{s.t.} \ & \begin{pmatrix} \mathcal{A} \\ \mathcal{B} \end{pmatrix}(X) - \begin{pmatrix} y \\ z \end{pmatrix} = \begin{pmatrix} b \\ d \end{pmatrix}, \quad \begin{pmatrix} y \\ z \end{pmatrix} \in \begin{pmatrix} \mathbf{0}^m \\ \mathcal{Q} \end{pmatrix}, \end{aligned} \tag{4.56}$$

where $y \in \mathbb{R}^m$ and $z \in \mathbb{R}^l$ are two auxiliary variables. For more detailed descriptions and convergence analysis of the proximal ADMM, the interested readers may refer to Appendix B of [52]. Given a penalty parameter $\beta > 0$, the augmented Lagrangian function of (4.56) takes the form

$$\begin{aligned} L_\beta(X, y, z, \xi, \zeta) := & \frac{1}{2}\|y\|_2^2 + \rho\mu\big(\|X\|_* - \langle G, X\rangle\big) - \langle \mathcal{A}(X) - y - b, \xi\rangle \\ & - \langle \mathcal{B}(X) - z - d, \zeta\rangle + \frac{\beta}{2}\|\mathcal{A}(X) - y - b\|_2^2 + \frac{\beta}{2}\|\mathcal{B}(X) - z - d\|_2^2. \end{aligned}$$

In the classical ADMM (see, e.g., [60, 67, 38]), the function L_β is minimized with respect to (y, z) and then with respect to X, followed by an update of the multiplier (ξ, ζ). While minimizing L_β with respect to (y, z) admits a closed-form solution, minimizing L_β with respect to X does not have a simple closed-form solution in general and could be costly. To overcome this difficulty, the proximal ADMM introduces an additional proximal term to "cancel out" the complicated parts. The iterative scheme of the proximal ADMM specified for the problem (4.56) can be described as

$$\begin{cases} y^{j+1} := \arg\min\limits_{z} \ L_\beta(X^j, y, z^j, \xi^j, \zeta^j), \\ z^{j+1} := \arg\min\limits_{y \in \mathcal{Q}} \ L_\beta(X^j, y^j, z, \xi^j, \zeta^j), \\ X^{j+1} := \arg\min\limits_{X} \ \Big\{ L_\beta(X, y^{j+1}, z^{j+1}, \xi^j, \zeta^j) + \frac{1}{2}\|X - X^j\|_{\mathcal{S}}^2 \Big\}, \\ \xi^{j+1} := \xi^j - \tau\beta\big(\mathcal{A}(X^{j+1}) - y^{j+1} - b\big), \\ \zeta^{j+1} := \zeta^j - \tau\beta\big(\mathcal{B}(X^{j+1}) - z^{j+1} - d\big), \end{cases}$$

where $\beta > 0$ is the penalty parameter, $\tau \in (0, (\sqrt{5}+1)/2)$ is the step length, $\mathcal{S}$ is a self-adjoint positive semidefinite (not necessary positive definite) operator

and $\|\cdot\|_{\mathcal{S}} := \langle \cdot, \mathcal{S}(\cdot)\rangle$. In particular, one can take $\tau \in (0,2)$ when the (y,z)-part vanishes. An elementary calculation yields the expression of z^{j+1} and y^{j+1} as

$$y^{j+1} = \frac{\beta}{\beta+1}\big(\mathcal{A}(X^j) - b\big) - \frac{1}{\beta+1}\xi^j,$$

and

$$z^{j+1} = \Pi_{\mathcal{Q}}\Big(\mathcal{B}(X) - d - \frac{1}{\beta}\zeta^j\Big),$$

where $\Pi_{\mathcal{Q}}$ is the projection operator onto $\mathcal{Q}$. Meanwhile, in order to "cancel out" the complicated part $\beta(\mathcal{A}^*\mathcal{A} + \mathcal{B}^*\mathcal{B})$, we choose

$$\delta \le \frac{1}{\beta\|\mathcal{A}^*\mathcal{A} + \mathcal{B}^*\mathcal{B}\|} \quad \text{and} \quad \mathcal{S} := \frac{1}{\delta}\mathcal{I} - \beta(\mathcal{A}^*\mathcal{A} + \mathcal{B}^*\mathcal{B}),$$

where $\|\cdot\|$ denotes the spectral norm of the operator. Then, we can explicitly express the update of X^{j+1} as

$$X^{j+1} = \delta\, \mathcal{P}^{\mathrm{soft}}_{\rho\mu}(Y^j), \tag{4.57}$$

where $\mathcal{P}^{\mathrm{soft}}_{\rho\mu}$ is the singular value soft-thresholding operator defined by (2.4) and

$$Y^j = \rho\mu G + \mathcal{A}^*\big(\xi^j - \beta(y^{j+1} - b)\big) + \mathcal{B}^*\big(\zeta^j - \beta(z^{j+1} - d)\big) + \mathcal{S}(X^j).$$

Now we can see that the trick of shifting $\mathcal{A}(X) - b$ from the objective function into the constraint allows more flexibility to control the singular value soft-thresholding to speed up the convergence since we know that the solution will be of low-rank. For a reasonable stopping criterion, we take a look at the dual problem of (4.56) taking the form

$$\begin{aligned} \min_{\xi,\zeta} \quad & \frac{1}{2}\|\xi\|^2 - \langle \xi, b\rangle - \langle \zeta, d\rangle \\ \text{s.t.} \quad & \mathcal{A}^*(\xi) + \mathcal{B}^*(\zeta) + \rho\mu G = \Lambda, \\ & \|\Lambda\| \le \rho\mu, \quad \zeta \in \mathcal{Q}^*, \end{aligned}$$

where $\mathcal{Q}^* = \mathbb{R}^{l_1} \times \mathbb{R}^{l_2}_+$ is the dual cone of $\mathcal{Q}$. It is easy to see from (4.57) that if we let $\Lambda^j = Y^j - X^{j+1}/\delta$, then $\|\Lambda^j\| \le \rho\mu$ and a simple calculation yields

$$\begin{aligned} \Delta^j &:= \mathcal{A}^*(\xi^j) + \mathcal{B}^*(\zeta^j) + \rho\mu G - \Lambda^j \\ &= \frac{1}{\delta}(X^{j+1} - X^j) + \beta\mathcal{A}^*\big(\mathcal{A}(X^j) - y^{j+1} - b\big) + \beta\mathcal{B}^*\big(\mathcal{B}(X^j) - z^{j+1} - d\big). \end{aligned}$$

In view of this, given a tolerance τ_{sub}, we terminate the primal ADMM if

$$\max\left\{R_p^j,\, 0.5R_d^j,\, R^j\right\} < \tau_{\text{sub}}, \tag{4.58}$$

where R_p^j and R_d^j denote the relative primal and dual feasibility of the problem (4.56) and R^j denotes the relative primal feasibility of the original problem (4.55), defined by

$$R_p^j := \sqrt{\frac{\|\mathcal{A}(X^j) - y^j - b\|_2^2 + (\operatorname{dist}(\mathcal{B}(X^j) - d, \mathcal{Q}))^2}{\max\{1, \|b\|_2^2 + \|d\|_2^2\}}},$$

$$R_d^j := \sqrt{\frac{\|\Delta^j\|_F^2}{\max\{1, \|\mathcal{A}^*(b)\|_F^2 + \|\mathcal{B}^*(d)\|_F^2\}} + \frac{(\operatorname{dist}(\zeta^j, \mathcal{Q}^*))^2}{\max\{1, \|b\|_2^2 + \|d\|/_2^2\}}},$$

and

$$R^j := \frac{\operatorname{dist}(\mathcal{B}(X^j) - d, \mathcal{Q})}{\max\{1, \|d\|_2^2\}},$$

where $\operatorname{dist}(\cdot,\cdot)$ denotes the Euclidean distance. The penalty parameter β plays an important role for the efficiency of the proximal ADMM. In our implementation, we let β be self-adjusted according to the ratio R_p^j/R_d^j.

When we apply Algorithm 4.2 to the problem (4.54) or apply Algorithm 4.3 if the loss function vanishes, the subproblem involved in the k-th step iteration can be concisely written as

$$\begin{aligned} \min \;\; & \frac{1}{2}\|X - C^k\|_F^2 + \varrho^k\|X\|_* \\ \text{s.t.} \;\; & \mathcal{B}(X) - d \in \mathcal{Q} \end{aligned} \tag{4.59}$$

for some $\varrho^k > 0$ and some $C^k \in \mathbb{R}^{n_1\times n_2}$. A simple calculation yields the dual problem of this strongly convex optimization problem (4.59) as

$$\min_{\eta\in\mathcal{Q}^*}\left\{\frac{1}{2}\|\mathcal{P}_{\varrho^k}^{\text{soft}}\big(C^k - \mathcal{B}(\eta)\big)\|_F^2 + \langle\xi, d\rangle\right\}. \tag{4.60}$$

To obtain the optimal solution η^* to the dual problem (4.60), we can use the semi-smooth Newton-CG method developed in [80] when only equalities are involved,

i.e., $l_2 = 0$; or use the smoothing Newton-BiCG method developed in [81] when inequalities is involved, i.e., $l_2 > 0$. After that, we can further obtain the unique optimal solution X^k to the subproblem (4.59) in the k-th step via the relation

$$X^k = \mathcal{P}^{\rm soft}_{\varrho^k}\big(C^k - \mathcal{B}^*(\eta^*)\big).$$

Let X^k be generated from the k-th (outer) iteration of Algorithms 4.1, 4.2, 4.3 or other algorithms for comparison. In all the following test problems, we set the initial point $X^0 = 0$. The relative primal feasibility (**relfea** for short) and the relative difference (**reldif** for short) are defined, respectively, by

$$\mathbf{relfea}^k := \frac{\mathrm{dist}(\mathcal{B}(X^k) - d, \mathcal{Q})}{\max\{1, \|d\|_2\}} \quad \text{and} \quad \mathbf{reldif}^k := \frac{\|X^k - X^{k-1}\|_F}{\max\{1, \|X^k\|_F\}}.$$

Let obj^k be the objective value of the problem (4.54) achieved after the k-th iteration, i.e.,

$$\mathrm{obj}^k := \frac{1}{2}\|\mathcal{A}(X^k) - b\|_2^2 + \rho\,\|F(X^k)\|_*.$$

The relative objective difference (**reldif_obj** for short) is defined by

$$\mathbf{reldif\_obj}^k := \frac{|\mathrm{obj}^k - \mathrm{obj}^{k-1}|}{\max\{1, \mathrm{obj}^k\}}.$$

For matrix recovery problems, the relative recovery error (**relerr** for short) is defined by

$$\mathbf{relerr}^k := \frac{\|X^k - \overline{X}\|_F}{\max\{1, \|\overline{X}\|_F\}},$$

where $\overline{X}$ is the unknown matrix to be recovered. We stress that unless specified, we did not perform any special technique in our implementation, especially the rank truncation technique.

4.6.1 Power of different surrogate functions

In this subsection, we test the power of different candidate functions for our proposed adaptive semi-nuclear norm technique for recovering a low-rank matrix from

a number of noiseless Gaussian linear measurements. Let $\overline{X} \in \mathbb{R}^{n_1 \times n_2}$ be the unknown matrix of rank r to be recovered and let $(A_i, b_i), i = 1, \ldots, m$ be the linear measurements with $b_i = \langle A_i, \overline{X} \rangle$. We aim to recover $\overline{X}$ via solving the problem

$$\begin{aligned} \min \quad & \|F(X)\|_* \\ \text{s.t.} \quad & \langle A_i, X \rangle = b_i, \quad i = 1, \ldots, m. \end{aligned} \tag{4.61}$$

In our experiments, we tested for a number of small problems on account of the computational cost. We set $n_1 = n_2 = 60$ and $r = 2$. For each trial, we randomly generated the true matrices $\overline{X}$, normalized in the spetral norm, by the command:

```
ML = randn(n1,r);     MR = randn(n2,r);     s = rand(r,1);
   X_bar = ML*diag(s)*MR';   X_bar = X_bar/norm(X_bar).
```

We also independently generated m random matrices $A_i \in \mathbb{R}^{n_1 \times n_2}$ with i.i.d. Gaussian entries of $N(0, 1/m)$ and then let $b_i := \langle A_i, \overline{X} \rangle$. To test for different numbers of linear measurements, we increased the ratio $\alpha := m/(r(n_1 + n_2 - r))$ from 1 to 3, in which the denominator is the degree of freedom. For each m, we run 20 trial for each candidate function with each given $\varepsilon > 0$.

We applied Algorithm 4.3 to the problem (4.61) and solved each subproblem (4.26) by using the the semi-smooth Newton-CG method. For each subproblem, we fixed the coefficient γ^k of the proximal term such that $\gamma^k \equiv 0.1$. In order to speed up the convergence, the subproblems were solved with tolerance increased from 10^{-3} to 10^{-6}. Algorithm 4.3 were terminated after 200 iterations unless the following stopping criterion is satisfied beforehand:

$$\mathbf{relfea}^k < 10^{-6} \quad \text{and} \quad \mathbf{reldif}^k < 10^{-6}.$$

We declared the true matrix $\overline{X}$ to be successfully recovered if the relative recovery error satisfies

$$\mathbf{relerr}^k \leq 10^{-5}.$$

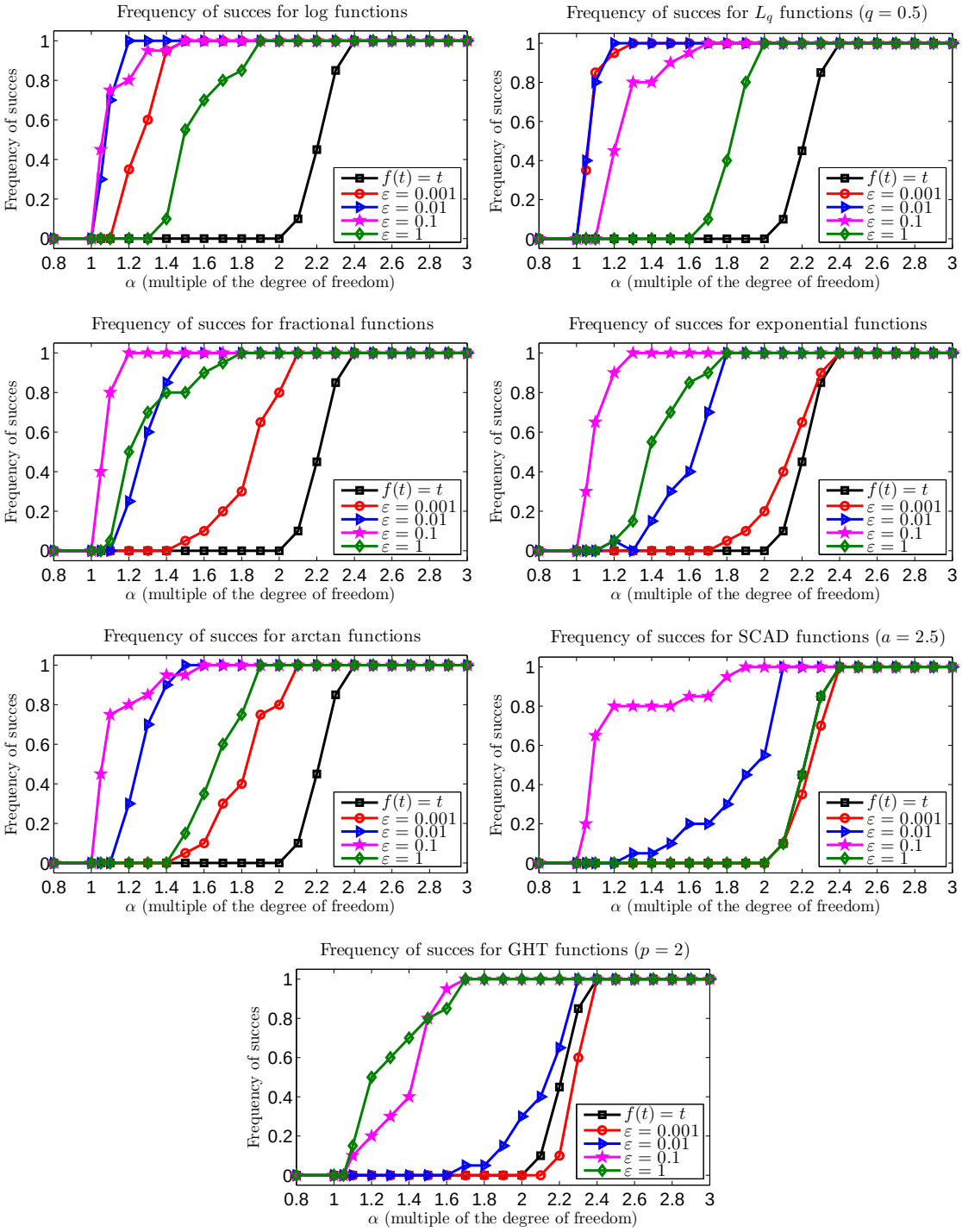
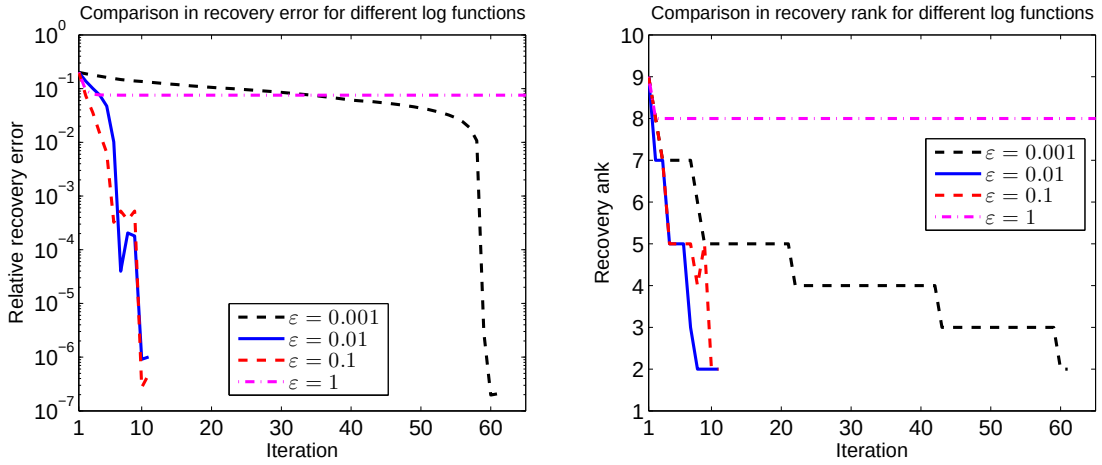


Figure 4.3: Frequency of success for different surrogate functions with different $\varepsilon > 0$ compared with the nuclear norm.

The comparison of recovery in terms of the frequency of success by using different surrogate functions with different $\varepsilon > 0$ and the nuclear norm are presented in Figure 4.3. We can see that compared with the nuclear norm, nonconvex surrogate functions substantially improve the recoverability, with their performances closely related to the chosen parameter ε which controls the function shape. Moreover, compared with the others, the performance of SCAD and GHT functions are more sensitive to the choice of parameter ε since they are not strictly increasing, as we discussed in Subsection 4.3.2. We can also see from Figure 4.3 that the parameter $\varepsilon > 0$ cannot be chosen too large or too small. The detailed sequential performances for one of the test problems using the log functions with different $\varepsilon > 0$ are plotted in Figure 4.4. As can be seen, a large ε ($\varepsilon = 1$) may lead to insufficient recoverability (or rank-promoting ability), while a small ε ($\varepsilon = 0.001$) may lead to very slow sequential progress. It is notable that the value of ε should be interpreted as the ratio to the spectral norm of the true matrix to be recovered because the true matrix has been normalized in spectral norm in our experiments. In addition, it is also interesting to notice that the curve of the relative recovery error has a very sharp decay once it passes through a threshold around 10^{-2}. This observation indicates that the iterative scheme of our proposed adaptive semi-nuclear norm minimization has a good "singular value filtering" effect.

4.6.2 Performance for exact matrix completion

In this section, we test the performance of our proposed adaptive semi-nuclear norm minimization for exact matrix completion problems, i.e., to recover a low-rank matrix from a number of its noiseless observations of entries. Let $\overline{X} \in \mathbb{R}^{n_1 \times n_2}$ be the unknown matrix to be recovered and let Ω denote the set of indices for the



(a) Comparison in relative recovery error

(b) Comparison in recovery rank

Figure 4.4: Comparison of log functions with different ε for exact matrix recovery

observed entries. We aim to recover $\overline{X}$ via solving the following problem

$$\begin{aligned} \min \quad & \|F(X)\|_* \\ \text{s.t.} \quad & X_{ij} = \overline{X}_{ij}, \quad (i,j) \in \Omega. \end{aligned}$$

Mohan and Fazel [130] proposed two iterative algorithms for exact matrix completion. One is called the IRLS-0 algorithm, which is an implementation of the iterative reweighted least squares minimization base on the log surrogate function discussed in Subsection 4.5.1, with its subproblem (4.43) solved by the gradient projection method. The other one is called the sIRLS-0 algorithm, which can be thought of as IRLS-0 but with each subproblem solved approximately using only one iteration of gradient projection. Both of theses two algorithms are feasible algorithms, taking advantage of easy manipulation of the projection onto the special feasible set. To speed up the performances of IRLS-0 and sIRLS-0, a randomized truncated SVD (e.g., see [182, 72]) together with the rank truncation technique are used in the implementation. The corresponding codes can be downloaded at `http://faculty.washington.edu/mfazel/IRLS_final.zip`.

We aim to compare our proposed adaptive semi-nuclear norm minimization approach (Algorithm 4.3, ASNN for short) with IRLS-0 and sIRLS_0. For consistency of comparison, we followed the experimental design of Mohan and Fazel in [130] and used the log function listed in Table 4.4 with $\varepsilon = 0.1$. It should be pointed out that different from our setting, the parameter ε in iterations of IRLS-0 and sIRLS-0 is not fixed but exponentially decreases from 10^{-2} to 10^{-10}. We tested both easy problems and hard problems in terms of different numbers of samples — divided into two categories with more than or less than 2.5 times the degree of freedoms respectively. We created the true matrix $\overline{X} \in \mathbb{R}^{n_1 \times n_2}$ of the form UV^T with $U \in \mathbb{R}^{n_1 \times r}$ and $V \in \mathbb{R}^{n_2 \times r}$ drawn from i.i.d. standard Gaussian entries and normalized to have the spectral norm one. We sampled the indices for observation by using i.i.d. Bernoulli random variables. As was done in [130], we terminated all the three algorithms when the relative recovery error

$$\mathbf{relerr}^k < 10^{-3}$$

was achieved and declared the success of recovery. Because the proximal ADMM is not a feasible algorithm, besides the relative recovery error, we also require one more stopping criterion on the relative primal feasibility for ASNN such that

$$\mathbf{relfea}^k \leq 10^{-4}.$$

It is reasonable to match the accuracy of feasibility to the relative recovery error declared for successful recovery. In addition, when the dimension of the test problem is no more than 200, we slightly modified Mohan and Fazel's codes of IRLS-0 and sIRLS-0 by replacing the randomized SVD with the default full SVD in MATLAB, as we observed that the latter one is much faster than the former one when the matrix size is small.

In our implementation of ASNN, we solved each subproblem (4.26) by using the proximal ADMM. Given that the proximal ADMM already introduces a

proximal term in each step, we simply set $\gamma^k \equiv 0$ in (4.26). To speed up the convergence, the subproblems (4.26) in the first few iterations were not solved with high accuracy, but with a moderate accuracy increased iteration by iteration. More specifically, let τ_{sub}^k be the tolerance in (4.58) for terminating the proximal ADMM in the k-th iteration. We set

$$\tau_{\text{sub}}^1 := 10^{-1} \quad \text{and} \quad \tau_{\text{sub}}^{k+1} := \max\left\{0.5\,\tau_{\text{sub}}^k, 10^{-4}\right\} \quad \forall\, k \geq 1.$$

One may take the first few iterations as finding a good starting point for ASNN. This strategy increases the number of outer iterations but reduces the total number of inner iterations as well as the number of SVDs. The main computational cost lies in computing a singular value decomposition (SVD). In our implementation, following [17, 168, 109], we used PROPACK [100] to compute a partial SVD whenever the matrix size is greater than 300 and the rank is recognized to be stable.

Tables 4.2 and 4.3 report the comparison among ASNN, IRLS-0 and sIRLS-0 for easy and hard test problems with different dimensions ($n_1 = n_2 = n$), ranks (r) and number of samples in terms of the sample ratio (**sr**), or alternatively, the ratio to the degree of freedom (α). The results of each test problem are reported to be the average of successful recoveries over 10 trials of random generations of the true matrix and its partial noiseless observations of entries. If no successful recovery was achieved, the results are reported to be that obtained at termination. In both Tables 4.2 and 4.3, **succ** means the number of successful recoveries; **iter** means the number of outer iterations and **initer** means the total number of inner iterations. The rank of a matrix was recognized under two different levels. The reported **rank1** and **rank2** refer to the numbers of singular values that are greater than 10^{-4} and 10^{-6} respectively. In addition, we also report the cumulative singular value residue (**sigres**), i.e., the sum of all singular values $\sigma_i(X)$ with i from $r+1$ to n. The computational time (**time**) is reported in seconds.

Table 4.2: Comparison of ASNN, IRLS-0 and sIRLS-0 on easy problems

Problem		Algorithm	succ	iter	initer	relerr	relfea	rank1|rank2	sigres	time
n	100	ASNN	10	7.8	55.8	4.70e-5	2.42e-5	10.0|10.0	1.41e-14	0.23
r	10	IRLS-0	10	47.1	300.9	9.53e-4	—	41.0|99.0	7.58e-3	0.22
sr	57%	sIRLS-0	10	51.8	—	9.40e-4	—	33.7|98.8	6.41e-3	0.16
α	3.00									
n	200	ASNN	10	7.6	58.9	5.34e-5	2.97e-5	10.0|10.0	2.24e-14	0.79
r	10	IRLS-0	10	50.7	363.2	9.67e-4	—	42.6|197.0	1.06e-2	0.93
sr	39%	sIRLS-0	10	58.0	—	9.57e-4	—	28.9|196.7	8.77e-3	0.68
α	4.00									
n	500	ASNN	10	8.0	76.5	6.05e-5	4.14e-5	10.0|10.0	5.06e-14	3.03
r	10	IRLS-0	10	70.8	616.3	9.65e-4	—	14.3|482.2	1.06e-2	8.44
sr	20%	sIRLS-0	10	90.4	—	9.71e-4	—	12.0|483.5	1.07e-2	6.86
α	5.05									
n	500	ASNN	10	10.0	164.5	1.80e-4	6.32e-5	10.0|10.0	5.50e-14	5.74
r	10	IRLS-0	10	221.5	1622.1	9.93e-4	—	11.5|466.0	5.28e-3	17.7
sr	12%	sIRLS-0	10	274.8	—	9.91e-4	—	10.8|469.8	5.62e-3	12.7
α	3.03									
n	1000	ASNN	10	8.4	110.3	5.83e-5	5.38e-5	10.0|10.0	1.02e-13	17.5
r	10	IRLS-0	10	103.7	943.8	9.76e-4	—	11.7|934.1	1.18e-2	42.9
sr	12%	sIRLS-0	10	135	—	9.82e-4	—	10.8|944.3	1.30e-2	22.6
α	6.03									
n	1000	ASNN	10	6.9	58.0	1.62e-5	1.03e-5	50.0|50.0	1.04e-13	19.4
r	50	IRLS-0	10	49.1	332.1	9.65e-4	—	257.1|991.2	6.00e-2	71.1
sr	39%	sIRLS-0	10	52.3	—	9.62e-4	—	214.6|990.2	5.54e-2	102.3
α	4.00									
n	1000	ASNN	10	9.0	141.2	9.62e-5	4.29e-5	20.0|20.0	1.04e-13	26.1
r	20	IRLS-0	10	180.0	1472.7	9.90e-4	—	20.5|957.5	1.39e-2	75.8
sr	12%	sIRLS-0	10	232.6	—	9.89e-4	—	20.2|939.3	1.42e-2	45.2
α	3.03									
n	2000	ASNN	10	8.0	122.1	3.51e-5	3.64e-5	20.0|20.0	2.08e-13	103.2
r	20	IRLS-0	10	92.6	897.7	9.74e-4	—	20.3|1907.3	2.87e-2	176.8
sr	12%	sIRLS-0	10	124.6	—	9.77e-4	—	20.0|1915.1	2.94e-2	131.3
α	6.03									
n	2000	ASNN	10	8.5	136.8	6.23e-5	2.91e-5	40.0|40.0	2.05e-13	141.9
r	40	IRLS-0	10	164.6	1406.8	9.84e-4	—	40.1|1924.7	2.98e-2	328.4
sr	12%	sIRLS-0	10	216.9	—	9.90e-4	—	40.0|1927.4	3.02e-2	258.0
α	3.03									

Table 4.3: Comparison of ASNN, IRLS-0 and sIRLS-0 on hard problems

Problem		Algorithm	succ	iter	initer	relerr	relfea	rank1|rank2	sigres	time
n	50	ASNN	6	31.8	552.3	8.88e-4	6.14e-5	5.0|5.2	2.66e-6	0.67
r	5	IRLS-0	4	2616.5	7664.0	9.99e-4	—	5.0|44.0	2.96e-4	2.53
sr	25%	sIRLS-0	1	2339.0	—	9.99e-4	—	5.0|45.0	3.30e-4	1.98
α	1.32									
n	100	ASNN	10	27.6	426.5	8.16e-4	8.37e-5	10.0|10.0	1.15e-14	1.57
r	10	IRLS-0	10	2028.4	6224.5	9.99e-4	—	10.0|90.2	6.69e-4	6.16
sr	25%	sIRLS-0	6	1858.2	—	9.99e-4	—	10.0|90.8	7.01e-4	5.17
α	1.32									
n	500	ASNN	10	31.0	1095.6	6.94e-4	7.51e-5	20.0|20.0	5.01e-14	44.7
r	20	IRLS-0	10	5019.2	16715.2	1.00e-3	—	20.0|407.1	1.66e-3	330.9
sr	10%	sIRLS-0	10	5039.7	—	1.00e-3	—	20.0|409.7	1.67e-3	287.9
α	1.28									
n	1000	ASNN	10	11.0	425.0	3.82e-4	7.52e-5	10.0|10.0	1.02e-13	60.0
r	10	IRLS-0	10	896.5	6595.1	9.98e-4	—	10.1|828.9	4.07e-3	268.0
sr	5%	sIRLS-0	10	1024.5	—	9.98e-4	—	10.2|834.7	4.13e-3	102.1
α	2.52									
n	1000	ASNN	10	39.4	2354.9	9.38e-4	5.44e-5	10.0|10.0	1.03e-13	256.5
r	10	IRLS-0	0	10000	36404.1	1.02e1	—	981.0|999.6	3.54e0	1400.4
sr	3%	sIRLS-0	0	10000	—	2.72e-3	—	10.1|583.3	1.82e-3	878.5
α	1.51									
n	1000	ASNN	10	10.0	209.3	1.52e-4	5.60e-5	20.0|20.0	1.03e-13	41.9
r	20	IRLS-0	10	331.6	2561.3	9.93e-4	—	20.0|949.0	1.12e-2	131.1
sr	10%	sIRLS-0	10	396.3	—	9.96e-4	—	20.1|948.9	1.11e-2	72.5
α	2.52									
n	1000	ASNN	10	30.0	846.8	4.91e-4	7.83e-5	20.0|20.0	1.03e-13	139.1
r	20	IRLS-0	6	2519.7	13547.3	9.99e-4	—	20.0|853.2	3.86e-3	648.5
sr	6%	sIRLS-0	10	2846.2	—	9.99e-4	—	20.0|828.8	3.46e-3	367.5
α	1.52									
n	1000	ASNN	10	31.1	3175.8	9.48e-4	8.20e-5	30.0|30.0	1.01e-13	601.5
r	30	IRLS-0	0	10000	34155.0	9.20e-1	—	976.2|999.6	2.48e0	2175.3
sr	7%	sIRLS-0	0	10000	—	1.63e-3	—	30.0|849.2	3.70e-3	1535.6
α	1.18									
n	2000	ASNN	10	31.0	1419.2	6.67e-4	8.09e-5	20.0|20.0	2.08e-13	938.8
r	20	IRLS-0	0	10000	40789.7	9.18e-1	—	1971|2000	8.06e0	5735.7
sr	3%	sIRLS-0	9	6184.9	—	1.00e-3	—	20.0|1300	3.58e-3	2102.2
α	1.51									

As can be observed from Table 4.2, for easy problems with high sample ratios, overall, the performances of ASNN, IRLS-0 and sIRLS-0 are comparable in terms of recovery error. Actually, due to a high sample ratio, the nuclear norm minimization already has enough recoverability for handling the recovery. That is why no much difference occurs in recovery among ASNN, IRLS-0 and sIRLS-0 since their advantages beyond the nuclear norm has not been involved yet. But one may still notice that ASNN appears more attractive than the other two when the sample ratio is relatively lower than the others. In Table 4.3 for hard problems with low sample ratios, the attraction of ASNN becomes more significant. We observe that ASNN substantially outperforms both IRLS-0 and sIRLS-0 in terms of high frequency of successful recovery, fewer iterations and less computational time. (The numbers of SVDs required in ASNN is equal to **initer**; while the number of SVDs required in IRLS-0 and sIRLS-0 is equal to **iter**.)

Another important observation is that **rank2** and **sigres** of ASNN is dramatically lower than that of IRLS-0 and sIRLS-0 for both easy and hard problems. This phenomenon is in concert with what we observed before in Figure 4.4 in the experiments of exact low-rank matrix recovery. This observation indicate that as a tool for "singular value filtering", ASNN possesses higher rank-promoting ability than that of IRLS-0 and sIRLS-0. Numerical results reported in Tables 4.2 and 4.3 validate our discussions in Subsection 4.5.1. We also point out that the performance of ASNN is much more stable than IRLS-0 and sIRLS-0 under different setting of parameters. In both IRLS-0 and sIRLS-0, the choice of the parameter ε is a critical issue in practical computations. For achieving the success of recovery, the parameter ε in IRLS-0 and sIRLS-0 is chosen to have an exponentially decay. Indeed, starting from a relatively larger ε serves for a quick decrease of the rank. However, due to the limited singular value filtering capability of the iterative scheme, ε has to be decreased to very small or tiny so that the weights of very

small singular values could be significantly larger than those of moderate singular values. Meanwhile, the delay speed of ε may heavily affect the performance. Differently, applying ASNN does not need to face such difficulty. The existence of the semi-nuclear norm endows the iterative scheme with strong singular value filtering capability so that one could simply fix a suitable ε to enjoy an exact low-rank solution.

4.6.3 Performance for finding a low-rank doubly stochastic matrix

In this subsection, we test the performance of our proposed adaptive semi-nuclear norm regularization approach for finding a low-rank doubly stochastic matrix based on a collection of observations of entries from a matrix which may or may not be of low rank or approximately low rank. A doubly stochastic matrix refers to a square matrix of nonnegative real numbers with each row and column summing to one. The goal of finding a doubly stochastic matrix is two-fold — having a low rank and deviating small from the given observations of entries. This consideration is a generalization of finding a nearest doubly stochastic matrix to a given square matrix that have been considered in the literature, e.g., see [68, 83, 8]. Some of entries could also be fixed according to possible available prior information. The problem of finding a low-rank doubly stochastic matrix was chosen to be a test problem in [81], formulated as a nuclear norm regularized least square problem.

In our experiments, we followed the setting in [81]. For each test problem, we generated a low-rank nonnegative matrix of the form UV^T with $U, V \in \mathbb{R}^{n\times r}$ drawn from i.i.d. uniform entries in $[0, 1]$ and then applied the Sinkhorn-Knopp algorithm [160] to convert it to a doubly stochastic matrix of rank r , denoted by M. We sampled partial entries of the generated matrix uniformly at random as

observations without noise or with i.i.d. Gaussian noise at noise level 10%. We let Ω_1 be the index set of observations and let $\widetilde{M}_{ij}$ be the noiseless or noisy observation of the (i,j)-th entry of M. We also randomly fixed a small number of entries of M to construct hard constraints and let Ω_2 be the corresponding index set (may be empty). The problem of finding a low-rank doubly stochastic matrix can be formulated as

$$\begin{aligned} \min \ & \frac{1}{2}\sum_{(i,j)\in\Omega_1}(X_{ij}-\widetilde{M}_{ij})^2+\rho\,\|F(X)\|_* \\ \text{s.t.} \ & X^Te=e, \quad Xe=e, \\ & X_{ij}=M_{ij}, \quad (i,j)\in\Omega_2, \\ & X\geq 0. \end{aligned} \tag{4.62}$$

We applied Algorithm 4.1 with $h^k \equiv h$ and $\mu^k \equiv f'_+(0)$ to the problem (4.62) and solved each subproblem (4.17) by using the proximal ADMM. Let $\tau^k_{\rm sub}$ be the tolerance in (4.58) for terminating the proximal ADMM in the k-th iteration. We set

$$\tau^1_{\rm sub} := 5\times 10^{-3} \quad \text{and} \quad \tau^{k+1}_{\rm sub} := \max\left\{0.8\,\tau^k_{\rm sub}, 10^{-5}\right\} \quad \forall\, k\geq 1.$$

We terminated Algorithm 4.1 by using the following stopping criterion:

$$\begin{cases} \mathrm{rank}(X^k)=\mathrm{rank}(X^{k-1})=\mathrm{rank}(X^{k-2}), \\ \textbf{relfea}^k<10^{-5} \ \text{ and } \ \textbf{reldif\_obj}^k<10^{-5}. \end{cases} \tag{4.63}$$

Tables 4.4 and 4.5 report the numerical comparison among the nuclear norm regularization approach (NN for short) and two adaptive semi-nuclear norm regularization approaches using the log function listed in Table 4.4 with different ε (ASNN1 and ASNN2 for short). For consistency of comparison, the parameters ρ for these three approaches were chosen such that $\rho\mu$ are equal to each other, where $\mu := 1$ in NN and $\mu := 1/\varepsilon$ in ASNN1 and ASNN2. Under this setting, with the initial point $X^0 = 0$, both ASNN1 and ASNN2 start from solving the nuclear norm regularized problem that is solved in NN. We set $\varepsilon := 0.05\,\sigma_1(X^1)$ in ASNN1

and $\varepsilon := 0.01\,\sigma_1(X^1)$ in ASNN2. For each test problem with different sample ratios (**sr**) and different numbers of fixed entries ($l = |\Omega_2|$), we report the number of iterations (**iter**: both the number of outer iterations and the total number of inner iterations), the value of the loss function (**loss**), the rank (**r**: the number of singular values greater than $10^{-6}\sigma_1(X)$), the relative primal feasibility (**feas**) and the running time (**time**: in seconds). In fact, under our particular experimental setting, finding a low-rank doubly stochastic matrix also fall into the category of matrix completion. Therefore, we also report the relative recovery error (**relerr**) for a reference.

As observed from Tables 4.4 and 4.5, for a smaller $\rho\mu$, both ASNN1 and ASNN2 outperform NN in terms of the solution quality — a smaller loss and a lower rank. For a larger $\rho\mu$, though NN, ASNN1 and ASNN2 produced the same rank due to the low-rank structure of the true matrix, ASNN1 and ASNN2 still lead to considerable decreases of loss. A notable observation is that the running time of the three approaches are overall comparable to each other. Furthermore, it is also notable that when a few entries are fixed, the total number of inner iterations and the running time grow greatly. This is because the additional hard constraints on partial entries largely increase the hardness of the problem so that the proximal ADMM has difficulty in achieving the target feasibility of constraints. We also comment here that Algorithm 4.2 with the subproblems solved by the smoothing Newton-BiCG method could be an alternative choice to achieve high feasibility for such hard problems. For a thorough comparison, we also run for each NN, ASNN1 and ASNN2 with different parameters ρ and plot three Loss vs. Rank paths in Figure 4.5. It is clear that ASNN2 shows the best performance in terms of the solution quality among the three algorithms under this experimental setting.

We also conducted another type of experiments in which the random matrix for observations of entries were drawn from i.i.d uniform entries in $[0, 2/n]$. Different

Table 4.4: Comparison of NN and ASNN with observations generated from a random low-rank doubly stochastic matrix without noise

Prob.		Alg.	$\rho\mu = 10^{-4}$						$\rho\mu = 10^{-3}$					
			iter	loss	r	feas	relerr	time	iter	loss	r	feas	relerr	time
n	500	NN	–|351	9.68e-7	64	2.63e-7	7.84e-2	49.6	–|168	7.21e-5	21	5.52e-7	8.14e-2	23.6
r	10	ASNN1	35|719	7.97e-7	35	4.26e-7	4.38e-2	89.4	22|126	4.80e-5	10	4.26e-6	7.15e-2	17.5
sr	5%	ASNN2	33|827	3.00e-7	10	8.90e-7	1.74e-2	110.3	19|124	1.64e-5	10	7.69e-6	5.78e-2	17.3
l	0													
n	500	NN	–|673	9.44e-7	63	6.72e-6	7.68e-2	87.8	–|1905	6.62e-5	23	1.00e-5	7.96e-2	255.9
r	10	ASNN1	34|1380	7.75e-7	35	5.76e-6	4.12e-2	170.8	27|1032	4.19e-5	10	9.01e-6	6.70e-2	139.2
sr	5%	ASNN2	33|1478	3.26e-7	10	6.24e-6	1.75e-2	191.2	29|836	9.74e-6	10	9.99e-6	4.57e-2	113.7
l	250													
n	1000	NN	–|404	1.14e-6	27	6.66e-7	1.08e-2	270.2	–|154	7.55e-5	10	1.85e-7	6.01e-2	104.6
r	10	ASNN1	28|315	3.90e-7	10	7.97e-7	5.69e-3	216.5	15|88	5.62e-5	10	4.90e-6	5.46e-2	61.7
sr	6%	ASNN2	24|281	8.06e-8	10	1.93e-6	3.06e-3	193.8	13|86	2.50e-5	10	7.26e-6	4.10e-2	60.2
l	0													
n	1000	NN	–|692	1.14e-6	33	3.69e-6	1.16e-2	465.6	–|2250	6.67e-5	10	6.21e-6	5.64e-2	1518.5
r	10	ASNN1	28|421	3.70e-7	10	2.24e-6	5.55e-3	287.8	26|1248	3.39e-5	10	8.51e-6	4.26e-2	828.1
sr	6%	ASNN2	29|460	9.63e-8	10	2.42e-6	3.37e-3	310.5	29|291	5.53e-6	10	1.00e-5	1.96e-2	201.5
l	1000													
n	1000	NN	–|136	2.47e-7	10	5.91e-7	1.72e-3	93.0	–|83	2.42e-5	10	3.26e-7	1.68e-2	58.5
r	10	ASNN1	19|123	7.59e-8	10	6.43e-6	9.40e-4	90.7	14|57	1.36e-5	10	9.78e-6	1.31e-2	42.0
sr	20%	ASNN2	18|132	1.12e-8	10	8.94e-6	3.56e-4	94.2	14|61	3.40e-6	10	9.23e-6	6.93e-3	44.8
l	0													
n	1000	NN	–|189	2.38e-7	10	8.96e-7	1.69e-3	128.9	–|423	2.37e-5	10	1.00e-5	1.66e-2	286.4
r	10	ASNN1	26|174	7.82e-8	10	1.24e-6	9.73e-4	125.1	16|61	1.14e-5	10	9.88e-6	1.19e-2	44.7
sr	20%	ASNN2	26|175	1.26e-8	10	1.45e-6	3.95e-4	124.5	15|63	2.77e-6	10	8.18e-6	6.23e-3	45.6
l	1000													
n	1000	NN	–|281	1.44e-6	55	6.90e-7	9.73e-3	191.5	–|91	8.80e-5	20	2.47e-7	4.94e-2	62.0
r	20	ASNN1	25|221	8.35e-7	20	1.64e-6	6.63e-3	154.8	15|52	8.13e-5	20	6.44e-6	4.85e-2	38.1
sr	10%	ASNN2	25|254	3.63e-7	20	1.54e-6	5.09e-3	176.2	14|57	4.02e-5	20	7.76e-6	3.78e-2	41.2
l	0													
n	1500	NN	–|167	5.19e-7	11	7.03e-7	3.75e-3	340.3	–|114	4.80e-5	10	2.08e-7	3.40e-2	231.7
r	10	ASNN1	19|144	1.62e-7	11	5.03e-6	2.26e-3	302.5	12|57	4.05e-5	10	9.93e-6	3.24e-2	121.5
sr	10%	ASNN2	19|149	2.69e-8	11	5.16e-6	1.09e-3	309.5	12|60	1.42e-5	10	9.46e-6	2.04e-2	126.7
l	0													

Table 4.5: Comparison of NN, ASNN1 and ASNN2 with observations generated from a random low-rank doubly stochastic matrix with 10% noise

Prob.		Alg.	$\rho\mu = 10^{-3}$						$\rho\mu = 3 \times 10^{-3}$					
			iter	loss	r	feas	relerr	time	iter	loss	r	feas	relerr	time
n	500	NN	–|175	9.22e-5	49	3.85e-7	1.03e-1	24.9	–|74	5.15e-4	7	4.62e-7	1.09e-1	10.5
r	10	ASNN1	26|273	7.21e-5	27	9.32e-7	1.04e-1	36.8	20|60	4.86e-4	7	3.50e-6	1.08e-1	9.0
sr	5%	ASNN2	30|573	2.46e-5	17	4.73e-7	1.23e-1	77.6	16|103	2.62e-4	7	8.64e-6	1.05e-1	14.4
l	0													
n	500	NN	–|1031	9.00e-5	49	9.98e-6	1.02e-1	138.5	–|2470	4.32e-4	18	1.00e-5	1.07e-1	322.6
r	10	ASNN1	29|1541	7.02e-5	24	9.50e-6	1.03e-1	206.5	29|3371	3.94e-4	7	1.00e-5	1.06e-1	454.1
sr	5%	ASNN2	29|2050	2.43e-5	17	9.98e-6	1.22e-1	275.6	27|2942	1.94e-4	7	8.86e-6	1.13e-1	395.6
l	250													
n	1000	NN	–|153	1.74e-4	80	4.29e-7	8.12e-2	103.5	–|62	6.84e-4	5	4.19e-7	1.12e-1	42.1
r	10	ASNN1	30|246	1.53e-4	50	4.89e-7	6.88e-2	174.0	15|30	6.82e-4	5	9.68e-6	1.12e-1	22.8
sr	6%	ASNN2	30|486	8.23e-5	26	5.12e-7	8.11e-2	330.8	16|99	4.74e-4	5	7.19e-6	9.48e-2	68.8
l	0													
n	1000	NN	–|1341	1.71e-4	80	1.00e-5	7.96e-2	903.5	–|4748	5.37e-4	17	6.32e-6	9.92e-2	3382.2
r	10	ASNN1	29|1453	1.51e-4	51	8.30e-6	6.73e-2	994.9	30|4584	5.18e-4	10	9.93e-6	9.61e-2	3219.7
sr	6%	ASNN2	29|1927	8.31e-5	26	9.99e-6	8.01e-2	1308.7	27|7223	3.07e-4	9	6.73e-6	7.23e-2	5091.0
l	1000													
n	1000	NN	–|153	2.34e-4	229	5.82e-7	5.43e-2	105.7	–|83	1.13e-3	10	1.37e-7	4.98e-2	59.9
r	10	ASNN1	23|147	2.12e-4	204	5.61e-7	5.18e-2	108.1	14|53	1.03e-3	10	4.08e-6	4.00e-2	39.8
sr	20%	ASNN2	34|568	1.46e-4	120	1.65e-7	6.66e-2	398.6	16|69	9.37e-4	10	2.40e-6	3.10e-2	52.0
l	0													
n	1000	NN	–|785	2.34e-4	228	1.00e-5	5.37e-2	538.1	–|3082	1.11e-3	23	7.75e-6	4.79e-2	2152.7
r	10	ASNN1	29|790	2.13e-4	201	1.00e-5	5.14e-2	542.8	29|2253	1.00e-3	10	9.01e-6	3.66e-2	1548.5
sr	20%	ASNN2	41|1699	1.46e-4	116	7.84e-6	6.71e-2	1176.2	29|1995	9.34e-4	10	1.00e-5	3.00e-2	1455.8
l	1000													
n	1000	NN	–|142	1.97e-4	136	5.15e-7	6.93e-2	98.3	–|42	8.01e-4	1	2.38e-6	7.72e-2	29.8
r	20	ASNN1	30|226	1.77e-4	105	5.96e-7	6.91e-2	160.7	15|29	8.00e-4	2	4.81e-6	7.72e-2	22.6
sr	10%	ASNN2	40|1011	1.01e-4	48	5.51e-7	8.70e-2	697.9	15|29	7.99e-4	3	4.83e-6	7.71e-2	23.0
l	0													
n	1500	NN	–|177	2.75e-4	155	1.28e-7	5.38e-2	364.8	–|68	8.28e-4	10	4.21e-7	8.61e-2	147.5
r	10	ASNN1	29|213	2.55e-4	124	1.31e-7	4.52e-2	465.4	11|42	7.95e-4	10	6.35e-6	8.30e-2	95.1
sr	10%	ASNN2	34|558	1.84e-4	59	1.43e-7	5.53e-2	1191.7	14|73	4.89e-4	10	9.58e-6	4.21e-2	165.0
l	0													

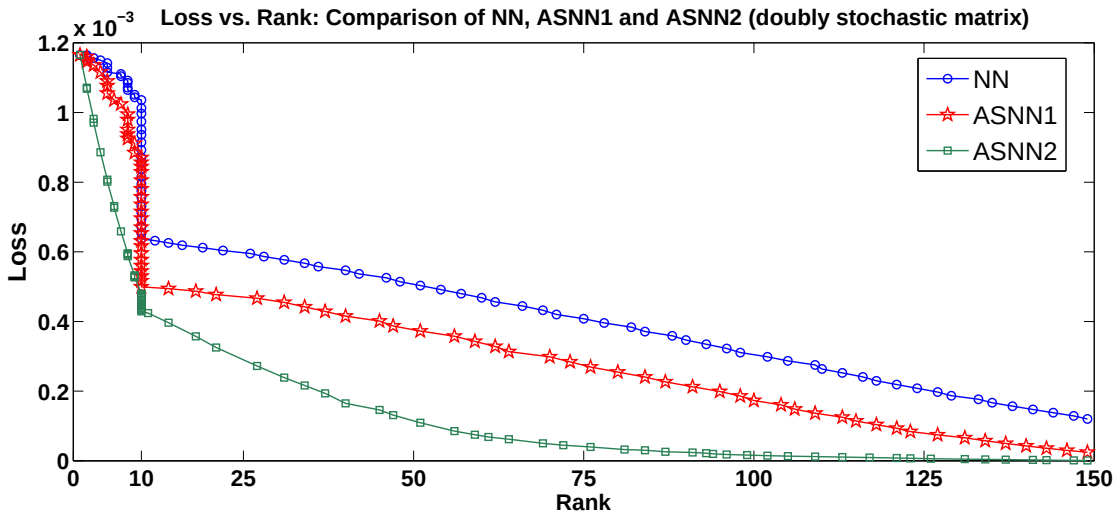


Figure 4.5: Loss vs. Rank: Comparison of NN, ASNN1 and ASNN2 with observations generated from a low-rank doubly stochastic matrix with noise ($n = 1000, r = 10,$ noise level $= 10\%,$ sample ratio $= 10\%$)

from the pervious type of experiments, the matrix generated in this way is only an approximate doubly stochastic matrix without any artificial low-rank structure. (Actually, due to the nonnegative restriction of entries, the largest singular value of the random matrix generated in either of these two ways dominates the other singular values so that the matrix has the approximate rank-1 structure to some extent in most cases.) Table 4.6 reports the comparison of NN, ASNN1 and ASNN2 for finding a low-rank doubly stochastic matrix under this setting. Besides the final results, we also report the intermediate results of ASNN1 and ASNN2 (in italic type) after a similar running time of NN.

We can see that both ASNN1 and ASNN2 outperform NN in terms of solution quality. But, due to the hardness of problems in this type of experiments, reducing the loss and the rank heavily conflicts with maintaining the feasibility of

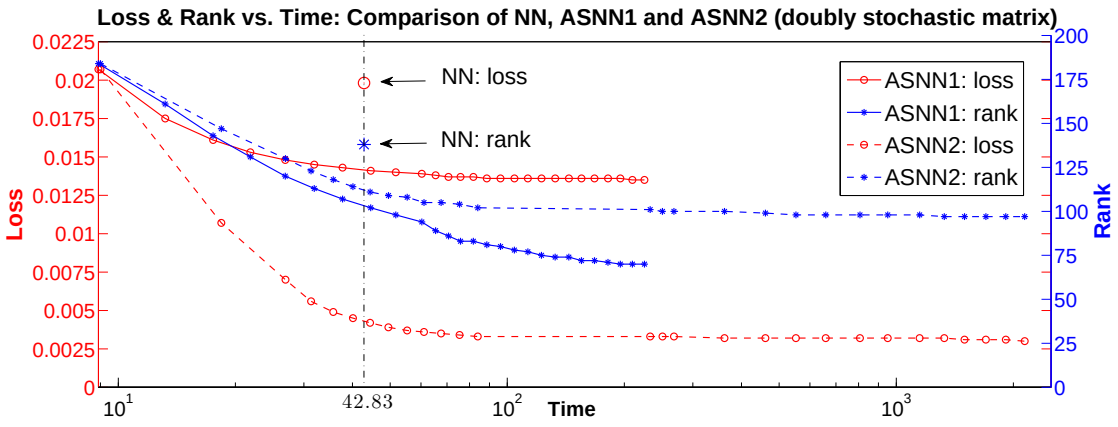


Figure 4.6: Loss & Rank vs. Time: Comparison of NN, ASNN1 and ASNN2 with observations generated from an approximate doubly stochastic matrix ($n = 1000, r = 10,$ sample ratio $= 20\%$)

constraints, especially the nonnegative constraints of entries. Therefore, comparing the intermediate and final results reported for each test problem in Table 4.6, we find that ASNN1 and ASNN2 spend much time on harmonizing this conflict for producing a solution with the loss and the rank being lower than that of NN. This can be found more clearly in Figure 4.6, which records the whole iterative processes of NN, ASNN1 and ASNN2 for the test problem with $n = 1000$ and sample ratio $= 20\%$ listed in Table 4.6. It is easy to see from Figure 4.6 that diminishing marginal utility of ASNN1 and ASNN2 are very obvious. Therefore, users should take into account the tradeoff between the solution quality and the computational time to make a decision, according to their demands.

Table 4.6: Comparison of NN, ASNN1 and ASNN2 with observations generated from an approximate doubly stochastic matrix ($\rho\mu = 10^{-2}$, no fixed entries)

sr	Alg.	$n = 500$					$n = 1000$				
		iter	loss	r	feas	time	iter	loss	r	feas	time
10%	NN	–|102	9.08e-3	50	4.11e-7	14.2	–|65	1.51e-2	36	4.53e-7	46.7
	ASNN1	27|1432	3.89e-3	20	7.19e-6	204.1	19|174	1.04e-2	20	4.34e-6	123.0
	intermed.	*10|90*	*4.48e-3*	*29*	*2.99e-5*	*13.3*	*10|65*	*1.18e-2*	*28*	*2.52e-5*	*46.6*
	ASNN2	28|3970	6.39e-4	31	8.33e-6	561.9	27|4625	2.81e-3	38	9.48e-6	3108.0
	intermed.	*11|94*	*7.93e-4*	*40*	*1.58e-4*	*13.9*	*5|69*	*5.45e-3*	*38*	*1.27e-4*	*49.0*
20%	NN	–|88	1.12e-2	101	4.67e-7	12.4	–|61	1.98e-2	138	4.42e-7	42.8
	ASNN1	28|1909	5.61e-3	47	9.58e-6	241.9	27|307	1.35e-2	70	3.76e-6	221.1
	intermed.	*10|83*	*5.89e-3*	*63*	*5.14e-5*	*11.9*	*8|61*	*1.41e-2*	*102*	*1.16e-5*	*44.9*
	ASNN2	28|3127	9.74e-4	61	9.82e-6	391.5	28|3050	3.05e-3	97	9.65e-6	2138.7
	intermed.	*10|84*	*1.19e-4*	*75*	*2.87e-6*	*11.9*	*7|61*	*4.19e-3*	*111*	*1.70e-4*	*44.1*
50%	NN	–|62	1.55e-2	218	5.94e-7	9.8	–|43	2.82e-2	360	6.43e-7	32.3
	ASNN1	29|933	8.79e-3	157	9.69e-6	141.2	29|484	1.98e-2	285	8.95e-6	353.6
	intermed.	*9|55*	*8.85e-3*	*176*	*1.44e-4*	*9.1*	*7|41*	*2.00e-2*	*314*	*2.42e-5*	*33.0*
	ASNN2	34|3722	2.16e-3	148	9.59e-6	560.7	29|3017	6.39e-3	253	8.47e-6	2183.3
	intermed.	*8|59*	*2.35e-3*	*174*	*5.64e-4*	*9.6*	*6|42*	*7.07e-3*	*297*	*3.10e-4*	*33.9*
100%	NN	–|22	2.16e-2	378	5.34e-7	4.1	–|22	3.87e-2	655	3.92e-7	17.9
	ASNN1	28|87	1.32e-2	378	6.50e-6	15.4	10|21	2.77e-2	655	7.78e-6	20.4
	intermed.	*9|20*	*1.33e-2*	*378*	*1.78e-5*	*4.0*	*8|19*	*2.78e-2*	*655*	*1.79e-5*	*18.0*
	ASNN2	29|376	5.42e-3	378	9.96e-6	63.6	29|641	1.25e-2	655	9.98e-6	511.2
	intermed.	*7|21*	*5.66e-3*	*378*	*7.44e-5*	*4.0*	*5|21*	*1.32e-2*	*655*	*5.05e-5*	*18.4*

4.6.4 Performance for finding a reduced-rank transition matrix

In this subsection, we test the performance of the adaptive semi-nuclear norm regularization approach for finding a reduced-rank transition matrix. A transition matrix is a square matrix that describes the probability of moving from one state to another in a dynamic system. Entries in each row represent the probabilities of moving from the state represented by that row to the other states. Thus, the sum of each row of a transition matrix is equal to one. Finding a reduced-rank approximation of a transition matrix has been considered by Lin [107] using the Latent Markov Analysis (LMA) approach. One of its applications is for computing the personalized PageRank, e.g. Google PageRank, which describes the backlink-based page quality around user-selected pages [12]. This problem was chosen to be a test problem in [81] recently after being formulated to be a nuclear norm regularized least squares problem.

In our experiments, we follow the setting in [81]. Given a set of n web pages as a directed graph whose nodes represent the web pages and edges represent the links between pages, let $\deg(i)$ denote the outdegree of Page i, i.e., the number of pages that can be reached by a direct line from Page i, excluding the self-referential links. Let $P \in \mathbb{R}^{n\times n}$ be the transition matrix whose (i,j)-th entriy describes the transition probability from Page i to Page j. If $\deg(i) > 0$, then $P_{ij} = 1/\deg(i)$ if there is a link from Page i to Page j and $P_{ij} = 0$ otherwise. If $\deg(i) = 0$, then we assume the uniform selection, i.e., $P_{ij} = 1/n$. In order to handle the convergence issue of the power method for computing the PageRank [141], the standard way is to replace the transition matrix with

$$P_c = cP + (1-c)ev^T,$$

where $c \in (0,1)$, $e = (1,\ldots,1) \in \mathbb{R}^n$ and $v \in \mathbb{R}^n$ is a probability vector, i.e., $v \geq 0$

and $e^T v = 1$. As what was did [81], we chose $c = 0.85$, $v_i = 1/n, i = 1, \ldots, n$ and added 10% i.i.d Gaussian noise to P_c, termed as $\widetilde{P}_c$.

In our experiments, we aim to find a reduced-rank transition matrix by solving the following optimization problem :

$$\begin{aligned} \min \ & \frac{1}{2}\|X - \widetilde{P}_c\|_F^2 + \rho \, \|F(X)\|_* \\ \text{s.t.} \ & Xe = e, \quad X \geq 0. \end{aligned} \tag{4.64}$$

The data for our experiments are identical to that used in [81], including "Harvard500" (available at http://www.mathworks.com/moler/numfilelist.html), "NUS500", "NUS1000", "NUS1500" (collected by Kaifeng Jiang, generated from a portion of web page starting at the root page http://www.nus.edu.sg), "automobile industries", "computational complexity", "computational geometry" and "randomized algorithms" (collected by Panayiotis Tasparas, available at http://www.cs.toronto.edu/~tsap/experiments/download/download.html).

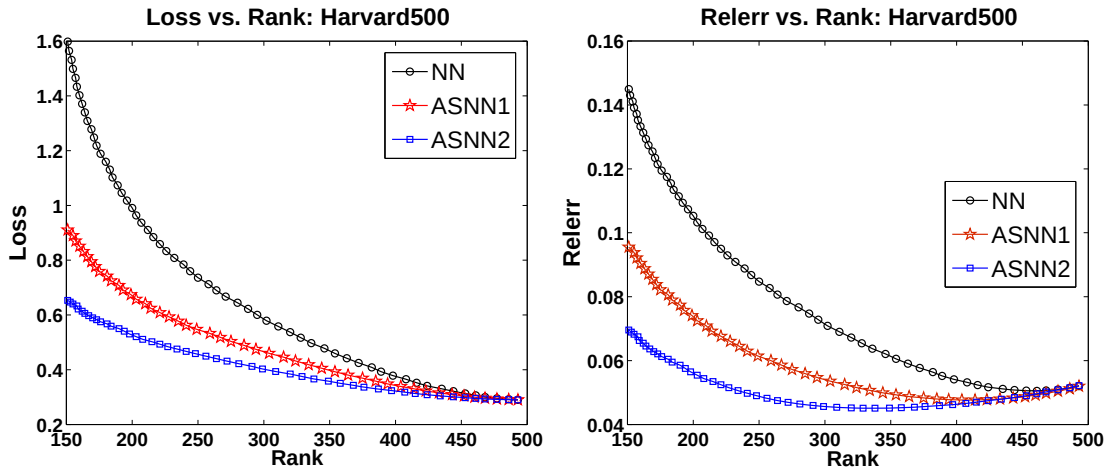
We applied Algorithm 4.1 with $h^k \equiv h$ and $\mu^k \equiv f'_+(0)$ to the problem (4.64) and solved each subproblem (4.17) by using the smoothing Newton-BiCG method. Due to the Euclidean distance loss function in the objective, there is no need to majorize the loss function so that the smoothing Newton-CG method is more appropriate than the proximal ADMM described in Subsection 4.6.3 as it can lead to high feasibility more easily because of its quadratic convergence under certain conditions. We solved the subproblems with the tolerance 10^{-5} and terminated Algorithm 4.1 by using the same stopping criterion (4.63) in Subsection 4.6.3.

The numerical results of NN, ASNN1 and ASNN2 on the data mentioned above are reported in Table 4.7, in which all the abbreviations mean the same as that in Subsection 4.6.3. Recall that given the same $\rho\mu$, both ASNN1 and ASNN2 start from solving a nuclear norm regularized problem solved in NN. (The first choice $\rho\mu = 5 \times 10^{-3} \, \|\widetilde{P}_c\|_F$ in Table 4.7 is consistent with that chosen in [81].) In Table 4.7, substantial decreases of the loss as well as the recovery error are found in

Table 4.7: Comparison of NN, ASNN1 and ASNN2 for finding a reduced-rank transition matrix

Prob.	Alg.	$\rho\mu = 5 \times 10^{-3} \|\widetilde{P}_c\|_F$						$\rho\mu = 1.25 \times 10^{-2} \|\widetilde{P}_c\|_F$					
		iter	loss	r	feas	relerr	time	iter	loss	r	feas	relerr	time
Harv.500	NN	–|7	4.36e-1	366	5.97e-6	5.86e-2	5.2	–|11	9.69e-1	202	8.41e-6	1.04e-1	11.0
$n = 500$	ASNN1	4|14	3.76e-1	369	1.44e-7	4.86e-2	10.0	4|22	6.45e-1	208	5.82e-6	7.12e-2	21.1
$r = 218$	ASNN2	6|19	3.40e-1	372	8.03e-9	4.55e-2	15.3	6|50	5.05e-1	214	3.92e-6	5.37e-2	89.5
NUS500	NN	–|8	1.35e-1	384	2.47e-6	5.70e-2	6.3	–|9	2.81e-4	210	3.34e-6	7.75e-2	7.0
$n = 500$	ASNN1	4|15	1.20e-1	386	9.10e-8	5.29e-2	11.2	4|18	2.10e-1	215	4.00e-7	5.64e-2	13.5
$r = 225$	ASNN2	5|17	1.09e-1	388	4.38e-8	5.47e-2	13.0	6|23	1.73e-1	221	3.64e-7	4.84e-2	18.6
NUS1000	NN	–|12	3.88e-1	657	3.10e-6	5.63e-2	59.3	–|15	8.30e-1	249	9.26e-6	9.64e-2	75.9
$n = 1000$	ASNN1	4|19	3.45e-1	663	5.34e-7	4.82e-2	93.8	6|32	6.12e-1	259	2.73e-6	7.04e-2	156.0
$r = 466$	ASNN2	5|21	3.07e-1	672	2.86e-6	4.59e-2	107.6	6|45	5.00e-1	268	5.77e-6	5.45e-2	305.4
NUS1500	NN	–|13	5.74e-1	957	1.11e-6	6.35e-2	181.7	–|25	1.27e0	357	9.10e-6	1.14e-1	502.0
$n = 1000$	ASNN1	4|20	5.05e-1	966	9.67e-6	5.37e-2	280.3	4|41	9.25e-1	368	3.85e-6	8.64e-2	801.5
$r = 807$	ASNN2	5|24	4.43e-1	982	1.78e-6	4.77e-2	326.1	9|66	7.39e-1	384	8.79e-6	6.82e-2	1312.2
Rand.Alg.	NN	–|11	7.87e-1	631	3.09e-6	4.48e-2	26.3	–|13	1.19e0	443	5.90e-6	6.03e-2	34.2
$n = 742$	ASNN1	4|18	7.40e-1	633	2.45e-7	4.06e-2	42.5	4|22	9.27e-1	450	8.30e-7	3.84e-2	63.9
$r = 216$	ASNN2	4|18	7.21e-1	636	1.01e-6	4.19e-2	42.5	6|28	8.49e-1	459	9.01e-7	3.67e-2	76.7
Complex.	NN	–|11	7.98e-1	711	3.67e-6	4.77e-2	42.5	–|14	1.39e0	440	4.68e-6	6.93e-2	57.9
$n = 884$	ASNN1	4|19	7.32e-1	715	7.66e-8	4.22e-2	67.6	7|32	1.02e0	448	4.95e-7	4.24e-2	142.3
$r = 255$	ASNN2	4|18	7.01e-1	719	1.53e-6	4.31e-2	67.5	8|37	9.01e-1	458	8.45e-7	3.74e-2	179.1
Auto.Ind.	NN	–|10	1.21e0	844	2.06e-6	4.05e-2	79.3	–|32	2.05e0	324	9.21e-6	6.67e-2	543.1
$n = 1196$	ASNN1	4|18	1.24e0	849	9.30e-7	3.42e-2	129.3	21|91	1.61e0	340	1.49e-6	4.14e-2	993.2
$r = 206$	ASNN2	4|18	1.07e0	858	7.08e-7	3.41e-2	135.7	9|67	1.47e0	348	3.11e-6	3.06e-2	842.0
Geometry	NN	–|10	9.18e-1	1019	1.66e-6	4.67e-2	81.3	–|13	1.52e0	682	9.24e-6	6.77e-2	130.2
$n = 1226$	ASNN1	4|18	8.50e-1	1021	2.69e-7	4.14e-2	140.5	4|22	1.13e0	694	9.33e-6	4.26e-2	204.8
$r = 416$	ASNN2	5|19	8.20e-1	1026	2.43e-7	4.21e-2	150.1	7|30	1.00e0	709	8.33e-6	3.77e-2	319.6

ASNN1 and ASNN2 compared with NN, in spite of a slight increase of the rank. To make a more clear vision, we also plot the Loss vs. Rank paths and the Relerr vs. Rank paths for NN, ASNN1 and ASNN2 in Figure 4.7 for comparison on the data "Harvard500". We can see that the solution qualities of ASNN1 and ASNN2 are substantially better than that of NN, especially when the rank attained is small.



(a) Loss vs. Rank

(b) Relerr vs. Rank

Figure 4.7: Loss vs. Rank and Relerr vs. Rank: Comparison of NN, ASNN1 and ASNN2 for finding a reduced-rank transition matrix on the data "Harvard500"

4.6.5 Performance for large noisy matrix completion with hard constraints

In this subsection, we test the performance of our proposed adaptive semi-nuclear norm regularization approach for large noisy matrix completion problems (of dimension at least 5000) with hard constraints. In our experiments, we created the true matrix $\overline{X} \in \mathbb{R}^{n\times n}$ as the product of two random $n \times r$ matrices with i.i.d. standard Gaussian entries. We sampled partial entries uniformly at random, most

of which were used as observations with 10% i.i.d. Gaussian noises and the others were used as fixed entries to construct equality constraints. Besides, we further sampled partial entries uniformly at random and identified their signs to construct inequalities constraints.

Let Ω_0 and Ω_i, $i = 1, 2, 3$ denote the set of indices of observed entries, fixed entries, identified non-negative entries and identified non-positive entries, respectively. Let $\widetilde{X}_{ij}$ denote the noisy observation of the (i, j)-th entry of $\overline{X}$. Then, the optimization problem to be solved for recovering the unknown low-rank matrix $\overline{X}$ can be formulated as

$$\begin{aligned} \min \ & \frac{1}{2} \sum_{(i,j)\in\Omega_0} (X_{ij} - \widetilde{X}_{ij})^2 + \rho \, \|F(X)\|_* \\ \text{s.t.} \ & X_{ij} = \overline{X}_{ij}, \quad (i,j) \in \Omega_1, \\ & X_{ij} \geq 0, \quad (i,j) \in \Omega_2, \\ & X_{ij} \leq 0, \quad (i,j) \in \Omega_3. \end{aligned} \tag{4.65}$$

We applied Algorithm 4.1 with $h^k \equiv h$ and $\mu^k \equiv f'_+(0)$ to the problem (4.65) and solved the each subproblem (4.17) by using the proximal ADMM. Given that the full SVD is not suitable to be applied for large matrices, we used PROPACK [100] to compute a partial SVD instead to reduce the computational time. Let sv^k be the number of singular values to be calculated in the k-th iteration. In our experiments, we simply set

$$\mathrm{sv}^1 = 1 \quad \text{and} \quad \mathrm{sv}^{k+1} = \min\left\{\mathrm{rank}(X^k) + 1, \mathrm{sv}_{\max}\right\} \quad \forall\, k \geq 1$$

where $\mathrm{sv}_{\max}$ is the maximum number of singular values allowed for partial SVDs. We set $\mathrm{sv}_{\max} = 100$ when $n \leq 10000$ and $\mathrm{sv}_{\max} = 50$ when $n > 10000$. These upper bounds can be regarded as the prior information for the rank of the unknown matrix. Let τ^k_{sub} be the tolerance in (4.58) for terminating the proximal ADMM in

the k-th iteration. We set

$$\tau_{\text{sub}}^1 := 5 \times 10^{-3} \quad \text{and} \quad \tau_{\text{sub}}^{k+1} := \max\left\{0.5\,\tau_{\text{sub}}^k, 10^{-5}\right\} \quad \forall\, k \geq 1.$$

We terminated Algorithm by using the same stopping criterion (4.63) in Subsection 4.6.3.

Table 4.8 reports the comparison of NN and ASNN1 for large noisy matrix completion with hard constraints, in which all the abbreviations mean the same as that in Subsection 4.6.3. In addition, α denotes the ratio of the total number of observed entries over the degree of freedom, i.e., $\alpha = |\Omega_0|/(r(n_1 + n_2 - r))$, l_{eq} denotes the number of equality constraints for fixed entries, i.e., $l_{\text{eq}} = |\Omega_1|$, and l_{ineq} denotes the number of inequality constraints for identified non-negative and non-positive entries, i.e., $l_{\text{ineq}} = |\Omega_2| + |\Omega_3|$. Table 4.8 indicates that the strategy for partial SVDs described above works so that in general large matrix completion problems with hard constraint (of good conditions) can be solved within a reasonable time. As can be observed, the advantage of ASNN1 over NN is quite apparent, not only in the solution quality, but also in the computational time for two main reasons. First, solving the subproblems in the first few iterations with moderate accuracy rather than high accuracy is a very efficient strategy. Second, the merit of the partial SVD in saving the computational time is more fully utilized since ASNN1 possesses more low-rank promoting ability. We also remark that the computational time of both NN and ASNN1 can be much reduced if $\text{sv}_{\max}$ is set to be smaller but appropriate. The rank truncation technique in [168] can also be applied to accelerated the convergence, especially when ρ is small. But the use of this technique should be careful since it may lead to erroneous results in certain instances, especially when the singular values of the matrix to be recovered are not clustered.

Table 4.8: Comparison of NN and ASNN1 for large matrix completion problems with hard constraints (noise level = 10%)

Problem					Algrithm	$\rho\mu = 10^{-1}\|\mathcal{A}^*(b)\|_2$					
n	r	α	l_{eq}	l_{ineq}		iter	loss	r	feas	relerr	time
5000	10	6	0	10000	NN	–|277	7.10e4	10	9.90e-6	1.54e-1	109.2
					ASNN1	10|355	2.52e4	10	9.94e-6	4.53e-2	129.8
		6	10000	0	NN	–|371	6.47e4	12	9.89e-6	1.42e-1	206.4
					ASNN1	10|412	2.52e4	10	9.93e-6	4.28e-2	140.5
		6	10000	9845	NN	–|372	6.47e4	12	1.00e-5	1.42e-1	215.8
					ASNN1	10|411	2.52e4	10	9.97e-6	4.28e-2	142.6
		6	100000	99576	NN	–|668	2.99e4	100	1.00e-5	9.25e-2	1445.0
					ASNN1	11|1214	2.37e4	55	1.00e-5	1.43e-2	2261.9
		3	10000	0	NN	–|902	3.26e4	100	9.98e-6	2.63e-1	1743.0
					ASNN1	10|1072	1.00e4	10	9.95e-6	6.87e-2	416.0
	50	5	10000	10000	NN	–|168	1.55e6	50	9.86e-6	1.61e-1	259.1
					ASNN1	10|210	5.00e5	50	9.86e-6	4.94e-2	351.8
10000	10	6	10000	0	NN	–|536	1.36e5	10	9.93e-6	1.49e-1	699.2
					ASNN1	10|634	5.03e4	10	9.99e-6	4.42e-2	514.4
		4	100000	100322	NN	–|2286	5.48e4	100	9.99e-6	1.69e-1	9397.1
					ASNN1	10|1639	3.01e4	15	9.97e-6	4.12e-2	2863.1
20000	10	6	20000	19738	NN	–|1193	2.65e5	10	9.95e-6	1.46e-1	2797.0
					ASNN1	11|1349	1.00e5	10	9.93e-6	4.42e-2	2357.0
50000	10	6	50000	0	NN	–|2331	6.59e5	11	9.97e-6	1.46e-1	18808.7
					ASNN1	11|2669	2.50e5	10	9.99e-6	4.42e-2	17746.2

Chapter 5

Conclusions and discussions

In this thesis, we proposed the semi-nuclear norm technique to address optimization problems with low-rank structures. Applying this novel technique yields a rank-corrected procedure for low-rank matrix completion with fixed basis coefficients and an adaptive semi-nuclear regularization approach for rank regularized problems with hard constraints. The introduced concept — semi-nuclear norm consists of the nuclear norm and a linear term. A proper semi-nuclear norm can possess significantly high rank-promoting ability beyond the reach of the widely-used nuclear norm. Thanks to a semi-nuclear norm, the rank-correction step for matrix completion with fixed basis coefficients produces an estimator of high accuracy and low rank. For this new estimator, we established non-asymptotic recovery error bounds and provided necessary and sufficient conditions for rank consistency in the sense of Bach [7]. The obtained results in these two aspects yield a consistent criterion for constructing a suitable rank-correction term (or semi-nuclear norm). For rank regularized problems, the adaptive semi-nuclear norm regularization approach iteratively solves a sequence of convex optimization problems, in which the objective functions are regularized by self-adjusted semi-nuclear norms. Each subproblem can be efficiently solved by recently developed methodologies.

Our proposed approach using adaptive semi-nuclear norms overcomes the difficultly in extending the reweighted trace minimization of Fazel, Hindi and Boyd [51] for rank minimization from the positive semidefinite case to the rectangular case. Again thanks to semi-nuclear norms, the iterative scheme of this approach has the advantages of achieving both high computational efficiency and the low-rank structure-preserving ability.

Many crucial issues of optimization problems with low-rank structures are still far from being settled. To conclude this thesis, we list some challenges that need to be explored in the further work.

- We believe that $a_m/b_m < 1$ should be guaranteed for the nuclear norm penalized least squares estimator with a reasonably small number of samples so that the error reduction of the rank-correction step could be confirmative. This problem is of paramount importance for practical applications.

- How to extend the rank consistency results for our proposed rank-correction step for matrix completion with noise to the high dimensional setting?

- How to establish a theoretical guarantee for the minimum-rank solution of our proposed adaptive semi-nuclear norm minimization for rank minimization problems?

- How to efficiently deal with large-scale low-rank optimization problems involving a large number of hard constraints that need to be satisfied with high accuracy?

- As the rank of a matrix reveals the relation between rows or columns more straightforward than singular values, it could be worthwhile looking for other surrogates of the rank function in this direction so that the costly SVDs could be avoided in computations even when hard constraints are involved.

Bibliography

[1] B.P.W. Ames and S.A. Vavasis. Nuclear norm minimization for the planted clique and biclique problems. *Mathematical programming*, 129(1):69–89, 2011. 2

[2] P.K. Andersen and R.D. Gill. Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, 10(4):1100–1120, 1982. 31, 32

[3] V.I. Arnold. On matrices depending on parameters. *Russian Mathematical Surveys*, 26(2):29–43, 1971. 73, 78

[4] H. Attouch. *Variational Convergence for Functions and Operators*. Applicable Mathematics Series. Pitman (Advanced Publishing Program), Boston, MA, 1984. 29

[5] H. Attouch and R.J.B. Wets. Approximation and convergence in nonlinear optimization. *Nonlinear programming*, 4:367–394, 1981. 28

[6] A. Auslender. Minimisation de fonctions localement lipschitziennes: applications a la programmation mi-convexe, mi-differentiable. *Nonlinear programming*, 3:429–460, 1981. 35, 36

[7] F.R. Bach. Consistency of trace norm minimization. *Journal of Machine Learning Research*, 9:1019–1048, 2008. ix, 5, 7, 10, 11, 43, 65, 67, 68, 172

[8] Z.J. Bai, D. Chu, and R.C.E. Tan. Computing the nearest doubly stochastic matrix with a prescribed entry. *SIAM Journal on Scientific Computing*, 29(2):635–655, 2007. 157

[9] L. Balzano, R. Nowak, and B. Recht. Online identification and tracking of subspaces from highly incomplete information. In *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, pages 704–711. IEEE, 2010. 1

[10] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009. 88

[11] M.P. Becker, I. Yang, and K. Lange. EM algorithms without missing data. *Statistical Methods in Medical Research*, 6(1):38–54, 1997. 33

[12] A.A. Benczúr, K. Csalogány, and T. Sarlós. On the feasibility of low-rank approximation for personalized pagerank. In *Special interest tracks and posters of the 14th international conference on World Wide Web*, pages 972–973. ACM, 2005. 165

[13] R. Bhatia. *Matrix Analysis*, volume 169. Springer Verlag, 1997. 22

[14] G. Birkhoff. Tres observaciones sobre el algebra lineal. *Univ. Nac. Tucumán. Revista A.*, 5:147–151, 1946. 16

[15] J.F. Bonnans and A. Shapiro. *Perturbation Analysis of Optimization Problems*. Springer Verlag, 2000. 73, 76

[16] P. Bühlmann and S. Van De Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer-Verlag New York Inc, 2011. 56

[17] J.F. Cai, E.J. Candès, and Z. Shen. A Singular Value Thresholding Algorithm for Matrix Completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010. 5, 19, 123, 153

[18] T.T. Cai and Z. Anru. Sharp RIP bound for sparse signal and low-rank matrix recovery. *Applied and Computational Harmonic Analysis*, 2012. 5

[19] E.J. Candès and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010. 6, 48

[20] E.J. Candès and Y. Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *Information Theory, IEEE Transactions on*, 57(4):2342–2359, 2011. 4

[21] E.J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009. 6, 48

[22] E.J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *Information Theory, IEEE Transactions on*, 56(5):2053–2080, 2010. 6, 48

[23] E.J. Candès, M.B. Wakin, and S.P. Boyd. Enhancing sparsity by reweighted l_1 minimization. *Journal of Fourier Analysis and Applications*, 14(5):877–905, 2008. 133

[24] R. Chartrand. Exact reconstruction of sparse signals via nonconvex minimization. *Signal Processing Letters, IEEE*, 14(10):707–710, 2007. 128

[25] R. Chartrand and V. Staneva. Restricted isometry properties and nonconvex compressive sensing. *Inverse Problems*, 24:035020, 2008. 128

[26] R. Chartrand and W. Yin. Iteratively reweighted algorithms for compressive sensing. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 3869–3872. IEEE, 2008. 133

[27] F. H. Clarke. *Optimization and Nonsmooth Analysis*, volume 5 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second edition, 1990. 20, 119

[28] G. Dal Maso. *An Introduction to* Γ*-convergence*. Progress in Nonlinear Differential Equations and their Applications, 8. Birkhäuser Boston Inc., Boston, MA, 1993. 29

[29] I. Daubechies, R. DeVore, M. Fornasier, and C.S. Güntürk. Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics*, 63(1):1–38, 2009. 133

[30] E. De Giorgi and T. Franzoni. Su un tipo di convergenza variazionale. *Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur. (8)*, 58(6):842–850, 1975. 28

[31] J. De Leeuw and W.J. Heiser. Convergence of correction matrix algorithms for multidimensional scaling. *Geometric representations of relational data*, pages 735–752, 1977. 33

[32] J. De Leeuw and G. Michailidis. Block relaxation algorithms in statistics. *Information systems and data analysis*, pages 308–325, 1994. 33

[33] C. Ding. *An introduction to a class of matrix optimization problems*. PhD thesis, National University of Singapore, 2012. 17, 18, 120

[34] C. Ding, D.F. Sun, and K.C. Toh. An introduction to a class of matrix cone programming. *Mathematical Programming, to appear*. 68, 71, 72, 75, 84

[35] T. Ding, M. Sznaier, and O.I. Camps. A rank minimization approach to video inpainting. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007. 2

[36] D.L. Donoho and J.M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994. 128

[37] K. Dvijotham and M. Fazel. A nullspace analysis of the nuclear norm heuristic for rank minimization. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 3586–3589. IEEE, 2010. 4

[38] J. Eckstein and D.P. Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1):293–318, 1992. 144

[39] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004. 10

[40] Y.M. Ermoliev, V.I. Norkin, and R.J.B. Wets. The minimization of discontinuous functions: mollifier subgradients. *SIAM Journal on Control and Optimization*, 33:149–167, 1995. 28

[41] S.M. Fallat and L. Hogben. The minimum rank of symmetric matrices described by a graph: a survey. *Linear Algebra and its Applications*, 426(2):558–582, 2007. 2

[42] J. Fan. Comments on "Wavelets in statistics: A review" by A. Antoniadis. *Statistical Methods and Applications*, 6(2):131–138, 1997. 10, 128, 133

[43] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001. 10, 128

[44] J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008. 10

[45] J. Fan and J. Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101, 2010. 10

[46] J. Fan, J. Lv, and L. Qi. Sparse high dimensional models in economics. *Annual Review of Economics*, 3:291, 2011. 10

[47] J. Fan and H. Peng. Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961, 2004. 10

[48] K. Fan. On a theorem of Weyl concerning eigenvalues of linear transformations I. *Proceedings of the National Academy of Sciences of the United States of America*, 35(11):652, 1949. 22

[49] M. Fazel. *Matrix rank minimization with applications*. PhD thesis, Stanford University, 2002. 2, 3, 7, 48, 102, 103, 104, 114, 134

[50] M. Fazel, H. Hindi, and S. Boyd. Rank minimization and applications in system theory. In *American Control Conference, 2004. Proceedings of the 2004*, volume 4, pages 3273–3278. IEEE, 2004. 2

[51] M. Fazel, H. Hindi, and S.P. Boyd. Log-det heuristic for matrix rank minimization with applications to hankel and euclidean distance matrices. In *American Control Conference, 2003. Proceedings of the 2003*, volume 3, pages 2156–2162. Ieee, 2003. 7, 13, 87, 114, 124, 134, 173

[52] M. Fazel, TK Pong, D. Sun, and P. Tseng. Hankel matrix rank minimization with applications in system identification and realization. *SIAM Journal on Matrix Analysis and Applications, to appear*. 2, 123, 144

[53] S.T. Flammia, D. Gross, Y.K. Liu, and J. Eisert. Quantum tomography via compressed sensing: error bounds, sample complexity, and efficient estimators. *Arxiv preprint* arXiv:1205.2300, 2012. 9, 46, 98

[54] M. Fornasier, H. Rauhut, and R. Ward. Low-rank matrix recovery via iteratively reweighted least squares minimization. *SIAM Journal on Optimization*, 21(4):1614–1640, 2011. 8, 129, 137

[55] S. Foucart and M.J. Lai. Sparsest solutions of underdetermined linear systems via l_q-minimization for $0 < q < 1$. *Applied and Computational Harmonic Analysis*, 26(3):395–407, 2009. 128

[56] R. Foygel, R. Salakhutdinov, O. Shamir, and N. Srebro. Learning with the weighted trace-norm under arbitrary sampling distributions. In *Advances in Neural Information Processing Systems (NIPS) 24*, volume 24, pages 2133–2141, 2011. 7

[57] R. Foygel and N. Srebro. Concentration-based guarantees for low-rank matrix reconstruction. In *24nd Annual Conference on Learning Theory (COLT)*, 20101. 7

[58] I.E. Frank and J.H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, pages 109–135, 1993. 128, 133

[59] M. Fukushima and H. Mine. A generalized proximal point algorithm for certain non-convex minimization problems. *International Journal of Systems Science*, 12(8):989–1000, 1981. 35, 36

[60] D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2(1):17–40, 1976. 144

[61] Y. Gao. *Structured low rank matrix optimization problems: A penalized approach*. PhD thesis, National University of Singapore, 2010. 11, 35, 36, 37

[62] Y. Gao and D.F. Sun. A majorized penalty approach for calibrating rank constrained correlation matrix problems. *Preprint available at* http://www.math.nus.edu.sg/~matsundf/MajorPen_May5.pdf, 2010. 11, 35, 36, 50, 84

[63] D. Geman and G. Reynolds. Constrained restoration and the recovery of discontinuities. *IEEE Transactions on pattern analysis and machine intelligence*, 14(3):367–383, 1992. 128

[64] C.J. Geyer. On the asymptotics of constrained M-estimation. *The Annals of Statistics*, pages 1993–2010, 1994. 32

[65] C.J. Geyer. On the asymptotics of convex stochastic optimization. *Unpublished manuscript*, 1996. 30

[66] H. Ghasemi, M. Malek-Mohammadi, M. Babaie-Zadeh, and C. Jutten. SRF: Matrix completion based on smoothed rank function. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 3672–3675. IEEE, 2011. 129

[67] R. Glowinski and A. Marrocco. *Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité, d'une classe de problèmes de Dirichlet non linéaires*, volume 140. Laboria, 1975. 144

[68] W. Glunt, T.L. Hayden, and R. Reams. The nearest doubly stochastic matrix to a real matrix with the same first moment. *Numerical Linear Algebra with Applications*, 5(6):475–482, 1998. 157

[69] R. Gribonval and M. Nielsen. Highly sparse representations from dictionaries are unique and independent of the sparseness measure. *Applied and Computational Harmonic Analysis*, 22(3):335–355, 2007. 110

[70] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *Information Theory, IEEE Transactions on*, 57(3):1548–1566, 2011. 6, 48, 60

[71] D. Gross, Y.K. Liu, S.T. Flammia, S. Becker, and J. Eisert. Quantum state tomography via compressed sensing. *Physical Review Letters*, 105(15):150401, 2010. 6, 9, 46, 98

[72] N. Halko, P.G. Martinsson, and J.A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011. 151

[73] C. Han and P.C.B. Phillips. GMM with many moment conditions. *Econometrica*, 74(1):147–192, 2006. 30

[74] G.H. Hardy, J.E. Littlewood, and G. Pólya. *Inequalities*. Univ. Press, Cambridge, 1934. 15, 16

[75] W.J. Heiser. Convergent computation by iterative majorization: Theory and applications in multidimensional data analysis. *Recent advances in descriptive multivariate analysis*, pages 157–189, 1995. 33

[76] A. Horn. Doubly stochastic matrices and the diagonal of a rotation matrix. *American Journal of Mathematics*, 76(3):620–630, 1954. 16

[77] J. Huang, S. Ma, and C.H. Zhang. Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, 18(4):1603, 2010. 10

[78] D.R. Hunter and K. Lange. A tutorial on MM algorithms. *The American Statistician*, 58(1):30–37, 2004. 33

[79] K. Jiang, D. Sun, and K.C. Toh. An inexact accelerated proximal gradient method for large scale linearly constrained convex SDP. *SIAM Journal on Optimization*, 22(3):1042–1064, 2012. 5, 9, 77, 88, 123

[80] K. Jiang, D.F. Sun, and K.C. Toh. Solving nuclear norm regularized and semidefinite matrix least squares problems with linear equality constraints. *Fields Institute Communications Series on Discrete Geometry and Optimization, K. Bezdek, Y. Ye, and A. Deza eds., to appear.* 5, 9, 123, 146

[81] K. Jiang, D.F. Sun, and K.C. Toh. A partial proximal point algorithm for nuclear norm regularized matrix least squares problems. *Preprint available at* http://www.math.nus.edu.sg/~matsundf/PPA_Smoothing-2.pdf, 2012. 5, 9, 123, 147, 157, 165, 166

[82] R.H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *Information Theory, IEEE Transactions on*, 56(6):2980–2998, 2010. 1, 6

[83] R.N. Khoury. Closest matrices in the space of generalized doubly stochastic matrices. *Journal of mathematical analysis and applications*, 222(2):562–568, 1998. 157

[84] A.J. King and R.J.B. Wets. Epi-consistency of convex stochastic programs. *Stochastics: An International Journal of Probability and Stochastic Processes*, 34(1-2):83–92, 1991. 32

[85] O. Klopp. Rank penalized estimators for high-dimensional matrices. *Electronic Journal of Statistics*, 5:1161–1183, 2011. 7

[86] O. Klopp. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli, to appear*, 2012. 6, 11, 48, 54, 55, 58, 61, 62

[87] K. Knight. Epi-convergence in distribution and stochastic equi-semicontinuity. *Unpublished manuscript*, 1999. 30, 32

[88] V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole Dâeté de Probabilités de Saint-FlourXXXVIII-2008*, volume 2033. Springer, 2011. 60

[89] V. Koltchinskii. Sharp oracle inequalities in low rank estimation. *Arxiv preprint* arXiv:1210.1144, 2012. 6

[90] V. Koltchinskii. Von Neumann entropy penalization and low-rank matrix estimation. *The Annals of Statistics*, 39(6):2936–2973, 2012. 7, 60

[91] V. Koltchinskii, K. Lounici, and A.B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011. 6, 48, 54, 60, 61

[92] L. Kong, L. Tunçel, and N. Xiu. S-goodness for low-rank matrix recovery, translated from sparse singal recovery. *Preprint available at* http://www.math.uwaterloo.ca/~ltuncel/publications/suff-LMR.pdf, 2011. 4

[93] L. Kong and N. Xiu. New bounds for restricted isometry constants in low-rank matrix recovery. *Preprint available at* http://www.optimization-online.org/DB_FILE/2011/01/2894.pdf, 2011. 5

[94] M.J. Lai, S. Li, L.Y. Liu, and H. Wang. Two results on the Schatten p-quasi-norm minimization for low rank matrix recovery. *Preprint available at*

http://www.math.uga.edu/~mjlai/papers/LaiLiLiuWang.pdf, 2012. 23, 128

[95] M.J. Lai and J. Wang. An unconstrained l_q minimization for sparse solution of underdetermined linear system. *SIAM Journal on Optimization*, 21:82–101, 2010. 128

[96] M.J. Lai, Y. Xu, and W. Yin. Improved iteratively reweighted least squares for unconstrained smoothed l_q minimization. *Preprint available at* http://www.caam.rice.edu/~wy1/paperfiles/Rice_CAAM_TR11-12_Mtx_Rcvry_ncvx_Lq.PDF, 2012. 137

[97] K. Lange. A gradient algorithm locally equivalent to the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 425–437, 1995. 35

[98] K. Lange. *Optimization*. Springer, New York, 2004. 35

[99] K. Lange, D.R. Hunter, and I. Yang. Optimization transfer using surrogate objective functions. *Journal of computational and graphical statistics*, 9(1):1–20, 2000. 33, 35

[100] R.M. Larsen. PROPACK-Software for large and sparse SVD calculations. *Available online.* http://soi.stanford.edu/~rmunk/PROPACK/, 2004. 153, 169

[101] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*, volume 23. Springer, 1991. 56

[102] K. Lee and Y. Bresler. Guaranteed minimum rank approximation from linear observations by nuclear norm minimization with an ellipsoidal constraint. *Arxiv preprint* arXiv:0903.4742, 2009. 5

[103] K. Lee and Y. Bresler. ADMIRA: Atomic decomposition for minimum rank approximation. *Information Theory, IEEE Transactions on*, 56(9):4402–4416, 2010. 1, 4

[104] C. Leng, Y. Lin, and G. Wahba. A note on the lasso and related procedures in model selection. *Statistica Sinica*, 16(4):1273, 2006. 10

[105] A.S. Lewis and H.S. Sendov. Nonsmooth analysis of singular values. Part II: Applications. *Set-Valued Analysis*, 13(3):243–264, 2005. 19, 68

[106] B. V. Lidskii. The proper values of the sum and the product of symmetric matrices (in Russian). *Dokl. Akad. Nauk SSSR*, 74:769—772, 1950. 24

[107] J. Lin. Reduced rank approximations of transition matrices. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, 2003. 165

[108] N. Linial, E. London, and Y. Rabinovich. The geometry of graphs and some of its algorithmic applications. *Combinatorica*, 15(2):215–245, 1995. 2

[109] Y.J. Liu, D. Sun, and K.C. Toh. An implementable proximal point algorithmic framework for nuclear norm minimization. *Mathematical Programming*, pages 1–38, 2009. 5, 123, 125, 153

[110] Y.K. Liu. Universal low-rank matrix recovery from Pauli measurements. *Arxiv preprint* arXiv:1103.2816, 2011. 54

[111] M.S. Lobo, M. Fazel, and S. Boyd. Portfolio optimization with linear and fixed transaction costs. *Annals of Operations Research*, 152(1):341–365, 2007. 133

[112] K. Löwner. Über monotone matrixfunktionen. *Mathematische Zeitschrift*, 38(1):177–216, 1934. 18

[113] Z. Lu and Y. Zhang. Penalty decomposition methods for rank minimization. *Arxiv preprint* arXiv:1008.5373, 2010. 8, 138

[114] S. Ma, D. Goldfarb, and L. Chen. Fixed point and Bregman iterative methods for matrix rank minimization. *Mathematical Programming*, pages 1–33, 2009. 5

[115] S. Ma, D. Goldfarb, and L. Chen. Fixed point and bregman iterative methods for matrix rank minimization. *Mathematical Programming*, 128(1):321–353, 2011. 19, 123

[116] A. Majumdar and R. Ward. Some empirical advances in matrix completion. *Signal Processing*, 91(5):1334–1338, 2011. 129

[117] A.S. Markus. The eigen- and singular values of the sum and product of linear operators. *Russian Mathematical Surveys*, 19:91–120, 1964. 16

[118] A.W. Marshall, I. Olkin, and B. Arnold. *Inequalities: Theory of Majorization and Its Applications*. Springer Verlag, 2010. 16, 24, 82

[119] B. Martinet. Breve communication. régularisation dinéquations variationelles par approximations successives. *Revue Française dInformatique et de Recherche Opérationelle*, 4:154–158, 1970. 42

[120] P. Massart. About the constants in Talagrand's concentration inequalities for empirical processes. *The Annals of Probability*, pages 863–884, 2000. 56

[121] N. Meinshausen. Relaxed lasso. *Computational Statistics & Data Analysis*, 52(1):374–393, 2007. 10

[122] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006. 10

[123] R. Meka, P. Jain, and I.S. Dhillon. Guaranteed rank minimization via singular value projection. In *Proceedings of the Neural Information Processing Systems Conference (NIPS)*, pages 937–945, 2010. 1, 4

[124] M. Mesbahi. On the rank minimization problem and its control applications. *Systems & Control Letters*, 33(1):31–36, 1998. 3

[125] M. Mesbahi and G.P. Papavassilopoulos. On the rank minimization problem over a positive semidefinite linear matrix inequality. *Automatic Control, IEEE Transactions on*, 42(2):239–243, 1997. 3

[126] R.R. Meyer. Sufficient conditions for the convergence of monotonic mathematical programming algorithms. *Journal of computer and system sciences*, 12(1):108–121, 1976. 34

[127] W. Miao, S. Pan, and D. Sun. A rank-corrected procedure for matrix completion with fixed basis coefficients. *Arxiv preprint* arXiv:1210.3709, 2012. 3

[128] W. Miao, S. Pan, and D. Sun. An adaptive semi-nuclear approach for rank optimization problems with hard constraints. *In preparation*, 2013. 3

[129] H. Mine and M. Fukushima. A minimization method for the sum of a convex function and a continuously differentiable function. *Journal of Optimization Theory and Applications*, 33(1):9–23, 1981. 35, 36

[130] K. Mohan and M. Fazel. Iterative reweighted algorithms for matrix rank minimization. *Journal of Machine Learning Research, to appear.* 8, 129, 136, 137, 151, 152

[131] K. Mohan and M. Fazel. New restricted isometry results for noisy low-rank recovery. In *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*, pages 1573–1577. IEEE, 2010. 5

[132] K. Mohan and M. Fazel. Reweighted nuclear norm minimization with application to system identification. In *American Control Conference (ACC), 2010*, pages 2953–2959. IEEE, 2010. 8, 10, 87, 135, 136

[133] S. Negahban, P. Ravikumar, M.J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012. 5

[134] S. Negahban and M.J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069–1097, 2011. 5

[135] S. Negahban and M.J. Wainwright. Restricted strong convexity and weighted matrix completion: optimal bounds with noise. *Journal of Machine Learning Research*, 13:1665–1697, 2012. 6, 7, 48, 54, 61

[136] Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. In *Soviet Mathematics Doklady*, volume 27, pages 372–376, 1983. 88

[137] D. Nettleton. Convergence properties of the EM algorithm in constrained parameter spaces. *Canadian Journal of Statistics*, 27(3):639–648, 1999. 35

[138] J.M. Ortega and W.C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press (New York), 1970. 33

[139] S. Oymak and B. Hassibi. New null space results and recovery thresholds for matrix rank minimization. *Arxiv preprint* arXiv:1011.6326, 2010. 4, 108

[140] S. Oymak, K. Mohan, M. Fazel, and B. Hassibi. A simplified approach to recovery conditions for low rank matrices. In *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*, pages 2318–2322. IEEE, 2011. 5, 23, 108, 128

[141] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. 1999. 165

[142] G.C. Pflug. Asymptotic dominance for solutions of stochastic programs. *Czechoslovak Journal for Operations Research*, 1(1):21–30, 1992. 32

[143] G.C. Pflug. Asymptotic stochastic programs. *Mathematics of Operations Research*, pages 769–789, 1995. 31, 32

[144] H. Qi and D.F. Sun. A quadratically convergent newton method for computing the nearest correlation matrix. *SIAM Journal on Matrix Analysis and Applications*, 28(2):360, 2006. 78

[145] H. Qi and D.F. Sun. An augmented Lagrangian dual approach for the *H*-weighted nearest correlation matrix problem. *IMA Journal of Numerical Analysis*, 31(2):491–511, 2011. 78

[146] Audenaert K. M. R. and F. Kittaneh. Problems and conjectures in matrix and operator inequalities. *Arxiv preprint* arXiv:1201.5232, 2012. 22

[147] R. Rado. An inequality. *J. London Math. Soc.*, 27:1–6, 1952. 16

[148] B.D. Rao, K. Engan, S.F. Cotter, J. Palmer, and K. Kreutz-Delgado. Subset selection in noise based on diversity measure minimization. *IEEE Transactions on Signal Processing*, 51(3):760–770, 2003. 133

[149] B. Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12:3413–3430, 2011. 6, 48, 60

[150] B. Recht, M. Fazel, and P.A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010. 3, 48

[151] B. Recht, W. Xu, and B. Hassibi. Null space conditions and thresholds for rank minimization. *Mathematical programming*, 127(1):175–202, 2011. 4

[152] S.M. Robinson. Local structure of feasible sets in nonlinear programming, Part II: Nondegeneracy. *Mathematical Programming at Oberwolfach II*, pages 217–230, 1984. 76

[153] R.T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970. 31, 41, 70

[154] R.T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976. 42

[155] R.T. Rockafellar and R.J.B. Wets. *Variational Analysis*, volume 317. Springer Verlag, 1998. 23, 28, 29, 73, 77

[156] A. Rohde and A.B. Tsybakov. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930, 2011. 7

[157] S. Ju. Rotfel'd. Remarks on the singular values of a sum of completely continuous operators. *Funkcional. Anal. i Priložen*, 1(3):95–96, 1967. 22

[158] R. Salakhutdinov and N. Srebro. Collaborative filtering in a non-uniform world: Learning with the weighted trace norm. In *Advances in Neural Information Processing Systems (NIPS)*, volume 23, pages 2056–2064, 2010. 7, 48

[159] E.D. Schifano, R.L. Strawderman, and M.T. Wells. Majorization-minimization algorithms for nonsmoothly penalized objective functions. *Electronic Journal of Statistics*, 4:1258–1299, 2010. 35

[160] R. Sinkhorn and P. Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific J. Math*, 21(2):343–348, 1967. 157

[161] N. Srebro. *Learning with matrix factorizations*. PhD thesis, Massachusetts Institute of Technology, 2004. 2

[162] N. Srebro, J.D.M. Rennie, and T. Jaakkola. Maximum-margin matrix factorization. *Advances in Neural Information Processing Systems (NIPS)*, 17(5):1329–1336, 2005. 2, 7

[163] N. Srebro and A. Shraibman. Rank, trace-norm and max-norm. *Learning Theory*, pages 599–764, 2005. 7

[164] T. Sun and C.H. Zhang. Calibrated elastic regularization in matrix completion. *Arxiv preprint* arXiv:1211.2264, 2012. 7

[165] R.C. Thompson. Convex and concave functions of singular values of matrix sums. *Pacific Journal of Mathematics*, 66(1):285–290, 1976. 22

[166] R.C. Thompson and L.J. Freede. On the eigenvalues of sums of Hermitian matrices. *Linear Algebra and Its Applications*, 4(4):369–376, 1971. 24

[167] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. 9

[168] K.C. Toh and S. Yun. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific J. Optim*, 6:615–640, 2010. 5, 153, 170

[169] J.A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, pages 1–46, 2011. 60

[170] P. Tseng and S. Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117(1):387–423, 2009. 35, 36

[171] M. Uchiyama. Subadditivity of eigenvalue sums. *Proceedings of the American Mathematical Society*, 134(5):1405–1412, 2006. 22

[172] F. Vaida. Parameter convergence for EM and MM algorithms. *Statistica Sinica*, 15(3):831, 2005. 35

[173] A.W. Van Der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes*. Springer Verlag, 1996. 56, 60

[174] J. von Neumann. Some matrix inequalities and metrization of matric space. *Mitt. Forsch.-Inst. Math. Mech. Univ. Tomsk*, 1:286–299, 1937. 116

[175] Y. Wang. Asymptotic equivalence of quantum state tomography and trace regression. *Preprint available at* `http://pages.stat.wisc.edu/~yzwang/paper/QuantumTomographyTrace.pdf`, 2012. 9

[176] G.A. Watson. Characterization of the subdifferential of some matrix norms. *Linear Algebra and its Applications*, 170:33–45, 1992. 21, 53, 69, 71, 121

[177] G.A. Watson. On matrix approximation problems with Ky Fan k norms. *Numerical Algorithms*, 5(5):263–272, 1993. 21, 71, 121

[178] Z. Wen, W. Yin, and Y. Zhang. Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm. *Mathematical Programming Computation*, pages 1–29, 2010. 1

[179] H. Weyl. Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung). *Mathematische Annalen*, 71(4):441–479, 1912. 24

[180] R. Wijsman. Convergence of sequences of convex sets, cones and functions. *Bulletin (New Series) of the American Mathematical Society*, 70(1):186–188, 1964. 27

[181] D. Wipf and S. Nagarajan. Iterative reweighted l_1 and l_2 methods for finding sparse solutions. *IEEE Journal of Selected Topics in Signal Processing*, 4(2):317–329, 2010. 133

[182] F. Woolfe, E. Liberty, V. Rokhlin, and M. Tygert. A fast randomized algorithm for the approximation of matrices. *Applied and Computational Harmonic Analysis*, 25(3):335–366, 2008. 151

[183] C.F.J. Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, pages 95–103, 1983. 34

[184] T.T. Wu and K. Lange. The MM alternative to EM. *Statistical Science*, 25(4):492–505, 2010. 33

[185] J. Yang and X. Yuan. Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization. *Mathematics of Computation, to appear*. 5

[186] Z. Yang. *A study on nonsymmetric matrix-valued functions*. Master's thesis, National University of Singapore, 2009. 18

[187] M. Yuan, A. Ekici, Z. Lu, and R. Monteiro. Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal*

Statistical Society: Series B (Statistical Methodology), 69(3):329–346, 2007. 2

[188] M.C. Yue and A.M.C. So. A perturbation inequality for the schatten-p quasi-norm and its applications to low-rank matrix recovery. *Arxiv preprint* arXiv:1209.0377, 2012. 23, 26, 27, 108, 128

[189] W.I. Zangwill. *Nonlinear Programming: A Unified Approach*. Prentice-Hall NJ:, 1969. 34

[190] C.H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010. 10, 128, 133

[191] P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7(2):2541, 2007. 10

[192] Y.B. Zhao. An approximation theory of matrix rank minimization and its application to quadratic equations. *Linear Algebra and its Applications*, 2012. 129

[193] S. Zhou, S. Van De Geer, and P. Bühlmann. Adaptive Lasso for high dimensional regression and Gaussian graphical modeling. *Arxiv preprint* arXiv:0903.2515, 2009. 10

[194] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006. 10, 133

[195] H. Zou and R. Li. One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, 36(4):1509, 2008. 10, 133

Name: Miao Weimin

Degree: Doctor of Philosophy

Department: Mathematics

Thesis Title: Matrix Completion Models with Fixed Basis Coefficients and Rank Regularized Problems with Hard Constraints

Abstract

The problems with embedded low-rank structures arise in diverse areas such as engineering, statistics, quantum information, finance and graph theory. This thesis is devoted to dealing with the low-rank structure via techniques beyond the widely-used nuclear norm for achieving better performance. In the first part, we propose a rank-corrected procedure for low-rank matrix completion problems with fixed basis coefficients. We establish non-asymptotic recovery error bounds and provide necessary and sufficient conditions for rank consistency. The obtained results, together with numerical experiments, indicate the superiority of our proposed rank-correction step over the nuclear norm penalization. In the second part, we propose an adaptive semi-nuclear norm regularization approach to address rank regularized problems with hard constraints via solving their nonconvex but continuous approximation problems instead. This approach overcomes the difficulty of extending the iterative reweighted l_1 minimization from the vector case to the matrix case. Numerical experiments show that the iterative scheme of our propose approach has advantages of achieving both the low-rank-structure-preserving ability and the computational efficiency.

Keywords: matrix completion, rank minimization, matrix recovery, low rank, error bound, rank consistency, semi-nuclear norm.

MATRIX COMPLETION MODELS WITH FIXED BASIS COEFFICIENTS AND RANK REGULARIZED PROBLEMS WITH HARD CONSTRAINTS

MIAO WEIMIN

NATIONAL UNIVERSITY OF SINGAPORE

2013

MATRIX COMPLETION MODELS WITH FIXED BASIS COEFFICIENTS AND RANK REGULARIZED PROBLEMS WITH HARD CONSTRAINTS

MIAO WEIMIN

2013