

Copyright Undertaking

This thesis is protected by copyright, with all rights reserved.

By reading and using the thesis, the reader understands and agrees to the following terms:

1. The reader will abide by the rules and legal ordinances governing copyright regarding the use of the thesis.
2. The reader will use the thesis for the purpose of research or private study only and not for distribution or further reproduction or any other purpose.
3. The reader agrees to indemnify and hold the University harmless from and against any loss, damage, cost, liability or expenses arising from copyright infringement or unauthorized usage.

IMPORTANT

If you have reasons to believe that any materials in this thesis are deemed not suitable to be distributed in this form, or a copyright owner having difficulty with the material being included in our database, please contact lbsys@polyu.edu.hk providing details. The Library will look into your claim and consider taking remedial action upon receipt of the written requests.

A PROXIMAL GENERATION BASED LEVEL
SET METHOD
WITH SECANT ITERATIONS FOR THE
LEAST-SQUARES
CONSTRAINED NUCLEAR NORM
MINIMIZATION

CHIYU MA

PhD

The Hong Kong Polytechnic University

2026

The Hong Kong Polytechnic University

Department of Applied Mathematics

A Proximal Generation based Level Set Method
with Secant Iterations for the Least-Squares
Constrained Nuclear Norm Minimization

Chiyu Ma

A thesis submitted in partial fulfilment of the requirements

for the degree of Doctor of Philosophy

August 2025

Certificate of Originality

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

_____(Signed)

MA Chiyu _____(Name of student)

Dedicate to my family.

Abstract

The least squares constrained nuclear norm minimization is a problem of practical significance but poses computational challenges. Its relevance spans numerous fields, including image processing, recommendation systems, system identification, classification, and clustering, among others. In this thesis, we propose a proximal generation based level set method with secant iterations to address this problem efficiently.

Our approach combines a level set method with secant iterations and proximal generation techniques. The level set method reformulates the problem as a univariate nonlinear equation $\varphi(\lambda) = \varrho$ derived from the regularized least squares problems. We employ the secant method for this equation and establish its fast convergence rates, which depend critically on two properties: the nonsingularity of $\partial\varphi(\lambda)$ and (strong) semismoothness of $\varphi(\lambda)$. We prove the nonsingularity of the Clarke generalized Jacobian for a broader class of univariate functions. For the semismooth analysis, we analyze the γ -order semismoothness through the semialgebraic property and derive a sufficient condition of the strong semismoothness through the implicit function theorem and the Lipschitz homeomorphism of the solution mapping.

For the nuclear norm regularized subproblems within the level set method, we develop a proximal generation method that effectively exploits low-rank structures without compromising convergence. The method employs low-rank decomposition to reduce subproblems' computation. Proximal gradient steps are utilized to gen-

erate high-quality initial points for nonconvex problems and guarantee the global convergence simultaneously.

Numerical experiments demonstrate the superior performance of our approach compared to existing methods. These experiments highlight our method’s robustness, efficiency, and ability to handle large-scale problems, making it a valuable tool for applications in machine learning, data science, and signal processing where low-rank matrix optimization is prevalent.

Acknowledgements

I wish to express my deepest gratitude to my advisor, Professor Defeng Sun, for his unwavering guidance and support throughout this research. His profound insight and patience were instrumental in shaping this work. He meticulously corrected many errors I had made, and his rigorous standards have profoundly shaped my analytical thinking. With his mentorship, I believe I am firmly on the road to becoming a capable applied mathematician.

My sincere thanks also go to the other faculty of our research group, for their invaluable feedback and for creating such a stimulating academic environment.

I extend my heartfelt thanks to my colleagues at Huawei. Collaborating with them provided me with the invaluable opportunity to apply mathematical optimization to real-world industrial challenges, profoundly enriching my perspective on research.

I am deeply thankful to my colleagues within our research group for their collaboration and camaraderie. I am also grateful to my classmates and friends in the wider program for their support and for many enlightening discussions. A special mention must be made of Mr. Jiaming Ma, whose expertise was invaluable and who provided indispensable help with the proofs in Section 3.2.

To my friends beyond these walls, thank you for your constant encouragement and for providing a joyful balance to my academic life.

I reserve my most profound appreciation for my family. Their unconditional love

and belief in me have been my foundation. I must pay a special tribute to my beloved grandfather, who passed away suddenly due to COVID-19 at the beginning of 2023. I am forever grateful for his care and support throughout my life, and for instilling in me a love of learning through his own dedicated reading habits; his influence remains with me always.

Finally, I would like to conclude with some lines from *The Wild Swans at Coole* by William Butler Yeats:

But now they drift on the still water,
Mysterious, beautiful;
Among what rushes will they build,
By what lake's edge or pool
Delight men's eyes when I awake some day
To find they have flown away?

Contents

Certificate of Originality	iii
Abstract	vii
Acknowledgements	ix
List of Figures	xiii
List of Tables	xv
List of Notations	xvii
1 Introduction	1
1.1 Motivating examples	3
1.1.1 Matrix completion	4
1.1.2 Low rank matrix regression	5
1.1.3 Low rank representation based subspace clustering	6
1.1.4 Applications with similar structure	7
1.2 Related approaches	10
1.2.1 Algorithms for the regularized problem $(\text{LS}(\lambda)\text{-P})$	10
1.2.2 Level set methods	13
1.2.3 Adaptive sieving strategies	14
1.3 Summary of contributions of the thesis	17
1.4 Organization of the thesis	19

2	Preliminaries	21
2.1	Semismooth analysis	21
2.2	Matrix analysis	27
2.3	Sensitivity analysis	32
3	A level set method with secant iterations for the least-squares constrained nuclear norm minimization	41
3.1	A level set method with the value function $\varphi(\lambda)$	42
3.2	Nonsingularity of the Clarke generalized Jacobian $\partial\varphi(\lambda)$	46
3.3	A secant method for the univariate equation (Uni.Eq.)	53
4	A proximal generation method for the nuclear norm regularized least-squares problem (LS(λ)-P)	61
4.1	A proximal generation method with low rank decomposition for the nuclear norm regularized least-squares problem (LS(λ)-P)	61
4.2	Convergence of the proximal generation method (Algorithm 4)	66
5	Numerical Experiment	71
5.1	Numerical issues in the proximal generation based level set method with secant iterations (Algorithm 5)	72
5.1.1	Numerical issues in the secant method (Algorithm 3)	72
5.1.2	Numerical issues in the proximal generation strategy (Algorithm 4)	73
5.2	ADMM algorithms for least-squares constrained nuclear norm minimization (CP(ρ)) and nuclear norm regularized least square problem (LS(λ)-D)	75
5.3	Matrix regression problems	78
5.4	Matrix completion problems	84
5.5	Numerical performance of secant iterations	91
6	Conclusions and future work	93
	Bibliography	95

List of Figures

3.1	A diagram for the proof of Theorem 3.1.	52
-----	---	----

List of Tables

5.1	Statistics of test data in low rank regressions	79
5.2	The performance of proximal generation based level set method with secant iterations (GLS_{sec}), SPGL1(SPG), and ADMM in solving the least-squares constrained nuclear norm minimization ($\text{CP}(\varrho)$) with the $\rho = c\ b\ $ in Table 5.1	80
5.3	The performance of proximal generation based level set method with secant iterations (GLS_{sec}) and with bisection iterations (GLS_{bis}) . . .	82
5.4	The performance of proximal generation based level set method with secant iterations (GLS_s) and bisection iterations (GLS_b), SPGL1(SPG), and ADMM(ADM) in solving the least square constrained nuclear norm minimization ($\text{CP}(\varrho)$) with the $\rho = c\ b\ $ in Table 5.1; and the performance of ADMM (ADM_{λ^*}) and APG (APG_{λ^*}) in solving the nuclear norm regularized least square problem ($\text{LS}(\lambda)$ -P) with the λ returned by GLS_{sec}	83
5.5	Statistics of test data in low rank matrix completion	85
5.6	The performance of proximal generation based level set method with secant iterations (GLS_{sec}), SPGL1(SPG), and ADMM in solving the least-squares constrained nuclear norm minimization ($\text{CP}(\varrho)$) with the $\rho = c\ b\ $ in Table 5.5.	87
5.7	The performance of proximal generation based level set method with secant iterations (GLS_{sec}) and with bisection iterations (GLS_{bis}) . . .	88
5.8	The performance of proximal generation based level set method with secant iterations (GLS_s) and bisection iterations (GLS_b), SPGL1(SPG), and ADMM(ADM) in solving the least square constrained nuclear norm minimization ($\text{CP}(\varrho)$) with the $\rho = c\ b\ $ in Table 5.5; and the performance of ADMM (ADM_{λ^*}) and APG (APG_{λ^*}) in solving the nuclear norm regularized least square problem ($\text{LS}(\lambda)$ -P) with the λ returned by GLS_{sec}	89

5.9	The performance of proximal generation based level set method with secant iterations (GLS_{sec}) and LoRADS	91
5.10	Numerical performance of secant iterations	92

List of Notations

- Let $\mathbb{R}$ be the set of real numbers, $\mathbb{R}^n$ be the real linear space of all n –dimensional vectors, and $\mathbb{R}^{m \times n}$ be the real linear space of all $m \times n$ matrices. Let $\mathbb{R}_+$ and $\mathbb{R}_{++}$ be the nonnegative real numbers and positive real numbers respectively.
- Vectors are denoted by lowercase letters (e.g., $x \in \mathbb{R}^n$), matrices by uppercase letters (e.g., $X \in \mathbb{R}^{m \times n}$), and linear operators/mappings by calligraphic letters (e.g., $\mathcal{A} : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^p$).
- For $x \in \mathbb{R}^n$, let $\|x\|_1$, $\|x\|$, and $\|x\|_\infty$ be the ℓ_1 norm, ℓ_2 norm, and ℓ_∞ norm respectively. For matrix $X \in \mathbb{R}^{m \times n}$, let $\|X\|_*$, $\|X\|$, and $\|X\|_2$ be the nuclear norm, Frobenius norm, and spectral norm respectively. The ℓ_1 norm of a matrix is defined as $\|X\|_1 = \sum_{1 \leq i \leq m, 1 \leq j \leq n} |X_{ij}|$, where X_{ij} is the (i, j) –th element in matrix X .
- The epigraph of spectral norm, a matrix cone, is denoted by $\text{epi } \|\cdot\|_2^{m \times n} := \{(t, X) \in \mathbb{R}_+ \times \mathbb{R}^{m \times n} \mid t \geq \|X\|_2\}$. The dual cone of epigraph of spectral norm is the epigraph of nuclear norm $\text{epi } \|\cdot\|_*^{m \times n} := \{(t, X) \in \mathbb{R}_+ \times \mathbb{R}^{m \times n} \mid t \geq \|X\|_*\}$. The cone of $n \times n$ positive semidefinite matrices is denoted as $\mathbb{S}_+^n$. All superscript dimensions may be omitted when clear from context.
- For a proper closed convex function $f : \mathbb{R}^n \rightarrow (-\infty, \infty]$, its proximal mapping is defined as $\text{prox}_{f(\cdot)}(z) = \text{argmin}_x f(x) + \frac{1}{2}\|x - z\|^2$ where $(\cdot)$ in the subscript could be omitted if there is no ambiguity.

- The indicator function of a nonempty closed convex set C is denoted as

$$\delta_C(x) := \begin{cases} 0 & \text{if } x \in C; \\ +\infty & \text{else.} \end{cases}$$

The proximal mapping of an indicator function $\delta_C(\cdot)$ is the projection onto C , written as $\text{prox}_{\delta_C}(z) = \Pi_C(z)$. The gauge function of a convex set C is defined as $\Upsilon(x \mid C) = \inf \{\lambda \geq 0 \mid x \in \lambda C\}$.

Chapter 1

Introduction

The nuclear norm has emerged as a fundamental tool in modern optimization, playing a pivotal role across diverse scientific and engineering disciplines. As the natural extension of the ℓ_1 norm from vector spaces to matrix spaces, the nuclear norm serves as a convex relaxation of the rank function as pointed in Recht et al. (2010). This convex surrogate property makes it particularly valuable in matrix optimization problems where low-rank solutions are desired, as it allows tractable computation while effectively promoting the underlying low-rank structure that appears in numerous applications.

The ubiquity of nuclear norm minimization is evident in its wide-ranging applications throughout contemporary research. In image processing, Cai et al. (2010) demonstrated its effectiveness for matrix completion tasks in visual data recovery. Recommendation systems leverage its properties for collaborative filtering, as summarized in Chen and Wang (2022), while in system identification, Fazel et al. (2013) established its utility with different structured matrices. Further applications span classification problems in Luo et al. (2015) and subspace clustering techniques in Liu et al. (2014), among many others. This remarkable versatility stems from the nuclear norm’s ability to simultaneously handle dimensionality reduction while maintaining problem convexity—a combination that is both theoretically sound and computa-

tionally practical.

In this thesis, we focus on a specific but important variant of nuclear norm optimization, the least-squares constrained nuclear norm minimization problem:

$$\begin{aligned} \min_{X \in \mathbb{R}^{m \times n}} \quad & \|X\|_* \\ \text{s.t.} \quad & \|\mathcal{A}X - b\| \leq \varrho, \end{aligned} \tag{CP}(\varrho)$$

where $\|\cdot\|_*$ denotes the nuclear norm (sum of singular values), $\|\cdot\|$ represents either the Euclidean norm for vectors or the Frobenius norm for matrices, $b \in \mathbb{R}^p$ is the observed data, $\mathcal{A}$ is a linear mapping from $\mathbb{R}^{m \times n}$ to $\mathbb{R}^p$, and $0 < \varrho < \|b\|$. We assume throughout that the problem is feasible and that $\|\mathcal{A}^*b\|_2 > 0$, conditions which ensure both mathematical well-posedness and practical relevance. An important observation is that the least-squares constraint is always active at optimal solutions—a property that will prove crucial in our subsequent analysis.

The formulation (CP)(ϱ) presents an elegant trade-off between two fundamental objectives: the least-squares constraint quantifies the model’s fidelity to observed data, while the nuclear norm in the objective function promotes the low-rank solutions that are often physically meaningful in applications. This differs from, yet complements, the more commonly studied regularized version:

$$\min_{X \in \mathbb{R}^{m \times n}} f_\lambda(X) := \lambda \|X\|_* + \frac{1}{2} \|\mathcal{A}X - b\|^2. \tag{LS}(\lambda)\text{-P}$$

While the regularized formulation (LS)(λ)-P has received more attention in the literature with solution methods including the Accelerated Proximal Gradient (APG) method (Toh and Yun, 2010), Alternating Direction Method of Multipliers (ADMM) (Fazel et al., 2013), and various Proximal Point Algorithms (PPA) (Liu et al., 2012; Jiang et al., 2013, 2014), the constrained version (CP)(ϱ) offers distinct practical advantages. The explicit least-squares constraint provides practitioners with direct

control over the noise level through the parameter ϱ , which can be more intuitive to calibrate than the regularization parameter λ in many applications.

The additional computational challenges posed by the constraint in $(\text{CP}(\varrho))$, particularly in maintaining feasibility while optimizing the nuclear norm, have limited its widespread adoption compared to the regularized formulation. However, developing efficient algorithms for this constrained problem holds significant promise: it would enable more precise control over solution quality in applications ranging from recommender systems to statistical modeling. Furthermore, a deeper understanding of this formulation may yield insights that benefit both theoretical analysis and practical implementation across the spectrum of nuclear norm optimization problems.

By focusing on this less-studied but practically important formulation, this thesis aims to bridge the gap between computational tractability and practical utility in constrained nuclear norm minimization—an advancement that could enhance solution accuracy and robustness across numerous application domains where these optimization techniques are employed.

1.1 Motivating examples

Nuclear norm minimization has been a widely studied topic in optimization and machine learning. It was established in Fazel et al. (2001) that the nuclear norm serves as the convex envelope of the rank function within the unit spectral norm ball. Initially applied to minimal-order system approximation in control theory, this convex relaxation gained broader recognition after its popularization by Recht et al. (2010). Since then, it has found extensive applications in regression analysis, recommendation systems, image reconstruction, and other domains. In this section, we review key applications of the formulation in $(\text{CP}(\varrho))$ and identify structurally similar problems that may benefit from our approach.

1.1.1 Matrix completion

Matrix completion is a fundamental problem with wide applications in recommendation systems, image analysis, and factor analysis. Its practical significance was highlighted by the Netflix Prize competition, while subsequent theoretical advances such as the singular value thresholding algorithm (Cai et al., 2010) provided key computational tools.

For a partially observed matrix $M \in \mathbb{R}^{m \times n}$ with corrupted entries indexed by $\Omega \in 1, 2, \dots, m \times 1, 2, \dots, n$, the matrix completion is to recover a low rank approximation X of M . By using the nuclear norm $\|X\|_*$ as a convex surrogate for rank minimization, the problem can be formulated as:

$$\begin{aligned} \min_{X \in \mathbb{R}^{m \times n}} \quad & \|X\|_* \\ \text{s.t.} \quad & \sum_{(i,j) \in \Omega} (X_{ij} - M_{ij})^2 \leq \varrho^2, \end{aligned} \tag{1.1}$$

where ϱ^2 bounds the permissible fitting error. This formulation, as shown in (1.1), fits the general framework of $(\text{CP}(\varrho))$ and illustrates the application of nuclear norm minimization to matrix completion. As discussed earlier, the least-squares constraint in (1.1) introduces non-smoothness, making the problem computationally challenging. Traditional approaches like proximal gradient methods or alternating direction methods (ADMM) are often employed, but they may struggle with scalability or tight error tolerances. Beyond standard matrix completion, this framework extends to related problems such as robust PCA, which decomposes a matrix into low-rank and sparse components, and tensor completion, which generalizes the approach to higher-dimensional data. These variants retain the core challenge of nuclear norm minimization while introducing additional structural constraints, further motivating the need for scalable and robust optimization techniques.

1.1.2 Low rank matrix regression

Low-rank regression addresses high-dimensional problems by imposing a low-rank structure on the coefficient matrix, simultaneously reducing model complexity and enhancing interpretability. This approach employs nuclear norm regularization as a convex surrogate for rank constraints, enabling tractable optimization while preserving the benefits of low-rank representations. Its effectiveness has been demonstrated across diverse domains, including dynamic stock return prediction (Fan et al., 2019), high-dimensional genetic analysis (Kong et al., 2020), and multi-task learning (Ji and Ye, 2009). Mathematically, this framework extends traditional least squares regression by incorporating a nuclear norm penalty—analogous to how Lasso promotes sparsity—yielding the regularized formulation in (LS(λ)-P). Notably, this regularized problem shares an intrinsic connection with the constrained formulation (CP(ρ)), as both balance low-rank structure with data fidelity through different but related mechanisms.

A key advantage of the least-squares constrained form (CP(ρ)), as discussed earlier, lies in its direct control over noise tolerance through the parameter ρ . Unlike the regularization parameter λ in penalized methods—whose effect depends on unknown problem scaling— ρ corresponds explicitly to the permissible noise level in observations. This makes the constrained formulation particularly appealing for applications where measurement error characteristics are known or can be reliably estimated, such as sensor calibration or physical system modeling. Both formulations nevertheless serve the same fundamental purpose: to recover structured, low-rank solutions while maintaining fidelity to observed data.

1.1.3 Low rank representation based subspace clustering

Subspace clustering addresses the challenge of grouping high-dimensional data distributed across multiple low-dimensional subspaces, which is a common scenario in applications ranging from motion segmentation to image classification (see Qu et al. (2023) for comprehensive surveys). The core optimization problem for clean data takes the form:

$$\begin{aligned} \min_C \quad & F(C) \\ \text{s.t.} \quad & X = YC, \end{aligned} \tag{1.2}$$

where X is the observed data, Y is the dictionary which could be generated from original data X , variable C is the coefficient matrix and the objective function F is the regularized term to induce structural properties. The choice of regularizer F is crucial as it determines how the subspace relationships are captured in the coefficient matrix. When $F(X) := \|X\|_*$, this yields low-rank representation (LRR) (Liu et al., 2010, 2014; Chen and Yang, 2013; Abdolali and Gillis, 2021), leveraging the nuclear norm to promote globally low-rank coefficients that reveal subspace affiliations.

In practical scenarios, the equality constraint $X = YC$ is often too restrictive due to data corruption. For cases with sparse outliers (e.g., occlusions in images or corrupted sensor readings), the problem can be reformulated by introducing an explicit outlier term E with ℓ_1 regularization :

$$\begin{aligned} \min_{C,E} \quad & \|C\|_* + \lambda \|E\|_1 \\ \text{s.t.} \quad & X = YC + E, \end{aligned}$$

where λ balances between low-rank structure and outlier sparsity. However, for general noise or dense corruption, a least-squares formulation becomes more appropriate:

$$\begin{aligned} \min_C \quad & \|C\|_* \\ \text{s.t.} \quad & \|X - YC\| \leq \rho, \end{aligned} \tag{1.3}$$

This formulation directly aligns with $(\text{CP}(\varrho))$, demonstrating the versatility of nuclear norm minimization under least-squares constraints for robust subspace recovery.

This method handles noisy real-world data while preserving interpretability. The parameter ρ directly balances accuracy and model complexity, proving valuable in medical imaging and sensor networks. It offers both theoretical guarantees and efficient computational solutions.

1.1.4 Applications with similar structure

This subsection presents applications that share structural similarities with problem $(\text{CP}(\varrho))$, identifying potential directions for future research.

Least squares constrained optimization for low rankness and sparsity

Many real-world applications, including hyperspectral image processing, traffic anomaly detection, matrix completion, and subspace clustering, require the simultaneous recovery of both low-rank and sparse components. To address this need, we consider an extension of least squares constrained nuclear norm minimization that incorporates explicit sparsity:

$$\begin{aligned} \min_{X, E \in \mathbb{R}^{m \times n}} \quad & \|X\|_* + g(E) \\ \text{s.t.} \quad & \|\mathcal{A}X + \mathcal{B}E - b\| \leq \varrho, \end{aligned} \tag{1.4}$$

where $g(E)$ denotes the sparsity-inducing regularizer (typically the ℓ_1 norm for element-wise sparsity or total variation norm for structured sparsity). This unified framework enables flexible modeling of various real-world scenarios where data consists of simultaneous low-rank and sparse components. The choice of $g(E)$ can be adapted to different sparsity patterns in specific applications.

Covariance estimation from quadratic sampling

Covariance estimation plays a fundamental role in statistical analysis and signal processing. The problem involves recovering a covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$ from quadratic measurements of the form:

$$y_i = a_i^\top \Sigma a_i + \eta_i, \quad i = 1, \dots, m \quad (1.5)$$

where η_i represent measurement noise. The associated linear operator $\mathcal{A}(M) : \mathbb{R}^{n \times n} \mapsto \mathbb{R}^m$ is defined by: $\mathcal{A}(M) = \{\langle M, a_i a_i^\top \rangle\}_{i=1}^m$. Following Chen et al. (2015), we examine two important formulations: the general low-rank covariance estimation problem

$$\begin{aligned} \min_M \quad & \text{tr}(M) \\ \text{s.t.} \quad & \|\mathcal{A}M - y\|_1 \leq \rho, \quad M \in \mathbb{S}_+, \end{aligned} \quad (1.6)$$

and its variant for wide-sense stationary processes that incorporates a Toeplitz constraint

$$\begin{aligned} \min_M \quad & \text{tr}(M) \\ \text{s.t.} \quad & \|\mathcal{A}M - y\|_2 \leq \rho, \\ & M \in \mathbb{S}_+, M \text{ is Toeplitz.} \end{aligned} \quad (1.7)$$

These formulations maintain convexity while addressing different structural assumptions about the underlying covariance matrix.

Mixed regression with two components

Mixed linear regression represents a fundamental problem in statistical learning that simultaneously addresses classification and regression tasks. The model considers observed data pairs $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$ generated from the mixture process:

$$y_i = z_i \cdot \langle x_i, \beta_1^* \rangle + (1 - z_i) \cdot \langle x_i, \beta_2^* \rangle + e_i, \quad i = 1, \dots, n,$$

where $z_i \in \{0, 1\}$ are latent class labels and e_i is the noise. This formulation captures scenarios where data originates from two distinct linear regimes, making it particularly relevant for applications such as behavioral analysis, sensor calibration, and econometric modeling where population heterogeneity exists. The joint estimation of both the regression coefficients β_1, β_2 and latent labels z_i presents significant computational and statistical challenges. Building on Chen et al. (2014) and Chen et al. (2017), a key theoretical advancement involves the convex relaxation:

$$\begin{aligned} \min_{K, g} \quad & \|K\|_* \\ \text{s.t.} \quad & \sum_{i=1}^n \left| -\langle x_i x_i^\top, K \rangle + 2y_i \langle x_i, g \rangle - y_i^2 \right| \leq \eta. \end{aligned} \tag{1.8}$$

This convex optimization approach provides robust estimates for mixed linear regression coefficients, with theoretical guarantees on recovery accuracy. By combining nuclear norm regularization with ℓ_1 constraints, it effectively handles noise and high-dimensional data while capturing the underlying mixture structure.

Low rank plus sparse matrix optimization with more constraints

To characterize matrix optimization problems more precisely, additional constraints beyond least squares are often required. For noisy low-rank and sparse matrix decomposition with fixed and sampled basis coefficients, the following convex optimization formulation was proposed in Wu (2014):

$$\begin{aligned} \min_{L, S \in \mathbb{R}^{n_1 \times n_2}} \quad & \frac{1}{2m} \|y - \mathcal{R}_\Omega(L + S)\|_2^2 + \rho_{L_0} \|L\|_* + \rho_{S_0} \|S\|_1 \\ \text{s.t.} \quad & \mathcal{R}_\mathcal{F}(L + S) = \mathcal{R}_\mathcal{F}(\bar{X}), \\ & \|L\|_\infty \leq b_L, \|S\|_\infty \leq b_S, \end{aligned} \tag{1.9}$$

where $\mathcal{R}_\mathcal{F}$ and $\mathcal{R}_\Omega$ are two linear mappings.

Alternatively, the least squares term can be incorporated as a constraint, turning the equality condition $\mathcal{RF}(L + S) = \mathcal{RF}(\bar{X})$ into a hard constraint (similar to constraints on diagonal elements or trace). Building on the adaptive nuclear semi-norm penalization technique from Miao (2013), we propose the following refined model:

$$\begin{aligned}
& \min_{L, S \in \mathbb{V}^{n_1 \times n_2}} \left(\|L\|_* - \langle F(\dot{L}), L \rangle \right) + \rho_o \left(\|S\|_1 - \langle G(\dot{S}), S \rangle \right) \\
& \text{s.t. } \mathcal{RF}(L + S) = \mathcal{RF}(\bar{X}), \\
& \|y - \mathcal{R}_\Omega(L + S)\|_2^2 \leq \rho_c, \\
& \|L\|_\infty \leq b_L, \|S\|_\infty \leq b_S.
\end{aligned} \tag{1.10}$$

This formulation enhances flexibility by decoupling the fidelity term from the objective and incorporating adaptive penalties for the low-rank and sparse components. The constraints ensure consistency with observed data, structural priors, and boundedness, making the model robust to noise and outliers. Such approaches are particularly useful in applications like robust PCA, matrix completion, and signal processing, where explicit control over rank, sparsity, and data fidelity is critical.

1.2 Related approaches

In this section, we review key algorithms for nuclear norm minimization problems, covering algorithms for regularized problems, level set methods, and adaptive sieving strategies. We focus on their theoretical foundations, computational trade-offs, and applicability to different problem scales.

1.2.1 Algorithms for the regularized problem (LS(λ)-P)

The nuclear norm regularized least squares problem (LS(λ)-P) has been extensively studied over the past decades. Below, we categorize algorithmic approaches based on problem scale and structural assumptions.

The nuclear norm was first introduced as a convex surrogate for rank minimization in control theory by Fazel et al. (2001), who reformulated it as a semidefinite program (SDP). This equivalence enabled the use of off-the-shelf SDP solvers, particularly for problems with linear equality constraints. The SDP framework remains a benchmark for testing solver efficiency, especially for methods exploiting primal low-rank decompositions (Han et al., 2024; Monteiro et al., 2024). However, SDP-based approaches may suffer from scalability limitations in modern huge scale problems.

The nuclear norm’s proximal operator admits an efficient closed-form solution through singular value thresholding (SVT) (Cai et al., 2010), making first-order methods particularly attractive for large-scale optimization problems. This key computational property enabled the development of efficient proximal gradient algorithms, most notably the accelerated proximal gradient (APG) method introduced by Toh and Yun (2010) specifically for solving $(\text{LS}(\lambda)\text{-P})$. Their implementation, which was made publicly available, established a benchmark for nuclear norm regularized problems and inspired numerous extensions, including applications to matrix completion (Ma et al., 2011). Another widely adopted first-order approach is the Alternating Direction Method of Multipliers (ADMM), which has proven effective for various nuclear norm minimization problems. The method’s versatility is demonstrated in applications ranging from system identification (Fazel et al., 2013) to noisy structured low-rank matrix recovery (Bi et al., 2020). Notably, ADMM serves as the foundational first-phase algorithm in the Semismooth Newton-based Augmented Lagrangian (SSNAL) method (Jiang et al., 2013, 2014). Each ADMM iteration requires solving one linear system and computing a proximal step, making it particularly suitable for problems with special structure that can be exploited for efficient linear system solutions. While these first-order methods are computationally efficient per iteration - typically requiring only partial eigenvalue decompositions through truncated SVD or randomized methods - they often suffer from slow asymp-

otic convergence. This characteristic manifests in the need for a large number of iterations to achieve high-precision solutions, particularly for ill-conditioned problems.

For larger scale instances, proximal point algorithms (PPAs) gained prominence due to their stability and convergence guarantees. Liu et al. (2012) proposed three implementable PPA variants, solving subproblems via proximal gradient methods. The semismooth/smoothing Newton augmented Lagrangian method (SSNAL) developed in Jiang et al. (2013, 2014) represents a significant advancement in solving nuclear norm regularized optimization problems with affine constraints. Theoretical analysis reveals that the outer iteration sequence, whether viewed as PPA for the primal or ALM for the dual, achieves asymptotic superlinear convergence under standard regularity conditions including Lipschitz homeomorphism or metric subregularity properties. The PPA/ALM subproblem is solved by the semismooth Newton method (smoothing Newton method if affine constraints exist), which exhibits local superlinear convergence rates. Computational efficiency is achieved through several key features. The method capitalizes on sparse Hessian structures that emerge naturally in the subproblem formulations, allowing for efficient solution of the Newton systems. Specialized linear algebra routines further enhance performance by exploiting problem-specific structure. These properties make SSNAL particularly effective for medium-scale problems where full eigen-decompositions remain computationally tractable. However, the method faces scalability limitations when applied to very large matrices, as the required eigen-decompositions become prohibitively expensive.

To circumvent the computation bottleneck of full matrix decompositions, a common reformulation replaces the nuclear norm with a non-convex factorization:

$$\min_{L \in \mathbb{R}^{m \times r}, R \in \mathbb{R}^{n \times r}} \frac{\lambda}{2} (\|L\|^2 + \|R\|^2) + \frac{1}{2} \|\mathcal{A}(LR^T) - b\|^2, \quad (1.11)$$

where $r \ll \min(m, n)$. Though this sacrifices convexity, it reduces storage from

$O(mn)$ to $O((m+n)r)$. Such factorization techniques have been employed not only in nuclear norm-related problems but also in semidefinite programming, as evidenced by works like Burer and Monteiro (2003), Burer and Monteiro (2005), and Lee et al. (2022). There are some papers studying the convergence of this nonconvex formulations. Actually we can still use the KKT conditions of original convex problem to verify the solution from nonconvex problems.

1.2.2 Level set methods

The level-set method has emerged as a powerful framework for solving constrained convex optimization problems by transforming them into root-finding problems for univariate nonlinear equations. This approach fundamentally reorganizes the optimization problem by interchanging the roles of the objective and constraint functions, creating a value function that can be analyzed through standard root-finding techniques. The methodology has found widespread application across diverse fields including compressed sensing, matrix completion, and inverse problems, as documented in the comprehensive treatment by Aravkin et al. (2019). Level set methods has been used in diverse fields like basis pursuit (van den Berg and Friedlander, 2009, 2011), Euclidean distance realization (Drusvyatskiy et al., 2017), and seismic data interpolation (Aravkin et al., 2014). When solving the original least-squares constrained problems ($\text{CP}(\varrho)$) with the traditional level set method, the univariate nonlinear equation is constructed as follow:

$$\begin{aligned} \phi(\tau) &:= \|\mathcal{A}X(\tau) - b\| = \varrho, \\ \text{where } X(\tau) &\in \arg \min_X \left\{ \frac{1}{2} \|\mathcal{A}X - b\|^2 \mid \|X\|_* \leq \tau \right\}. \end{aligned} \tag{1.12}$$

While this formulation is theoretically sound, its practical implementation faces computational challenges. The projection steps required by standard solution methods

like projected gradient descent can become prohibitively expensive for large-scale problems, and the constrained nature of the subproblems may lead to slow convergence in practice.

Recent developments have introduced a regularized variant of the level-set method that offers improved computational properties (Li et al., 2018b). By reformulating the problem using a penalized objective with the nuclear norm as a regularizer, this approach transforms the root-finding problem into one based on solutions of regularized subproblems as follow:

$$\varphi(\lambda) := \|\mathcal{A}X^*(\lambda) - b\| = \varrho, \quad (\text{Uni.Eq.})$$

where $X^*(\lambda)$ is an optimal solution of (LS(λ)-P). This univariate function φ is often referred to as the value function. Notably, the nuclear norm regularized least-squares problem in (Uni.Eq.) is often more computationally tractable than its counterpart with a nuclear norm constraint in (1.12). To solve the univariate nonlinear equation (Uni.Eq.), standard root-finding methods such as the bisection or secant method may be employed.

1.2.3 Adaptive sieving strategies

The adaptive sieving strategy, as elucidated in Yuan et al. (2023), capitalizes on the sparsity of solutions to construct small-scale problems, effectively reducing dimensions by restrict the decision variable to a subspace. It functions akin to column generation in linear programming. Considering a typical sparse optimization problem

$$\min_{x \in \mathbb{R}^n} f(x) + \lambda p(x), \quad (1.13)$$

where $f(x)$ is a convex twice continuously differentiable function and $p(x)$ is proper closed convex function used to prompt the sparsity of the solution, adaptive sieving

strategy formulate following subspace constrained problem

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) + \lambda p(x), \\ \text{s.t.} \quad & x_I = 0, \end{aligned} \tag{1.14}$$

where I is an index set. The index set will be adjusted according to the KKT residual. Besides, it has been generalized to structure sparse optimizations, i.e.

$$\begin{aligned} \min_{x \in \mathbb{R}^n, y \in \mathbb{R}^m} \quad & f(x) + \lambda p(y), \\ \text{s.t.} \quad & y = Bx. \end{aligned} \tag{1.15}$$

This approach, known for its aggressive restriction, surpasses conventional screening rules such as those outlined in Tibshirani et al. (2012), Ghaoui et al. (2010), and Zeng et al. (2021) in terms of computational efficiency. Widely utilized across various data science domains, including convex clustering (Yuan et al., 2022), precision matrix estimation (Li et al., 2023), and risk-sensitive statistical learning (Wu et al., 2023), adaptive sieving excels in scenarios seeking sparse solutions.

The adaptive sieving technique has been successfully integrated with the regularized level-set method by Li et al. (2024) to address least-squares constrained sparse optimization problems, which represent the sparse vector counterpart of the nuclear norm constrained formulation (CP(ϱ)). This combined approach specifically targets problems of the form:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & p(x) \\ \text{s.t.} \quad & \|Ax - b\| \leq \rho, \end{aligned}$$

where $p(x)$ denotes a sparsity-inducing regularizer. The regularized level-set method naturally lends itself to this adaptive sieving strategy, as it inherently requires solving a sequence of regularized subproblems (1.13). The sieving mechanism proves

particularly effective in this context by progressively identifying and eliminating inactive variables, thereby substantially reducing the computational burden of each subproblem solution. A significant theoretical contribution of Li et al. (2024) establishes the fast convergence rate of the secant method when applied to solve the resulting univariate equation. The convergence is guaranteed under the condition that the sparsity-inducing objective $p(x)$ constitutes a polyhedral gauge function. The polyhedral structure enables favorable geometric properties that facilitate rapid convergence in the root-finding process. This theoretical advance, combined with the practical efficiency gains from adaptive sieving, makes the hybrid approach particularly attractive for large-scale sparse optimization problems where computational efficiency is paramount.

While adaptive sieving methods have proven effective for sparse vector problems, their extension to low-rank matrix optimization remains challenging due to the more complicated geometry of nuclear norm constraints. Several studies, including those by Hsieh and Olsen (2014), Zhou and Zhao (2015), and Zhong and Xu (2023), have developed screening rules for matrices by projecting the problem onto subspaces spanned by factor matrices $U \in \mathbb{R}^{m \times r}$ and $V \in \mathbb{R}^{n \times r}$. This reduces the original problem to a smaller-scale optimization

$$\min_{S \in \mathbb{R}^{r \times r}} \lambda^k \|S\|_* + \frac{1}{2} \|\mathcal{A}(U' S V'^T) - b\|^2. \quad (1.16)$$

However, these methods tend to be conservative, often requiring careful tuning of the subspace dimension r and potentially missing critical directions in the solution space. While they provide a viable framework, further improvements are needed to enhance their adaptability and efficiency for large-scale low-rank problems.

1.3 Summary of contributions of the thesis

In this paper, we propose an innovative proximal generation based level set method integrated with secant iterations, specifically designed for the least-squares constrained nuclear norm minimization problem, as formulated in equation (CP(ϱ)). Our approach builds upon foundational concepts from prior work on sieving-based secant methods for least-squares constrained minimization, where the objective involves a polyhedral gauge function in Li et al. (2024). Our methodology further characterizes the non-singularity of the value function’s Jacobian at the optimal parameter λ , demonstrating a fast convergence rate for the secant method. Unlike earlier studies that primarily addressed polyhedral settings, our work extends to the non-polyhedral domain, driven by the nuclear norm in the objective function. We provide a comprehensive analysis of the secant method’s convergence properties across a broader class of problems, ensuring applicability to complicated scenarios. Additionally, we introduce a tailored proximal generation strategy to optimize low-rank matrix structures, enhancing computational efficiency and scalability. This approach not only improves performance for large-scale problems but also offers practical insights into matrix optimization tasks commonly encountered in machine learning, signal processing, and related fields.

We summarize the main contributions of this paper as follows:

1. We establish the nonsingularity of the Clarke generalized Jacobian $\partial\psi(\cdot)$ for the projection norm function $\psi(\lambda) := \|\Pi_{\lambda C}(b)\|$, where C is a closed convex set and $0 < \lambda < \Upsilon(b \mid C) < +\infty$. Specifically, we demonstrate that the value function $\varphi(\lambda)$ in (Uni.Eq.), a particular instance of $\|\Pi_{\lambda C}(b)\|$, exhibits desirable properties. We further provide a sufficient condition for the strong semismoothness of $\varphi(\lambda)$, which underpins the rapid convergence of the secant method when applied to (Uni.Eq.). This theoretical advancement ensures robust and efficient

solutions for optimization problems involving projection-based objectives, offering a foundation for broader applications in constrained minimization.

2. To address the computational challenges of solving a sequence of nuclear norm regularized problems (LS(λ)-P), we introduce a novel proximal generation method that exploits the low-rank structure inherent in nuclear norm minimization. By leveraging low-rank decomposition, our approach reduces the dimensionality of the problem, while the incorporation of proximal gradient techniques ensures global convergence. This work represents the first extension of the adaptive sieving strategy from Yuan et al. (2023) to nonpolyhedral and potentially nonconvex settings, marking a significant departure from prior polyhedral-focused methods. Combined with secant iterations, our proximal generation based level set method provides a computationally efficient framework for tackling complicated matrix optimization problems, enhancing scalability and performance.
3. We develop a suite of practical implementation techniques to operationalize our proposed method and conduct extensive numerical experiments on real-world tasks, including matrix regression and matrix completion. Our proximal generation based level set method with secant iterations demonstrates superior performance when benchmarked against several state-of-the-art algorithms. These experiments highlight the method’s robustness, efficiency, and ability to handle large-scale problems, making it a valuable tool for applications in machine learning, data science, and signal processing where low-rank matrix optimization is prevalent.

1.4 Organization of the thesis

The organization of the thesis is as follows. Chapter 2 provides some necessary preliminaries for the subsequent analysis. In Chapter 3 our level set method with secant iterations is proposed. This section also includes the proof of the nonsingularity of the Clarke generalized Jacobian of a general class of univariate functions including $\varphi(\lambda)$ in (Uni.Eq.). This nonsingularity is a crucial aspect for the convergence of our secant method with its inherent significance. Chapter 4 delves into the introduction of a proximal generation method tailored for subproblems within the level set paradigm. In Chapter 5, we delve into conducting numerical experiments focused on low-rank matrix regressions and completions. The thesis is concluded in the final section.

Chapter 2

Preliminaries

In this chapter, we introduce some preliminary definitions and propositions to facilitate further discussions.

2.1 Semismooth analysis

In this section, we introduce some concepts related with semismoothness, directional differentiability, and secant method.

Let spaces $\mathcal{X}$ and $\mathcal{Y}$ be two real Euclidean spaces. The set of linear operators from $\mathcal{X}$ to $\mathcal{Y}$ is denoted as $\mathcal{L}(\mathcal{X}, \mathcal{Y})$. Let $F : \mathcal{O} \rightarrow \mathcal{Y}$ be a locally Lipschitz continuous function where $\mathcal{O} \subset \mathcal{X}$ is an open set. We first discuss its directional differentiability and F(réchet)-differentiability.

Definition 2.1. (*Bonnans and Shapiro, 2013, Definition 2.44*) We say that F is *directionally differentiable at a point $x \in \mathcal{O}$ in a direction $h \in X$ if the limit*

$$F'(x, h) := \lim_{t \downarrow 0} \frac{F(x + th) - F(x)}{t}$$

exists. This limit is called the directional derivative at x in direction h . If F is directionally differentiable at x in every direction $h \in X$, we say that F is directionally differentiable at x .

According to the definition, we know the directional derivative $F'(x, h)$ is positively homogeneous of degree 1 in h , i.e. $F'(x, \lambda h) = \lambda F'(x, h)$ for all $h \in \mathcal{X}$ and all $\lambda \geq 0$.

Definition 2.2. (*Bonnans and Shapiro, 2013, Definition 2.48*) We say that F is directionally differentiable at x in the $F(\text{réchet})$ sense if F is directionally differentiable at x and

$$F(x + h) = F(x) + F'(x, h) + o(\|h\|), \quad h \in X.$$

If, in addition, $F'(x, \cdot)$ is linear and continuous, it is said that F is $F(\text{réchet})$ differentiable at x and $F'(x, \cdot)$ is the $F(\text{réchet})$ derivative at x .

For locally Lipschitz functions, we have the following proposition from Shapiro (1990) on the relationship between directional differentiability and F -directional differentiability.

Proposition 2.1 (Shapiro (1990)). *For a locally Lipschitz function F , directional differentiability and $F(\text{réchet})$ directional differentiability are equivalent.*

The local Lipschitz continuity plays an important role in optimization. Besides the directional differentiability, it also supports the definition of the B-subdifferential and the Clarke generalized Jacobian thorough the following Rademacher's theorem.

Definition 2.3. (*Rockafellar and Wets, 2009, Theorem 9.60, Rademacher's theorem*) Let $F : \mathcal{O} \rightarrow \mathcal{Y}$ be a locally Lipschitz continuous function where $\mathcal{O} \subset \mathcal{X}$ is an open set. Let D be the subset of $\mathcal{O}$ consisting of the points where F is differentiable. Then $\mathcal{O} \setminus D$ is a negligible set. In particular D is dense in $\mathcal{O}$, i.e., $\text{cl } D \supset \mathcal{O}$.

Denote D_F the set of all differentiable points of F on $\mathcal{O}$ and $F'(x)$ as the corresponding Jacobian. The B-subdifferential of $F(\cdot)$ at $x \in \mathcal{O}$ is defined as

$$\partial_B F(x) := \left\{ V \in \mathcal{L}(\mathcal{X}, \mathcal{Y}) \mid \exists \{x^k\} \subseteq D_F \text{ s.t. } \lim_{k \rightarrow \infty} x^k = x \text{ and } \lim_{k \rightarrow \infty} F'(x^k) = V \right\}; \quad (2.1)$$

the Clarke generalized Jacobian of $F(\cdot)$ at $x \in \mathcal{O}$ is defined as the convex hull of B-subdifferentials at this point. i.e.

$$\partial F(x) := \text{conv}(\partial_B F(x)). \quad (2.2)$$

We list some properties of above B-subdifferential and the Clarke generalized Jacobian.

Proposition 2.2. *Let $F : \mathcal{O} \subseteq \mathcal{X} \rightarrow \mathcal{Y}$ be a locally Lipschitz continuous function defined on an open set $\mathcal{O}$, and let $x \in \mathcal{O}$. The following statements hold:*

- (i) $\partial_B F(x)$ is a nonempty compact subset of $\mathbb{R}^{m \times n}$, for any $x \in \mathcal{O}$;
- (ii) $\partial_B F(x)$ is upper semicontinuous at $x \in \mathcal{O}$, i.e., for any $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\partial_B F(y) \subseteq \partial_B F(x) + \varepsilon \mathcal{B}_{m \times n}, \forall y \text{ satisfying } \|y - x\| < \delta$$

where $\mathcal{B}_{m \times n} \subseteq \mathbb{R}^{m \times n}$ is the open unit ball centered at the origin;

- (iii) $\partial F(x)$ is a nonempty compact convex subset of $\mathbb{R}^{m \times n}$, for any $x \in \mathcal{O}$;
- (iv) $\partial F(x)$ is upper semicontinuous at $x \in \mathcal{O}$, i.e., for any $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\partial F(y) \subseteq \partial F(x) + \varepsilon \mathcal{B}_{m \times n}, \forall y \text{ satisfying } \|y - x\| < \delta$$

where $\mathcal{B}_{m \times n} \subseteq \mathbb{R}^{m \times n}$ is the open unit ball centered at the origin;;

- (v) for any convex function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, the Clarke generalized Jacobian is same as the subdifferential in the sense of convex analysis.

Proof. Properties (i) and (ii) follow from the definition of B-subdifferential and the locally Lipschitz continuity. Properties (iii) and (iv) come from (Clarke, 1990, Proposition 2.6.2). Properties (v) comes from (Clarke, 1990, Proposition 2.2.7). $\square$

With above preparations, we review the concept related with semismoothness.

Definition 2.4. Let $F : \mathcal{O} \subset \mathcal{X} \rightarrow \mathcal{Y}$ be a locally Lipschitz continuous function on the open set $\mathcal{O}$. Function F is said to be G -semismooth at $x \in \mathcal{O}$ if for any $h \rightarrow 0$ and $V \in \partial F(x + h)$,

$$F(x + h) - F(x) - Vh = o(\|h\|). \quad (2.3)$$

Let γ be a positive constant. Function F is said to be γ -order G -semismooth at $x \in \mathcal{O}$ if for any $h \rightarrow 0$ and $V \in \partial F(x + h)$,

$$F(x + h) - F(x) - Vh = O(\|h\|^{1+\gamma}). \quad (2.4)$$

Function F is said to be G -semismooth (γ -order G -semismooth) on $\mathcal{O}$ if it is G -semismooth (γ -order G -semismooth) at any point $x \in \mathcal{O}$. Function F is called strongly G -semismooth at x if it is 1-order G -semismooth at x .

If function $F : \mathcal{O} \subset \mathcal{X} \rightarrow \mathcal{Y}$ is both G -semismooth and directionally differentiable at $x \in \mathcal{O}$, it is said to be semismooth at $x \in \mathcal{O}$. The concept of semismoothness on a set is as defined in Definition 2.4. The name G -semismoothness is from Pang et al. (2003).

Remark 2.1. In Definition 2.4, the mapping V is restricted in the Clarke generalized Jacobian ∂F . Actually, the Clarke generalized Jacobian ∂F could be replaced with any nonempty and compact valued, upper-semicontinuous set-valued mapping $T_F : \mathcal{O} \rightrightarrows \mathcal{L}(\mathcal{X}, \mathcal{Y})$.

Similarly as the γ -order G -semismooth (2.4), we can strengthen the directional differentiability as follow. The name calm B-differentiability is from Ding et al. (2014b).

Definition 2.5. Function $F : \mathcal{O} \subset \mathcal{X} \rightarrow \mathcal{Y}$ is said to be γ -order directionally differentiable at $x \in \mathcal{O}$ if for any $h \rightarrow 0$

$$F(x + h) - F(x) - F'(x; h) = O(\|h\|^{1+\gamma}). \quad (2.5)$$

Function F is said to be directionally differentiable (γ -order directionally differentiable) on $\mathcal{O}$ if it is directionally differentiable (γ -order directionally differentiable) at any point $x \in \mathcal{O}$. Function F is called calmly B-differentiable at x if (2.5) holds with $\gamma = 1$.

If function $F : \mathcal{O} \subset \mathcal{X} \rightarrow \mathcal{Y}$ is both γ -order G-semismooth and γ -order directionally differentiable at $x \in \mathcal{O}$, it is said to be γ -order semismooth at $x \in \mathcal{O}$. If function F is both strongly G-semismooth and calmly B-differentiable at $x \in \mathcal{O}$, it is said to be strongly semismooth at $x \in \mathcal{O}$.

Next proposition discusses some equivalent definitions of semismoothness.

Proposition 2.3. *Let $F : \mathcal{O} \subset \mathcal{X} \rightarrow \mathcal{Y}$ be a locally Lipschitz continuous function on the open set $\mathcal{O}$ and $x \in \mathcal{O}$. Following statements are equivalent:*

- (i) F is γ -order semismooth at x , i.e. (2.4) and (2.5) hold ;
- (ii) for any $h \rightarrow 0$ and $V \in \partial F(x + h)$, it holds

$$Vh - F'(x; h) = O(\|h\|^{1+\gamma}); \quad (2.6)$$

- (iii) for any $h \rightarrow 0$ and $x + h \in D_F$, it holds

$$F'(x + h; h) - F'(x; h) = O(\|h\|^{1+\gamma}). \quad (2.7)$$

Proof. The implications (i) $\Rightarrow$ (ii) and (ii) $\Rightarrow$ (iii) are obvious. By the derivation of (Facchinei and Pang, 2003, Theorem 7.4.3) with the Lebesgue integral, we have (iii) $\Rightarrow$ (2.5). With (Sun and Sun, 2002, Theorem 3.7), we can get (2.4) from (2.7) and (2.5). Therefore we have (iii) $\Rightarrow$ (i). This completes the proof. $\square$

Remark 2.2. *Above equivalences also hold if we replace the γ -order semismoothness with semismoothness and substitute $O(\|h\|^{1+\gamma})$ with $o(\|h\|)$.*

Next proposition is about the composition of semismooth functions.

Proposition 2.4. *Suppose that the function $\mathcal{F} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is (γ -order semismooth) semismooth at x and that the function $\mathcal{G} : \mathbb{R}^m \rightarrow \mathbb{R}^n$ is (γ -order semismooth) semismooth at $\mathcal{F}(x)$. Then, the composite function $H = \mathcal{G} \circ \mathcal{F}$ is (γ -order semismooth) semismooth at x .*

Proof. The composition rule of semismooth functions is from (Mifflin, 1977, Theorem 5). The rule of γ -order semismooth functions is from (Fischer, 1997, Proposition 19). $\square$

In the last part of this section, we review some approximation properties on semismoothness and the secant method.

Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be an univariate function. For $x, y \in \mathbb{R}, x \neq y$, the divided difference is defined as

$$\delta_f(x, y) := \frac{f(x) - f(y)}{x - y}.$$

In secant method, we use $\delta_f(x, y)$ rather than $f'(x)$. Following lemma adapted from Potra et al. (1998) analyzes some approximation properties of $\delta_f(x, y)$, which are crucial for the convergence of the secant method.

Lemma 2.1. *Suppose $f : \mathbb{R} \rightarrow \mathbb{R}$ is semismooth at $\bar{x} \in \mathbb{R}$. The lateral derivatives of f at $\bar{x}$ are defined as*

$$\bar{d}^- := -f'(\bar{x}; -1) \quad \text{and} \quad \bar{d}^+ := f'(\bar{x}; 1). \quad (2.8)$$

Then following results hold:

- (i) *the lateral derivatives $\bar{d}^-$ and $\bar{d}^+$ exist, and $\partial_B f(\bar{x}) = \{\bar{d}^-, \bar{d}^+\}$;*
- (ii) *it holds that*

$$\begin{aligned} \bar{d}^- - \delta_f(u, v) &= o(1) && \text{for all } u \uparrow \bar{x}, v \uparrow \bar{x}, \\ \bar{d}^+ - \delta_f(u, v) &= o(1) && \text{for all } u \downarrow \bar{x}, v \downarrow \bar{x}; \end{aligned} \quad (2.9)$$

moreover, if $f(\cdot)$ is γ -order semismooth at $\bar{x}$, then

$$\begin{aligned}\bar{d}^- - \delta_f(u, v) &= O(|u - \bar{x}|^\gamma + |v - \bar{x}|^\gamma) \quad \text{for all } u \uparrow \bar{x}, v \uparrow \bar{x}, \\ \bar{d}^+ - \delta_f(u, v) &= O(|u - \bar{x}|^\gamma + |v - \bar{x}|^\gamma) \quad \text{for all } u \downarrow \bar{x}, v \downarrow \bar{x}.\end{aligned}\tag{2.10}$$

Proof. Part (i) and (2.9) in part (ii) are directly from (Potra et al., 1998, Lemma 2.2) and (Potra et al., 1998, Lemma 2.3). Using the equivalent characterization (2.6) of the γ -order semismoothness, we can derive (2.10) through a similar process as in (Potra et al., 1998, Lemma 2.3). $\square$

2.2 Matrix analysis

In this section, we review some results on nuclear norm from matrix analysis.

Let matrix $X \in \mathbb{R}^{m \times n}$ ($m \leq n$) admit the following singular value decomposition (SVD):

$$X = U \begin{bmatrix} \Sigma & 0 \end{bmatrix} V^T$$

where $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are orthogonal matrices, $\Sigma = \text{Diag}(\sigma_1, \dots, \sigma_m)$, and $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_m \geq 0$ are singular values of X being arranged in non-increasing order. The nuclear norm of X is the sum of singulars, denoted as $\|X\|_*$.

With the aforementioned SVD, the proximal mapping of nuclear norm (with a regularized constant) has the following explicit formula:

$$\begin{aligned}\text{Prox}_{\lambda \|\cdot\|_*}(X) &:= \argmin_Z \frac{1}{2} \|X - Z\|^2 + \lambda \|Z\|_* \\ &= U \begin{bmatrix} \Sigma_\lambda & 0 \end{bmatrix} V^T,\end{aligned}\tag{2.11}$$

where $\Sigma_\lambda = \text{Diag}((\sigma_1 - \lambda)_+, \dots, (\sigma_m - \lambda)_+)$ and $(x)_+ := \max\{x, 0\}$. Splitting the right singular matrix V into $\begin{bmatrix} V_1 & V_2 \end{bmatrix}$ where $V_1 \in \mathbb{R}^{n \times m}$ and $V_2 \in \mathbb{R}^{n \times (n-m)}$, fol-

following identities hold:

$$\begin{bmatrix} 0 & X \\ X^T & 0 \end{bmatrix} = Q \begin{bmatrix} \Sigma & 0 & 0 \\ 0 & -\Sigma & 0 \\ 0 & 0 & 0 \end{bmatrix} Q^T,$$

$$\text{where } Q := \frac{1}{\sqrt{2}} \begin{bmatrix} U & U & 0 \\ V_1 & -V_1 & \sqrt{2}V_2 \end{bmatrix}$$

and

$$\begin{aligned} & \begin{bmatrix} 0 & \text{Prox}_{\lambda\|\cdot\|_*}(X) \\ \text{Prox}_{\lambda\|\cdot\|_*}(X)^T & 0 \end{bmatrix} \\ &= \Pi_{\mathbb{S}_+} \left(\begin{bmatrix} 0 & X \\ X^T & 0 \end{bmatrix} - \lambda \text{Id}_{m+n} \right) - \Pi_{\mathbb{S}_+} \left(- \begin{bmatrix} 0 & X \\ X^T & 0 \end{bmatrix} - \lambda \text{Id}_{m+n} \right). \end{aligned}$$

These identities are straightforward. Then we have the following proposition on the proximal mapping of nuclear norm (also called the soft thresholding operator).

Proposition 2.5. *The proximal mapping of the nuclear norm is strongly semismooth.*

Proof. The projection onto the positive semidefinite cone is strongly G-semismooth (Sun and Sun, 2002) and calmly B-differentiable (Ding et al., 2014a). Then the strong semismoothness of the proximal mapping of nuclear norm is followed from the above identities. $\square$

It is well-known that the dual norm of the nuclear norm is the spectral norm $\|\cdot\|_2$, the largest singular value of the matrix, i.e.

$$\sigma_{\max}(Z) =: \|Z\|_2 = \sup_{\|X\|_* \leq 1} \{\langle Z, X \rangle\}.$$

Before characterizing the conjugate of the nuclear norm, we demonstrate the following proposition on the conjugate of a gauge function $k(x)$, i.e. a non-negative positively homogeneous convex function such that $k(0) = 0$.

Proposition 2.6. *Let $k(x)$ be a gauge function on $\mathbb{R}^n$, it holds that its conjugate function $k^*(y)$ is the indicator function $\delta_{\partial k(0)}(y)$; moreover if $k(x)$ is a norm $\|x\|$, it holds that $\partial\|0\| = \{y \mid \|y\|^* \leq 1\}$, where $\|\cdot\|^*$ is the dual norm of $k(x) = \|x\|$.*

Proof. According to the conjugate function, we have

$$k^*(y) = \sup_x \{\langle y, x \rangle - k(x)\} \geq \langle y, 0 \rangle - k(0) = 0.$$

If $k^*(y) > 0$, it hold $k^*(y) = +\infty$ according to the positive homogeneity. If $k^*(y) = 0$, then $\{\langle y, x \rangle - k(x)\}$ achieves its supreme at $x = 0$, which is equivalent to $y \in \partial k(0)$ according to (Rockafellar, 1970, Theorem 23.5). This prove the first argument in the proposition.

If the gauge function $k(x)$ is a norm $\|x\|$, according to the definition of subgradient, we have the following equivalences:

$$\begin{aligned} \partial\|0\| &\Leftrightarrow \{y \mid \langle x, y \rangle \leq \|x\|, \forall x \in \mathbb{R}^n\} \\ &\Leftrightarrow \{y \mid \langle x, y \rangle \leq 1, \forall \|x\| \leq 1\} \\ &\Leftrightarrow \{y \mid \|y\|^* \leq 1\}. \end{aligned}$$

This completes the proof. □

Thus we have the conjugate function of the nuclear norm is such an indicator function

$$\sup_X \{\langle Y, X \rangle - \|X\|_*\} = \delta_{\partial\|0\|_*}(Y) = \delta_{\{Y \mid \|Y\|_2 \leq 1\}}(Y).$$

With this conjugation, we can write the dual problem of (LS(λ)-P) as follow. First we write an equivalent formula of the regularized problem.

$$\begin{aligned} (\text{LS}(\lambda)\text{-P}) &\Leftrightarrow \min_{X \in \mathbb{R}^{m \times n}, u \in \mathbb{R}^p} \lambda \|X\|_* + \frac{1}{2} \|u\|^2 \\ &\text{s.t. } u = b - \mathcal{A}X. \end{aligned}$$

Then the dual problem is from following Lagrange duality.

$$\begin{aligned}
& \max_{y \in \mathbb{R}^p} \inf_{X \in \mathbb{R}^{m \times n}, u \in \mathbb{R}^p} \frac{1}{2} \|u\|^2 + \lambda \|X\|_* + \langle y, b - \mathcal{A}X - u \rangle \\
& \Leftrightarrow \max_{y \in \mathbb{R}^p} \langle b, y \rangle + \inf_{u \in \mathbb{R}^p} \left\{ \frac{1}{2} \|u\|^2 - \langle y, u \rangle \right\} + \inf_{X \in \mathbb{R}^{m \times n}} \{ \lambda \|X\|_* - \langle \mathcal{A}^* y, X \rangle \} \\
& \Leftrightarrow \max_{y \in \mathbb{R}^p} -\frac{1}{2} \|y\|^2 + \langle b, y \rangle - \delta_{\|\cdot\|_2 \leq \lambda}(\mathcal{A}^* y) \\
& \Leftrightarrow \min_{y \in \mathbb{R}^p} -\langle b, y \rangle + \frac{1}{2} \|y\|^2 \\
& \text{s.t.} \quad \mathcal{A}^* y = \lambda Z, \quad \|Z\|_2 \leq 1.
\end{aligned}$$

From the Moreau decomposition, we also have the explicit formula of the projection onto the spectral norm ball (with the SVD $X = U[\Sigma]V$):

$$\begin{aligned}
\Pi_{\|\cdot\|_2 \leq \lambda}(X) &= \text{Prox}_{\delta_{\{Y \mid \|Y\|_2 \leq 1\}}}(X) \\
&= X - \text{Prox}_{\lambda \|\cdot\|_*}(X) \\
&= U[\min(\Sigma, \lambda)0]V^T,
\end{aligned}$$

where $\min(\Sigma, \lambda) = \text{Diag}(\min(\sigma_1, \lambda), \dots, \min(\sigma_m, \lambda))$. We also know the projection $\Pi_{\|\cdot\|_2 \leq \lambda}(\cdot)$ is strongly semismooth from the Moreau decomposition. Note that every projection onto a closed convex set is the proximal mapping of the corresponding indicator function, we know every projection onto a closed convex set is globally Lipschitz with modulus 1 (also called nonexpansive).

Besides the nuclear norm itself, we also review some results on the epigraphs of nuclear norm and spectral norm. These epigraphs are actually matrix cones:

$$\begin{aligned}
\text{epi } \|\cdot\|_* &:= \{(t, X) \in \mathbb{R}_+ \times \mathbb{R}^{m \times n} \mid t \geq \|X\|_*\} \\
\text{and } \text{epi } \|\cdot\|_2 &:= \{(t, X) \in \mathbb{R}_+ \times \mathbb{R}^{m \times n} \mid t \geq \|X\|_2\}.
\end{aligned}$$

As nuclear norm and spectral norm are dual to each other, we can have the duality relationship between these two cones.

Proposition 2.7. *Cone induced by epigraph of nuclear norm $\text{epi } \|\cdot\|_*$ and cone induced by epigraph of spectral norm $\text{epi } \|\cdot\|_2$ are both closed convex pointed solid cones. The dual cone of $\text{epi } \|\cdot\|_*$ is $\text{epi } \|\cdot\|_2$.*

The strong semismoothness of projection of $\Pi_{\text{epi } \|\cdot\|_*}$ and $\Pi_{\text{epi } \|\cdot\|_2}$ has been established.

Proposition 2.8. *The metric projections $\Pi_{\text{epi } \|\cdot\|_*}$ and $\Pi_{\text{epi } \|\cdot\|_2}$ are strongly semismooth.*

Proof. The claim B-differentiability and strong G-semismoothness of $\Pi_{\text{epi } \|\cdot\|_2}$ are from (Ding et al., 2014b, Theorem 2) and (Ding et al., 2014b, Theorem 4). The strong semismoothness of $\Pi_{\text{epi } \|\cdot\|_*}$ follows from the Moreau decomposition

$$\Pi_{\text{epi } \|\cdot\|_*}(X) = X + \Pi_{\text{epi } \|\cdot\|_2}(-X).$$

□

With the matrix cone of epigraph of spectral norm, the dual problem of the nuclear norm regularized least squares problem can directly be reformulated as follow:

$$\begin{aligned} \min_{y \in \mathbb{R}^p} \quad & \frac{1}{2} \|y\|^2 - b^T y \\ \text{s.t.} \quad & (\lambda, \mathcal{A}^* y) \in \text{epi } \|\cdot\|_2. \end{aligned}$$

And the primal problem can also be formulated from the duality of above conic

programming.

$$\begin{aligned}
& \max_{(s,X) \in \text{epi } \|\cdot\|_*} \inf_y \frac{1}{2} \|y\|^2 - b^T y - \langle (s, -X), (\lambda, \mathcal{A}^* y) \rangle \\
& \Leftrightarrow \max_{(s,X) \in \text{epi } \|\cdot\|_*} -\lambda s + \inf_y \frac{1}{2} \|y\|^2 - b^T y + \langle \mathcal{A} X, y \rangle \\
& \Leftrightarrow \max_{(s,X) \in \text{epi } \|\cdot\|_*} -\lambda s - \frac{1}{2} \|b - \mathcal{A} X\|^2 \\
& \Leftrightarrow \min_{(s,X) \in \mathbb{R}_+ \times \mathbb{R}^{m \times n}} \frac{1}{2} \|\mathcal{A} X - b\|^2 + \lambda s \\
& \text{s.t. } (s, X) \in \text{epi } \|\cdot\|_*,
\end{aligned}$$

which is equivalent to the original nuclear norm regularized least squares problem (LS(λ)-P).

2.3 Sensitivity analysis

In this section, we discuss some optimality conditions and sensitivity analysis results related with our nonpolyhedral constraint system, especially the optimization constrained by the epigraph of spectral norm.

Let $G : \mathcal{X} \rightarrow \mathcal{Y}$ be a twice continuously differentiable mapping and $K \subset \mathcal{Y}$ is a closed convex set. We mainly consider the constraint system

$$G(x) \in K. \tag{2.12}$$

For later discussions, we first introduce some definitions,

Definition 2.6. *For a closed convex set K and a point $y \in K$, we define the tangent cone of K at y as follow:*

$$T_K(y) := \{d \in \mathcal{Y} \mid \text{dist}(y + td, K) = o(t), t \geq 0\};$$

the second order tangent cone is defined as

$$T_K^2(y, d) = \left\{ w \in \mathcal{Y} \mid \exists t_k \downarrow 0, \text{dist} \left(y + t_k d + \frac{1}{2} t_k^2 w, K \right) = o(t_k^2) \right\}$$

Definition 2.7. The linearity space of a closed convex cone C , denoted as $\text{lin}(C)$, is the largest linear space in cone C . It can be expressed explicitly as follow:

$$\text{lin}(C) := C \cap (-C).$$

With above preparations, we can introduce the concept of constraint nondegeneracy, which is the generalization of linear independent constraint qualification (LICQ) to general constraint system (2.12).

Definition 2.8. We say a feasible point $\bar{x}$ to constraint system $G(x) \in K$ is constraint nondegenerate if

$$G'(\bar{x})\mathcal{X} + \text{lin}(T_K(G(\bar{x}))) = \mathcal{Y}. \quad (2.13)$$

We now focus on the matrix optimization constrained by the epigraph of the spectral norm. The conic form of the dual of the nuclear norm regularized least squares (LS(λ)-P) is actually in this form. Following theoretical results build primarily upon the seminal work Ding (2012) which established fundamental results on Ky Fan k -norms (note that the spectral norm is the Ky Fan 1-norm and the nuclear norm is the Ky Fan n -norm) and Guo et al. (2015) which focused on the spectral norm constrained matrix optimization.

Following we consider the following spectral norm constrained optimization:

$$\begin{aligned} \min_{y \in \mathbb{R}^p} \quad & f(y) \\ \text{s.t.} \quad & (g(y), G(y)) \in \text{epi} \|\cdot\|_2^{m \times n}, \end{aligned} \quad (2.14)$$

where $f : \mathbb{R}^p \rightarrow \mathbb{R}$, $g : \mathbb{R}^p \rightarrow \mathbb{R}$, and $G : \mathbb{R}^p \rightarrow \mathbb{R}^{m \times n}$ are twice continuously differentiable functions. The corresponding Lagrangian function is

$$L(y, s, X) = f(y) - \langle (s, -X), (g(y), G(y)) \rangle \text{ for } (y, (s, X)) \in \mathbb{R}^p \times \text{epi } \|\cdot\|_*;$$

and the KKT conditions are

$$\begin{aligned} L'_y(y, s, X) &= 0; \\ (-s, X) &\in N_{\text{epi } \|\cdot\|_2}(g(y), G(y)). \end{aligned}$$

The KKT conditions can be equivalently formulated as the following nonsmooth equation:

$$F(y, s, X) = \begin{bmatrix} L'_y(y, s, X) \\ -(g(y), G(y)) + \Pi_{\text{epi } \|\cdot\|_2}(g(y) - s, G(y) + X) \end{bmatrix} = 0, \quad (2.15)$$

Let matrix $X \in \mathbb{R}^{m \times n}$ admit the SVD in Section 2.2. We first define following matrix operators for convenience.

$$\begin{aligned} S(X) &:= \frac{1}{2} (X + X^T) \\ D(X) &:= \frac{1}{2} (X - X^T), \forall X \in \mathbb{R}^{m \times m}, \\ \text{and } \mathcal{B}(X) &:= \begin{bmatrix} 0_{m \times m} & X \\ X^T & 0_{n \times n} \end{bmatrix}, \forall X \in \mathbb{R}^{m \times n}. \end{aligned}$$

The tangent cone of the epigraph of the spectral graph is characterized as follow.

Proposition 2.9 (Ding (2012); Guo et al. (2015)). *Suppose $\bar{X}$ has the SVD as in Section 2.2, then the tangent cone of $\text{epi } \|\cdot\|_2$ is characterized by*

$$T_{\text{epi } \|\cdot\|_2}(\bar{t}, \bar{X}) = \begin{cases} \mathbb{R} \times \mathbb{R}^{m \times n}, & \text{if } (\bar{t}, \bar{X}) \in \text{int } \text{epi } \|\cdot\|_2 \\ \text{epi } \|\cdot\|_2, & \text{if } (\bar{t}, \bar{X}) = (0, 0_{m \times n}) \\ \{(\tau, H) : S(U_{\alpha_1}^T H V_{\alpha_1}) \preceq \tau I_{|\alpha_1|}\}, & \text{if } (\bar{t}, \bar{X}) \in \text{bd } \text{epi } \|\cdot\|_2 \setminus \{(0, 0_{m \times n})\} \end{cases}$$

where $\alpha_1 = \{i : \sigma_i(\bar{X}) = \bar{t}\}$. And the largest linear space contained in $T_{\text{epi} \|\cdot\|_2}(\bar{t}, \bar{X})$, denoted by $\text{lin}(T_{\text{epi} \|\cdot\|_2}(\bar{t}, \bar{X}))$, can be characterized by

$$\text{lin}(T_{\text{epi} \|\cdot\|_2}(\bar{t}, \bar{X})) = \begin{cases} \mathbb{R} \times \mathbb{R}^{m \times n}, & \text{if } (\bar{t}, \bar{X}) \in \text{int epi } \|\cdot\|_2 \\ \{(0, 0_{m \times n})\}, & \text{if } (\bar{t}, \bar{X}) = (0, 0_{m \times n}) \\ \{(\tau, H) : S(U_{\alpha_1}^T H V_{\alpha_1}) = \tau I_{|\alpha_1|}\}, & \text{if } (\bar{t}, \bar{X}) \in \text{bd epi } \|\cdot\|_2 \setminus \{(0, 0_{m \times n})\} \end{cases}$$

For the spectral norm constrained optimization, the constraint nondegeneracy has a following characterization.

Proposition 2.10 (Zhang et al. (2013); Guo et al. (2015)). *Let $\bar{y}$ be a feasible point of $(g(y), G(y)) \in \text{epi } \|\cdot\|_2^{m \times n}$ and $(g(\bar{y}), G(\bar{y})) \in \text{bd epi } \|\cdot\|_2 \setminus \{(0, 0_{m \times n})\}$. Suppose $G(\bar{y})$ has SVD as in Section 2.2. Then, $\bar{y}$ is constraint nondegenerate if and only if the following p -dimensional vectors are linearly independent:*

$$e_{ij} := \begin{pmatrix} U_i^T \frac{\partial G(\bar{y})}{\partial y_1} V_j + U_j^T \frac{\partial G(\bar{y})}{\partial y_1} V_i \\ U_i^T \frac{\partial G(\bar{y})}{\partial y_2} V_j + U_j^T \frac{\partial G(\bar{y})}{\partial y_2} V_i \\ \vdots \\ U_i^T \frac{\partial G(\bar{y})}{\partial y_i} V_j + U_j^T \frac{\partial G(\bar{y})}{\partial y_i} V_i \end{pmatrix}, \quad i < j, i, j \in \alpha_1,$$

and

$$e_{ii} := \begin{pmatrix} U_i^T \frac{\partial G(\bar{y})}{\partial y_1} V_i - \frac{\partial g(\bar{y})}{\partial y_1} \\ U_i^T \frac{\partial G(\bar{y})}{\partial y_2} V_i - \frac{\partial g(\bar{y})}{\partial y_2} \\ \vdots \\ U_i^T \frac{\partial G(\bar{y})}{\partial y_i} V_i - \frac{\partial g(\bar{y})}{\partial y_i} \end{pmatrix}, \quad i \in \alpha_1,$$

where $\alpha_1 = \{i : \sigma_i(\bar{X}) = \bar{t}\}$.

Before we demonstrate the strong second order sufficient condition, we define the critical cone of the epigraph of spectral norm at the KKT point $(\bar{y}, \bar{s}, \bar{X})$ as follow:

$$C_{\text{epi} \|\cdot\|_2}(\bar{y}, \bar{s}, \bar{X}) = T_{\text{epi} \|\cdot\|_2}(g(\bar{y}), G(\bar{y})) \cap (-\bar{s}, \bar{X})^\perp;$$

the critical cone of the programming (2.14) at the KKT point $(\bar{y}, \bar{s}, \bar{X})$ is

$$C(\bar{y}) = \{d \in \mathbb{R}^p : (D_y g_0(\bar{y})d, D_y G(\bar{y})d) \in \mathcal{C}_{\text{epi } \|\cdot\|_2}(\bar{y}, \bar{s}, \bar{X})\};$$

and an outer approximation set of $\text{aff}(C(\bar{y}))$ as

$$\text{app}(\bar{s}, \bar{X}) = \{d \in \mathbb{R}^p : (D_y g_0(\bar{y})d, D_y G(\bar{y})d) \in \text{aff } \mathcal{C}_{\text{epi } \|\cdot\|_2}(\bar{y}, \bar{s}, \bar{X})\}.$$

Definition 2.9 (Guo et al. (2015)). *Let $\bar{y}$ be a stationary point of the spectral norm constrained optimization (2.14). We say that the strong second order sufficient condition holds at $\bar{y}$ if*

$$\begin{aligned} & \sup_{(s, X) \in \Lambda(\bar{y})} \left\{ \langle d, L''_{yy}(\bar{y}, s, X)d \rangle \right. \\ & \left. - \sigma \left((-s, X), T_{\text{epi } \|\cdot\|_2}^2((g(\bar{y}), G(\bar{y})), (g'(\bar{y})d, G'(\bar{y})d)) \right) \right\} > 0, \quad \forall d \in C'(\bar{y}) \setminus \{0\} \end{aligned}$$

where

$$C'(\bar{y}) := \bigcap_{(s, X) \in \Lambda(\bar{y})} \text{app}(s, X)$$

and $\Lambda(\bar{y})$ is the set of (s, X) s.t. $(\bar{y}, s, X)$ is a KKT point of (2.14).

It is obvious that if $f(y)$ is strongly convex, then the strong second order sufficient condition is always held at the KKT point.

With the nondegeneracy and the strong second order sufficient condition, we can consider the stability of the KKT system of (2.14).

Proposition 2.11 (Guo et al. (2015)). *Let $\bar{y}$ be a locally optimal solution of (2.14) and $(\bar{s}, \bar{X})$ be a Lagrange multiplier at $\bar{y}$. If the strong second-order sufficient condition holds at $\bar{y}$ and $\bar{y}$ is constraint nondegenerate, then any element in $\partial F(\bar{y}, \bar{s}, \bar{X})$ is nonsingular.*

To analyze the superlinear convergence, we also give an inverse function theorem on local Lipschitz homeomorphism.

Definition 2.10. *A function $\Phi : \mathcal{O} \subseteq \mathcal{X} \rightarrow \mathcal{X}$ is said to be a locally Lipschitz homeomorphism near $x \in \mathcal{O}$ if there exists an open neighborhood $\mathcal{N} \subseteq \mathcal{O}$ of x such that the restricted map $\Phi|_{\mathcal{N}} : \mathcal{N} \rightarrow \Phi(\mathcal{N})$ is Lipschitz continuous and bijective, and its inverse is also Lipschitz continuous.*

From Clarke's inverse function theorem (Clarke, 1976, 1990), we know the non-smooth function $F(y, s, X)$ in (2.15) is a locally Lipschitz homeomorphism near $(\bar{y}, \bar{s}, \bar{X})$ if any element in $\partial F(\bar{y}, \bar{s}, \bar{X})$ is nonsingular.

To further facilitate our analysis in subsequent chapters, we also demonstrate an inverse function theorem on semismoothness.

Proposition 2.12. *Let function $\Phi : \mathcal{O} \subseteq \mathcal{X} \rightarrow \mathcal{X}$ be a locally Lipschitz homeomorphism near $x \in \mathcal{O}$. Then there exists an open neighborhood $\mathcal{N} \subseteq \mathcal{O}$ of x such that*

(i) *Φ is G -semismooth (γ -order G -semismooth) at some point $\bar{x} \in \mathcal{N}$ if and only if Φ^{-1} , the local inverse mapping of Φ near x_0 , is G -semismooth (γ -order G -semismooth) at $\bar{y} := \Phi(\bar{x})$;*

(ii) *Φ is directionally differentiable (γ -order directionally differentiable) at some point $\bar{x} \in \mathcal{N}$ if and only if Φ^{-1} , the local inverse mapping of Φ near x_0 , is directionally differentiable (γ -order directionally differentiable) at $\bar{y} := \Phi(\bar{x})$*

Proof. The inverse function theorems on G -semismoothness and directional differentiability are sourced from (Meng et al., 2005, Theorem 2). On γ -order G -semismoothness, the derivation is same as the strong G -semismoothness case in (Meng et al., 2005, Theorem 2). Our current focus lies in establishing the γ -order directionally differentiability.

Note that Φ is also a locally Lipschitz homeomorphism near $\bar{x}$. We denote $\Psi(\cdot) := \Phi'(\bar{x}; \cdot)$. Shrinking the open neighborhood $\mathcal{N}$ if necessary, for any $y \in \Phi(\mathcal{N})$ and $y \rightarrow \bar{y}$, we have

$$\begin{aligned}
& \Phi^{-1}(y) - \Phi^{-1}(\bar{y}) - (\Phi^{-1})'(\bar{y}; y - \bar{y}) \\
&= x - \bar{x} - \Psi^{-1}(y - \bar{y}) \\
&= -\Psi^{-1}(y - \bar{y} - \Psi(x - \bar{x})) \\
&= -\Psi^{-1}(\Phi(x) - \Phi(\bar{x}) - \Phi'(\bar{x}; x - \bar{x})) \\
&= O(\Phi(x) - \Phi(\bar{x}) - \Phi'(\bar{x}; x - \bar{x})),
\end{aligned} \tag{2.16}$$

where $x = \Phi^{-1}(y)$. The relation between $(\Phi^{-1})'(\bar{y}; \cdot)$ and $\Psi^{-1}(\cdot)$ is from (Kummer, 1992, Lemma 2). The last equation is guaranteed by the Lipschitz homeomorphism property of $\Psi(\cdot)$ (Pang et al., 2003, Theorem 6).

The other direction of the theorem on the γ -order directionally differentiability can be obtained by revising the derivation (2.16). This completes the proof of the proposition. $\square$

Every inverse function theorem has its implicit function theorem counterpart for $H(x, u)$ by applying the inverse function theorem to

$$\Psi(x, u) := \begin{bmatrix} H(x, u) \\ u \end{bmatrix}.$$

The implicit function theorem counterpart of Proposition 2.12 is as follow.

Corollary 2.1. *Let $H(x, u) : \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{X}$ be a Lipschitz function. Suppose that every element in $\partial_x H(\bar{x}, \bar{u})$ is nonsingular and $H(\bar{x}, \bar{u}) = 0$. Then, there exist an open neighborhood $\mathcal{N} \subseteq Y$ of $\bar{u}$ and a Lipschitz continuous function $x(\cdot)$ defined on $\mathcal{N}$ such that $H(x(u), u) = 0$ for every $u \in \mathcal{N}$. Moreover, $x(\cdot)$ is G -semismooth (γ -order G -semismooth, directional differentiable, γ -order directional differentiable) at some*

point $u \in \mathcal{N}$ if H is G -semismooth (γ -order G -semismooth, directional differentiable, γ -order directional differentiable) at $(x(u), u)$.

Remark 2.3. *This section reviews preliminary results on optimality conditions and sensitivity analysis for spectral norm constrained optimization problems. As the dual norm of the nuclear norm, the spectral norm belongs to the broader class of Ky Fan k -norms. Fundamental variational analysis results including tangent cones, critical cones, and second-order optimality conditions for general Ky Fan k -norm constrained problems were established in Ding (2012, 2017). For the specific case of spectral norm constraints (Ky Fan 1-norm), Guo et al. (2015) further investigated strong second-order sufficient conditions and proved the nonsingularity of Clarke's generalized Jacobian for the associated nonsmooth mappings. The sensitivity analysis framework builds on early work by Robinson (1980); Dontchev and Rockafellar (1996) for polyhedral cases and Sun (2006) for semidefinite programming. Recent advances have extended these results to weaker Lipschitz properties such as robust isolated calmness (Ding et al., 2017) and Hölderian stability (Mordukhovich et al., 2015). Several equivalent characterizations have been developed, including the Aubin property (Chen et al., 2025), strong variational sufficiency (Wang et al., 2023), and full stability (Mordukhovich et al., 2015). For comprehensive treatments of nuclear and spectral norm optimization problems, we refer readers to Cui and Sun (2018); Cui et al. (2024).*

Chapter 3

A level set method with secant iterations for the least-squares constrained nuclear norm minimization

Recall that we aim to solve the least-squares constrained nuclear norm minimization (CP(ϱ)).

$$\begin{aligned} \min_{X \in \mathbb{R}^{m \times n}} \quad & \|X\|_* \\ \text{s.t.} \quad & \|\mathcal{A}X - b\| \leq \varrho. \end{aligned} \tag{CP(\varrho)}$$

We adopt the level set method from Li et al. (2018b) which transfer the least squares constrained problem into a univariate nonlinear equation

$$\varphi(\lambda) := \|\mathcal{A}X^*(\lambda) - b\| = \varrho, \tag{Uni.Eq.}$$

where $X^*(\lambda)$ is an optimal solution of (LS(λ)-P):

$$\min_{X \in \mathbb{R}^{m \times n}} f_\lambda(X) := \lambda \|X\|_* + \frac{1}{2} \|\mathcal{A}X - b\|^2. \tag{LS(\lambda)-P}$$

In this chapter, we begin by demonstrating several properties of the value function $\varphi(\lambda)$ in the univariate equation (Uni.Eq.) and the level set method from Li et al. (2018b). Then we prove the nonsingularity of the Clarke generalized Jacobian of a

class of univariate function including $\varphi(\lambda)$. This nonsingularity is a crucial aspect for the convergence of our secant method with its inherent significance. Finally we analyze the semismoothness of $\varphi(\lambda)$ and develop a secant method for (Uni.Eq.).

3.1 A level set method with the value function $\varphi(\lambda)$

The dual problem of the regularized problem (LS(λ)-P) takes the following form (derivation could be found in Chapter 2.2):

$$\begin{aligned} \min_{y \in \mathbb{R}^p, Z \in \mathbb{R}^{m \times n}} \quad & \frac{1}{2} \|y\|^2 - b^T y \\ \text{s.t.} \quad & \mathcal{A}^* y = \lambda Z, \|Z\|_2 \leq 1. \end{aligned} \tag{LS(\lambda)-D}$$

The KKT conditions of (LS(λ)-P) and (LS(λ)-D) are:

$$\begin{aligned} Z &\in \partial \|X\|_*, \\ y - b + \mathcal{A}X &= 0, \\ \text{and} \quad \mathcal{A}^* y &= \lambda Z. \end{aligned} \tag{3.1}$$

Besides, the dual problem can be equivalently formulated as

$$\begin{aligned} \min_{y \in \mathbb{R}^p} \quad & \frac{1}{2} \|y\|^2 - b^T y \\ \text{s.t.} \quad & \lambda^{-1} y \in Q, \end{aligned} \tag{3.2}$$

where

$$Q := \{z \in \mathbb{R}^p \mid \mathcal{A}^* z \in \partial \|0\|_*\} = \{z \in \mathbb{R}^p \mid \|\mathcal{A}^* z\|_2 \leq 1\}.$$

We denote the primal dual solutions of (LS(λ)-P) and (LS(λ)-D) as $(X(\lambda), y(\lambda), Z(\lambda))$.

With the formula (3.2), the dual solution $y(\lambda)$ can be explicitly expressed as

$$y(\lambda) = \lambda \Pi_Q(\lambda^{-1} b) = \Pi_{\lambda Q}(b).$$

Then the value function $\varphi(\lambda)$ can be equivalently expressed as

$$\varphi(\lambda) = \|\mathcal{A}X(\lambda) - b\| = \|y(\lambda)\| = \|\Pi_{\lambda Q}(b)\|. \quad (3.3)$$

To justify the level set method for $(\text{CP}(\varrho))$, we recall key properties of the value function $\varphi(\lambda)$, which align with those established in (Li et al., 2018b, Proposition 1) and (Li et al., 2024, Propositions 3.1, 3.3). Here the set Q is not restricted as a polyhedral set. Notably, the bound $\|\mathcal{A}^*b\|_2$ corresponds exactly to the gauge function $\Upsilon(b \mid Q)$. For completeness, we provide an independent proof of these properties, offering a more direct argument for part (iii) in the nuclear norm scenario.

Proposition 3.1. *Suppose that $0 < \|\mathcal{A}^*b\|_2$. The following statements hold:*

- (i) *For any $\lambda \geq 0$, the quantity $\varphi(\lambda)$ is independent of the particular choice of $X(\lambda)$ solving (LS(λ)-P);*
- (ii) *For any $\lambda > \|\mathcal{A}^*b\|_2$, $y(\lambda) = b$ and $X(\lambda) = 0$;*
- (iii) *For any $0 < \lambda_1 < \lambda_2 \leq \|\mathcal{A}^*b\|_2$, $\varphi(\lambda_1) < \varphi(\lambda_2)$.*

Proof. (i) The results follow from the KKT conditions (3.1) and the strong convexity of the dual objective function (LS(λ)-D).

(ii) For the dual objective function, it holds that

$$\frac{1}{2}\|y\|^2 - b^T y \geq -\frac{1}{2}\|b\|^2.$$

Then it could be verified that $y(\lambda) = b$ and $X(\lambda) = 0$ are the solution of (LS(λ)-P) and (LS(λ)-D).

(iii) For any $0 < \lambda_1 < \lambda_2 \leq \|\mathcal{A}^*b\|_2$, let $X(\lambda_1)$ and $X(\lambda_2)$ be two corresponding solutions.

We first prove the monotonicity (maybe not strict) of the value function $\phi(\lambda)$. From the optimality of $X(\lambda_1)$ and $X(\lambda_2)$, following inequalities hold:

$$\begin{aligned}\frac{1}{2} \|\mathcal{A}X(\lambda_1) - b\|^2 + \lambda_1 \|X(\lambda_1)\|_* &\leq \frac{1}{2} \|\mathcal{A}X(\lambda_2) - b\|^2 + \lambda_1 \|X(\lambda_2)\|_*, \\ \frac{1}{2} \|\mathcal{A}X(\lambda_2) - b\|^2 + \lambda_2 \|X(\lambda_2)\|_* &\leq \frac{1}{2} \|\mathcal{A}X(\lambda_1) - b\|^2 + \lambda_2 \|X(\lambda_1)\|_*.\end{aligned}$$

Combine the above inequalities, we have

$$(\lambda_1 - \lambda_2) (\|X(\lambda_1)\|_* - \|X(\lambda_2)\|_*) \leq 0.$$

Thus we have the nuclear norm $\|X(\lambda)\|_*$ is nonincreasing. With the first above inequality, we have the monotonicity of the value function:

$$\begin{aligned}\frac{1}{2} \|\mathcal{A}X(\lambda_1) - b\|^2 &\leq \frac{1}{2} \|\mathcal{A}X(\lambda_2) - b\|^2 + \lambda_1 (\|X(\lambda_2)\|_* - \|X(\lambda_1)\|_*) \\ &\leq \frac{1}{2} \|\mathcal{A}X(\lambda_2) - b\|^2.\end{aligned}\tag{3.4}$$

Then we prove the strict monotonicity by contradiction. Suppose that $\varphi(\lambda_1) = \varphi(\lambda_2)$. From (3.4), we have

$$\|X(\lambda_2)\|_* = \|X(\lambda_1)\|_*,$$

which means $X(\lambda_1)$ is also a solution of (LS(λ)-P) with $\lambda = \lambda_2$. Therefore we have

$$y(\lambda_1) = \mathcal{A}X(\lambda_2) - b = y(\lambda_2).$$

Note that

$$\|\mathcal{A}^*y(\lambda_2)\|_2 = \|\mathcal{A}^*y(\lambda_1)\|_2 \leq \lambda_1 < \lambda_2 \leq \|\mathcal{A}^*b\|_2;$$

we know $y(\lambda_2)$ is an interior point of the constraint set λQ . Therefore we have

$$y(\lambda_2) - b \in N_{\lambda_2 Q}(y(\lambda_2)) = \{0\},$$

Algorithm 1 A level set method for $(\text{CP}(\varrho))$

Require: Initial points λ^{-1} , λ^0 , λ^1 and bounds λ_{lo} , λ_{up} such that $0 \leq \lambda_{\text{lo}} \leq \lambda^{-1} < \lambda^0 \leq \lambda_{\text{up}} \leq \|\mathcal{A}^*b\|_2$ and $\lambda_{\text{lo}} \leq \lambda^1 \leq \lambda_{\text{up}}$.

Ensure: A solution λ^* to $\varphi(\lambda) = \varrho$ and the corresponding $X(\lambda^*)$.

- 1: **for** $k = 1, 2 \dots$ **do**
- 2: Compute

$$X^k = \underset{X \in \mathbb{R}^{m \times n}}{\operatorname{argmin}} \lambda^k \|X\|_* + \frac{1}{2} \|\mathcal{A}X - b\|^2 \quad (\text{LS}(\lambda^k)\text{-P})$$

- 3: Compute $\varphi(\lambda^k) = \|\mathcal{A}X^k - b\|$.
 - 4: Generate λ^{k+1} with the bisection method or the secant method.
 - 5: **end for**
-

which means

$$y(\lambda_1) = y(\lambda_2) = b \in \lambda_1 Q.$$

It contradicts with the condition $\lambda_1 < \lambda_2 \leq \Upsilon(b \mid Q)$. This completes the proof. $\square$

With the strict monotonicity of value function $\varphi(\lambda)$, we have the following proposition.

Proposition 3.2. *Suppose that the constrained problem $(\text{CP}(\varrho))$ is feasible and its least-squares constraint is active at every optimal solution. Then solving $(\text{CP}(\varrho))$ is equivalent to finding a root λ^* of the univariate nonlinear equation (Uni.Eq.). Specifically, any solution of $(\text{CP}(\varrho))$ is given by $X^*(\lambda^*)$ in (Uni.Eq.).*

This proposition justifies the level set method (Algorithm (1)) from Li et al. (2018b). Algorithm 1, first introduced in Li et al. (2018b), offers substantial computational advantages over conventional level set approaches (Aravkin et al., 2019; van den Berg and Friedlander, 2009, 2011). The reformulated subproblem $(\text{LS}(\lambda^k)\text{-P})$ combines two key benefits: its regularized least-squares structure facilitates our proximal generation strategy (developed in Chapter 4), while simultaneously enabling efficient implementation of the secant method (presented in Section 3.3). These

advantages have the potential to significantly reduce both the number of required subproblems to solve and the computational cost per subproblem.

3.2 Nonsingularity of the Clarke generalized Jacobian $\partial\varphi(\lambda)$

In this section, we investigate the nonsingularity of the the Clarke generalized Jacobian $\partial\varphi(\lambda)$, which plays an important role in the superlinear convergent rate of the secant method.

We consider a broader univariate function

$$\psi(\lambda) := \|\lambda\Pi_C(\lambda^{-1}b)\| = \|\Pi_{\lambda C}(b)\|, \quad (3.5)$$

where $b \in \mathbb{R}^p$ is a constant vector and $C \subset \mathbb{R}^p$ is a closed convex set such that $0 \in C$ and $0 < \Upsilon(b \mid C) < +\infty$. The value function $\varphi(\lambda)$ is one example of (3.5). Besides, according to the Lipschitz continuity of the projection onto the closed convex set C and the Rademacher theorem, $\psi(\lambda)$ is differentiable almost everywhere on $(0, \Upsilon(b \mid C))$.

We first state some technical results on $\psi(\lambda)$.

Lemma 3.1. *For any $\lambda \in (0, \Upsilon(b \mid C))$, we have*

$$\begin{aligned} \Pi_{\lambda C}(b) &\neq 0, \\ b - \Pi_{\lambda C}(b) &\neq 0 \\ \text{and } \langle \Pi_{\lambda C}(b), b - \Pi_{\lambda C}(b) \rangle &> 0. \end{aligned}$$

Proof. Since $\Upsilon(b \mid C) > 0$ and $0 \in C$, we have $b \neq 0$. We will prove above three inequalities by contradiction.

If $\Pi_{\lambda C}(b) = 0$, we have

$$\begin{aligned} &\langle \Pi_{\lambda C}(b) - \lambda\Upsilon(b \mid C)^{-1}b, b - \Pi_{\lambda C}(b) \rangle \\ &= -\lambda\Upsilon(b \mid C)^{-1}\|b\|^2 < 0. \end{aligned}$$

However, according to $\lambda \Upsilon(b \mid C)^{-1}b \in \lambda C$ and the definition of projection, we have

$$\langle \Pi_{\lambda C}(b) - \lambda \Upsilon(b \mid C)^{-1}b, b - \Pi_{\lambda C}(b) \rangle \geq 0. \quad (3.6)$$

This contradiction shows the first inequality.

If $b - \Pi_{\lambda C}(b) = 0$, we have $b \in \lambda C$, which conflicts $\lambda < \Upsilon(b \mid C)$. Therefore the second inequality holds.

According to $0 \in C$ and the property of projection, we have $\langle \Pi_{\lambda C}(b), b - \Pi_{\lambda C}(b) \rangle \geq 0$. If $\langle \Pi_{\lambda C}(b), b - \Pi_{\lambda C}(b) \rangle = 0$, we have

$$\begin{aligned} & \langle \Pi_{\lambda C}(b) - \lambda \Upsilon(b \mid C)^{-1}b, b - \Pi_{\lambda C}(b) \rangle \\ &= -\lambda \Upsilon(b \mid C)^{-1} \|b - \Pi_{\lambda C}(b)\|^2 < 0, \end{aligned}$$

which conflicts the property of projection (3.6). This completes the proof. $\square$

Following theorem gives bounds for the gradient of $\psi(\lambda)$ at its differentiable points.

Theorem 3.1. *If $\lambda_0 \in (0, \Upsilon(b \mid C))$ is a differentiable point, we have*

$$\psi'(\lambda_0) \in \left[\frac{\langle \Pi_{\lambda_0 C}(b), b - \Pi_{\lambda_0 C}(b) \rangle^2}{\lambda_0 \|\Pi_{\lambda_0 C}(b)\| \|b - \Pi_{\lambda_0 C}(b)\|^2}, \frac{1}{\lambda_0} \|\Pi_{\lambda_0 C}(b)\| \right]. \quad (3.7)$$

Proof. To get the bounds of derivative $\psi'(\lambda_0)$, we construct two functions $\psi_1(\lambda)$, $\psi_2(\lambda)$ such that

$$\psi_1(\lambda) \leq \psi(\lambda) \leq \psi_2(\lambda) \text{ on } (0, \lambda_0)$$

and

$$\psi_1(\lambda_0) = \psi(\lambda_0) = \psi_2(\lambda_0)$$

to estimate the bounds of $\psi'(\lambda)$.

We denote $y_0 := \Pi_{\lambda_0 C}(b)$ for convenience. To construct ψ_1 and ψ_2 , we consider such two programs where $\lambda \in (0, \lambda_0]$ is a parameter.

$$\begin{aligned}
& \min_{z \in \mathbb{R}^p} \text{ or } \max_{z \in \mathbb{R}^p} \|z\| \\
& \text{s.t. } \exists \text{ closed convex set } 0 \in S \subseteq \mathbb{R}^p, \\
& \Pi_{\lambda_0 S}(b) = y_0, \\
& \Pi_{\lambda S}(b) = z.
\end{aligned} \tag{3.8}$$

It is obvious that $\Pi_{\lambda C}(b)$ is a feasible point for either minimization or maximization (let the closed convex set S be the set C). Optimal values of these two programs could be used as the bounds of $\psi(\lambda)$.

Program (3.8) is equivalent with the following program.

$$\begin{aligned}
& \min_{z \in \mathbb{R}^p} \text{ or } \max_{z \in \mathbb{R}^p} \|z\| \\
& \text{s.t. } \langle z - \frac{\lambda}{\lambda_0} y_0, b - y_0 \rangle \leq 0, \\
& \langle z - \frac{\lambda}{\lambda_0} y_0, z - b \rangle \leq 0, \\
& \langle z, z - b \rangle \leq 0.
\end{aligned} \tag{3.9}$$

To show the equivalence between (3.8) and (3.9), we only need to show the feasible set are same. If z is a feasible point of (3.9), it is also a feasible set of (3.8) with the closed convex

$$S := \text{conv} \left\{ 0, \frac{1}{\lambda_0} y_0, \frac{1}{\lambda} z \right\}.$$

For a feasible point z of (3.8), three inequalities in (3.9) hold because of the two projections in (3.8).

For the minimization in (3.9), we consider the following relaxation,

$$\begin{aligned} \min_{z \in \mathbb{R}^p} \quad & \|z\| \\ \text{s.t.} \quad & (\|z\| - \frac{\lambda}{\lambda_0} \|y_0\|)(\|z\| - \|y_0\|) \leq 0. \end{aligned} \quad (3.10)$$

Since $\lambda \leq \lambda_0$, $\frac{\lambda}{\lambda_0} \|y_0\|$ is the minimal value of (3.10) (actually $z_1 = \frac{\lambda}{\lambda_0} y_0$ is a minimal point of (3.9) as it is feasible for (3.9)). Therefore, we have

$$\psi_1(\lambda) := \frac{\lambda}{\lambda_0} \|y_0\| \leq \|\Pi_{\lambda C}(b)\| = \psi(\lambda) \text{ for any } \lambda \in (0, \lambda_0]. \quad (3.11)$$

For the maximization in (3.9), we consider the following relaxation,

$$\begin{aligned} \max_{z \in \mathbb{R}^p} \quad & \|z\| \\ \text{s.t.} \quad & \langle z - \frac{\lambda}{\lambda_0} y_0, b - y_0 \rangle \leq 0, \\ & \langle z - \frac{\lambda}{\lambda_0} y_0, z - y_0 \rangle \leq 0. \end{aligned} \quad (3.12)$$

The convex maximization (3.12) can be equivalently reformulated as the following program

$$\max_{\beta \in [\frac{\lambda}{\lambda_0}, 1]} \nu(\beta), \quad (3.13)$$

where

$$\begin{aligned} \nu(\beta) = \max_{z \in \mathbb{R}^p} \quad & \|z\| \\ \text{s.t.} \quad & \langle z - \frac{\lambda}{\lambda_0} y_0, b - y_0 \rangle \leq 0, \\ & \langle z - \frac{\lambda}{\lambda_0} y_0, z - \beta y_0 \rangle = 0. \end{aligned} \quad (3.14)$$

It is obvious that $\nu(\frac{\lambda}{\lambda_0}) = \frac{\lambda}{\lambda_0} \|y_0\|$. For $\beta \in (\frac{\lambda}{\lambda_0}, 1]$, we can also give an explicit solution as follows.

In the program (3.14), with the equality constraint, the objective $\|z\|$ can be transferred as $\|z - \frac{\lambda}{\lambda_0}y_0\|$ for

$$\|z - \frac{\lambda}{\lambda_0}y_0\|^2 = \frac{\beta\lambda_0 - \lambda}{\beta\lambda_0 + \lambda}\|z\|^2 + \left(\frac{\lambda^2}{\lambda_0^2} - \frac{2\beta\lambda^2}{(\beta\lambda_0 + \lambda)\lambda_0}\right)\|y_0\|^2,$$

which could be derived from the following identity

$$\|z - \frac{\lambda}{\lambda_0}y_0\|^2 = \|z - \frac{\lambda}{\lambda_0}y_0\|^2 - \frac{2\lambda}{\lambda + \beta\lambda_0}\langle z - \frac{\lambda}{\lambda_0}y_0, z - \beta y_0 \rangle.$$

We claim the maximizer of (3.14) is

$$z_2(\beta) = \beta y_0 - \frac{\beta\lambda_0 - \lambda}{\lambda_0} \frac{\langle y_0, b - y_0 \rangle}{\|b - y_0\|^2} (b - y_0).$$

The feasibility of $z_2(\beta)$ can be verified. To verify the optimality, we first combine the two constraints in (3.14) as follows

$$\begin{aligned} 0 &\geq \langle z - \frac{\lambda}{\lambda_0}y_0, z - \beta y_0 \rangle + \frac{\beta\lambda_0 - \lambda}{\lambda_0} \frac{\langle y_0, b - y_0 \rangle}{\|b - y_0\|^2} \langle z - \frac{\lambda}{\lambda_0}y_0, b - y_0 \rangle \\ &= \left\langle z - \frac{\lambda}{\lambda_0}y_0, z - \left(\frac{\lambda}{\lambda_0}y_0 + \frac{\beta\lambda_0 - \lambda}{\lambda_0}y_0\right) + \frac{\beta\lambda_0 - \lambda}{\lambda_0} \frac{\langle y_0, b - y_0 \rangle}{\|b - y_0\|^2} (b - y_0) \right\rangle \end{aligned}$$

Then the optimality is followed by

$$\begin{aligned} \left\| z - \frac{\lambda}{\lambda_0}y_0 \right\|^2 &\leq \left\langle z - \frac{\lambda}{\lambda_0}y_0, \frac{\beta\lambda_0 - \lambda}{\lambda_0}y_0 - \frac{\beta\lambda_0 - \lambda}{\lambda_0} \frac{\langle y_0, b - y_0 \rangle}{\|b - y_0\|^2} (b - y_0) \right\rangle \\ &\leq \left\| z - \frac{\lambda}{\lambda_0}y_0 \right\| \left\| \frac{\beta\lambda_0 - \lambda}{\lambda_0}y_0 - \frac{\beta\lambda_0 - \lambda}{\lambda_0} \frac{\langle y_0, b - y_0 \rangle}{\|b - y_0\|^2} (b - y_0) \right\|. \end{aligned}$$

It is obvious $\|z - \frac{\lambda}{\lambda_0}y_0\|$ takes its maximal value at $z_2(\beta)$, the corresponding maximal value in (3.14) is

$$\|z_2(\beta)\| = \sqrt{\beta^2\|y_0\|^2 + \frac{\lambda^2 - \beta^2\lambda_0^2}{\lambda_0^2} \frac{\langle y_0, b - y_0 \rangle^2}{\|b - y_0\|^2}}.$$

With

$$\|y_0\|^2 \geq \frac{\langle y_0, b - y_0 \rangle^2}{\|b - y_0\|^2},$$

we know the maximal value of (3.13) is taken at $\beta = 1$. Similarly, $z_2(1)$ is actually feasible for (3.9).

The maximal value $\|z_2(1)\|$ is the upper bound function $\psi_2(\lambda)$ we construct:

$$\begin{aligned} \psi_2(\lambda) &:= \sqrt{\|y_0\|^2 + \frac{\lambda^2 - \lambda_0^2}{\lambda_0^2} \frac{\langle y_0, b - y_0 \rangle^2}{\|b - y_0\|^2}} \\ &\geq \|\Pi_{\lambda C}(b)\| = \psi(\lambda) \text{ for any } \lambda \in (0, \lambda_0]. \end{aligned} \tag{3.15}$$

Moreover, we have $\psi_1(\lambda_0) = \psi(\lambda_0) = \psi_2(\lambda_0)$.

With the bound function $\psi_1(\lambda)$ and $\psi_2(\lambda)$, we can estimate the bounds of derivative $\psi'(\lambda_0)$:

$$\psi'(\lambda_0) = \lim_{\lambda \uparrow \lambda_0} \frac{\psi(\lambda) - \psi(\lambda_0)}{\lambda - \lambda_0} \in [\psi'_2(\lambda_0), \psi'_1(\lambda_0)].$$

The explicit form of the above interval is

$$[\psi'_2(\lambda_0), \psi'_1(\lambda_0)] = \left[\frac{\langle \Pi_{\lambda_0 C}(b), b - \Pi_{\lambda_0 C}(b) \rangle^2}{\lambda_0 \|\Pi_{\lambda_0 C}(b)\| \|b - \Pi_{\lambda_0 C}(b)\|^2}, \frac{1}{\lambda_0} \|\Pi_{\lambda_0 C}(b)\| \right].$$

The fraction in the lower bound is well defined according to Lemma 3.1. This completes the proof. $\square$

Remark 3.1. *The auxiliary program (3.9) can be easily understood with Euclidean geometry. Considering the two dimension plane spanned by y_0 and $(b - y_0)$, we can express program (3.9) in Figure 3.1. The feasible set of (3.9) is the intersection of blue circles and the left half plane of the vertical dashed line. The feasible set of (3.12) is the intersection of the green circle and the left half plane of the vertical dashed line.*

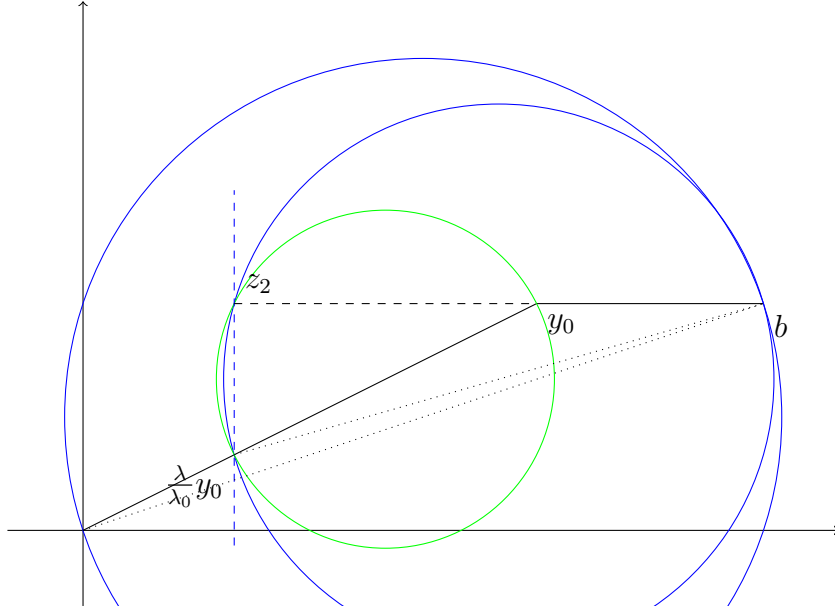


Figure 3.1: A diagram for the proof of Theorem 3.1.

Remark 3.2. *The proof of Theorem 3.1 proceeds by constructing two auxiliary functions on $(0, \lambda_0)$. This approach provides an estimate for $\psi'(\lambda_0)$, given that ψ is differentiable at λ_0 . An analogous construction for $\lambda > \lambda_0$ is possible but not required here.*

Theorem 3.2. *For any $\bar{\lambda} \in (0, \Upsilon(b \mid C))$, every element of the Clarke generalized Jacobian $\partial\psi(\bar{\lambda})$ is positive, i.e.*

$$\partial\psi(\bar{\lambda}) \subset \mathbb{R}_{++}.$$

Proof. If $\bar{\lambda}$ is a differentiable point, we have $\partial\psi(\bar{\lambda}) = \{\psi'(\bar{\lambda})\}$, whose unique element is a positive number according to Lemma 3.1 and Theorem 3.1.

For nondifferentiable $\bar{\lambda} \in (0, \Upsilon(b \mid C))$, we can find a closed neighborhood $\mathcal{B}$ of $\bar{\lambda}$ such that

$$0 < m_\psi \leq \psi'_2(\lambda) \leq \psi'(\lambda) \leq \psi'_1(\lambda) \leq M_\psi < +\infty$$

for any differentiable point $\lambda \in \mathcal{B}$. The existence of such uniform bounds m_ψ and M_ψ is due to the continuity of $\psi'_1(\lambda)$ and $\psi'_2(\lambda)$. By the definition of the Clarke generalized Jacobian, any element $v \in \partial\psi(\bar{\lambda})$ satisfies $v \in [m_\psi, M_\psi] \subset \mathbb{R}_{++}$. This completes the proof. $\square$

The nonsingularity of the Clarke generalized Jacobian of value function $\varphi(\lambda)$ follows from this theorem.

Corollary 3.1. *For any $\lambda \in (0, \|\mathcal{A}^*b\|_2)$, every element of the Clarke generalized Jacobian $\partial\varphi(\lambda)$ is positive, i.e.*

$$\partial\psi(\bar{\lambda}) \subset \mathbb{R}_{++}.$$

Proof. The value function $\varphi(\lambda)$ is an example of (3.5). And $\|\mathcal{A}^*b\|_2 = \Upsilon(b \mid Q)$. This completes the proof. $\square$

3.3 A secant method for the univariate equation (Uni.Eq.)

This section develops a secant method for the univariate equation $\varphi(\lambda) = \varrho$ and analyzes its convergence. The analysis includes studying the semismoothness of $\varphi(\lambda)$.

Algorithm 2 A secant method for $\varphi(\lambda) - \varrho = 0$

Require: Initial points $\lambda^{-1}, \lambda^0 \in \mathbb{R}$.

1: **for** $k = 0, 1, 2, \dots$ **do**

2: Compute

$$\lambda^{k+1} = \lambda^k - (\delta_\varphi(\lambda^k, \lambda^{k-1}))^{-1} (\varphi(\lambda^k) - \varrho). \quad (3.16)$$

3: **end for**

The classic secant method for univariate equation is shown in Algorithm 2. The local convergence of the secant method for the semismooth function $\varphi(\lambda)$ is established in the following proposition, adapted from (Potra et al., 1998, Theorem 3.2).

Proposition 3.3. *Suppose that $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ is semismooth at the solution λ^* such that $\varphi(\lambda^*) = \varrho$. Denote d^- and d^+ as the lateral derivatives of φ at λ^* , as defined in (2.8). Suppose that d^- and d^+ are both positive (or negative), then there is two neighborhoods $\mathcal{U}_1, \mathcal{U}_2$ of λ^* such that $\mathcal{U}_1 \subset \mathcal{U}_2$ for each $\lambda^{-1}, \lambda^0 \in \mathcal{U}_1$ the Algorithm 2 is well defined and produces a sequence of iterates $\{\lambda^k\}$ s.t. $\{\lambda^k\} \in \mathcal{U}_2$ and $\{\lambda^k\}$ converges to λ^* 3-step Q -superlinearly, i.e.*

$$|\lambda^{k+3} - \lambda^*| = o(|\lambda^k - \lambda^*|).$$

Moreover, it also holds that:

(i) if

$$\alpha := \frac{|d^+ - d^-|}{\min\{|d^+|, |d^-|\}} < 1,$$

then $\{\lambda^k\}$ is Q -linearly convergent with Q -factor α ;

(ii) if φ is γ -order semismooth at λ^* with some $\gamma > 0$, then

$$|\lambda^{k+3} - \lambda^*| = O(|\lambda^k - \lambda^*|^{1+\gamma}).$$

If φ is differentiable at λ^* with a nonzero derivative, we can establish an explicit rate of the superlinear convergence.

Proposition 3.4. *Let λ^* be a solution to $\varphi(\lambda) = \varrho$. Let $\{\lambda^k\}$ be the sequence generated by Algorithm 2 for solving $\varphi(\lambda) = \varrho$. For $k \geq -1$, denote $e_k := \lambda^k - \lambda^*$, and assume that $|e_k| > 0$. Assume that $\partial\varphi(\lambda^*)$ is a singleton $\{v^*\}$ and $v^* \neq 0$. We have following results.*

(i) *If φ is semismooth at λ^* , the sequence $\{\lambda^k\}$ converges to λ^* Q -superlinearly;*

(ii) *Suppose that φ is γ -order semismooth at λ^* with some $\gamma > 0$. For $k \geq -1$, denote $c_k := |e_k| / (|e_{k-1}| |e_{k-2}|^\gamma)$. Then either one of the following two properties is satisfied:*

(1) *$\{\lambda^k\}$ converges to λ^* superlinearly with Q -order at least $(1 + \sqrt{1 + 4\gamma})/2$;*

(2) $\{\lambda^k\}$ converges to λ^* superlinearly with R -order at least $(1 + \sqrt{1 + 4\gamma})/2$, and for any constant $\underline{C} > 0$, there exists a subsequence $\{c_{i_k}\}$ satisfying $c_{i_k} < \underline{C} i_k^{-i_k}$.

Proof. (i) The result follows from (Potra et al., 1998, Lemma 2.3) and (Li et al., 2024, Proposition 5.2).

(ii) From Lemma 2.1 and (Li et al., 2024, Proposition 5.2), we have that there exists a constant $\bar{C} > 0$ and a positive integer K such that

$$|e_{k+1}| \leq \bar{C} |e_k| |e_{k-1}|^\gamma \quad \forall k \geq K.$$

Then the convergent rates follow from (Potra, 1989, Theorem 2.2) and (Potra, 1989, Corollary 3.1). $\square$

Like Newton methods, secant methods achieve fast convergence when the Clarke generalized Jacobian is nonsingular and the function is semismooth at the optimal point. The nonsingularity condition was examined earlier; we now analyze the semismoothness of $\varphi(\lambda)$.

Proposition 3.5. *For $Q = \{z \in \mathbb{R}^p \mid \|\mathcal{A}^* z\|_2 \leq 1\}$, the projection function $\Pi_Q(\cdot)$ is γ -order semismooth and the value function $\varphi(\lambda) = \|\lambda \Pi_Q(\lambda^{-1}b)\|$ is also γ -order semismooth on $(0, +\infty)$ with some $\gamma > 0$.*

Proof. With its linear matrix inequality representation, we know set Q is a semialgebraic set. Then its projection $\Pi_Q(\cdot)$ is a semialgebraic function. From (Bolte et al., 2009, Remark 4), we know $\Pi_Q(\cdot)$ is γ -order semismooth in the sense of (2.7). The γ -order semismoothness of the value function $\varphi(\lambda)$ is due to the composition of γ -order semismooth functions (Proposition 2.4). This completes the proof. $\square$

Proposition 3.5 establishes the existence of some $\gamma > 0$. The following theorem provides sufficient conditions for the stronger result $\gamma = 1$, corresponding to strong

semismoothness of $\varphi(\lambda)$. To facilitate further discussion, we reformulate the dual of nuclear norm regularized problem (LS(λ)-D) as the following conic program:

$$\begin{aligned} \min_{y \in \mathbb{R}^p} \quad & \frac{1}{2} \|y\|^2 - b^T y \\ \text{s.t.} \quad & (\lambda, \mathcal{A}^* y) \in \text{epi } \|\cdot\|_2. \end{aligned} \quad (3.17)$$

Corresponding primal program, an equivalent formula of (LS(λ)-P) is as follow:

$$\begin{aligned} \min_{(s, X) \in \mathbb{R} \times \mathbb{R}^{m \times n}} \quad & \lambda s + \frac{1}{2} \|\mathcal{A}X - b\|^2 \\ \text{s.t.} \quad & (s, X) \in \text{epi } \|\cdot\|_*. \end{aligned} \quad (3.18)$$

The KKT conditions in the conic form is as follow:

$$\begin{aligned} y - b + \mathcal{A}X &= 0 \\ (-s, X) &\in N_{\text{epi } \|\cdot\|_2}((\lambda^*, \mathcal{A}^* y)). \end{aligned}$$

Proposition 3.6. *Let $y(\lambda^*)$ be the optimal solution of (3.17) with $\lambda = \lambda^*$. If $y(\lambda^*)$ is constraint nondegenerate to (3.17), i.e.,*

$$(0, \mathcal{A}^* \mathbb{R}^p) + \text{lin} \left(T_{\text{epi } \|\cdot\|_2}(\lambda^*, \mathcal{A}^* y(\lambda^*)) \right) = \mathbb{R} \times \mathbb{R}^{m \times n}, \quad (3.19)$$

then the value function $\varphi(\cdot)$ is strongly semismooth at λ^* .

Proof. The primal dual solution $(y(\lambda^*), s(\lambda^*), X(\lambda^*))$ of (3.17) and (3.18) with $\lambda = \lambda^*$ satisfies the nonsmooth equation

$$\begin{aligned} 0 &= G(\lambda^*, y, s, X) \\ &= F_{\lambda^*}(y, s, X) \\ &= \left[\begin{array}{c} y - b + \mathcal{A}X \\ (\lambda^*, \mathcal{A}^* y) - \Pi_{\text{epi } \|\cdot\|_2}((\lambda^*, \mathcal{A}^* y) - (s, -X)) \end{array} \right], \end{aligned} \quad (3.20)$$

which is equivalent to the KKT condition (3.1). Given the strong convexity of objective in (3.17) and the constraint nondegeneracy (3.19), we know any element of $\partial F_{\lambda^*}(y(\lambda^*), s(\lambda^*), X(\lambda^*))$ is nonsingular from Proposition 2.11.

According to (Meng et al., 2005, Theorem 3), there exist an open interval $\mathcal{N}$ containing λ^* and a Lipschitz continuous function $(y(\lambda), s(\lambda), X(\lambda))$ on $\mathcal{N}$ such that $G(\lambda, y(\lambda), s(\lambda), X(\lambda)) = 0$. Additionally, the projection $\Pi_{\text{epi}\|\cdot\|_2}(\cdot)$ is strongly semismooth from Proposition 2.8. Utilizing Proposition 2.12 and (Meng et al., 2005, Theorem 3), we conclude that $(y(\lambda), s(\lambda), X(\lambda))$ exhibits strong semismoothness at λ^* .

Besides, given $y(\lambda) = \Pi_{\lambda Q}(b) \neq 0$ and $\varphi(\lambda) = \|y(\lambda)\|$, the strong semismoothness of $\varphi(\cdot)$ at λ^* is followed from the composition of strongly semismooth functions. This completes the proof. $\square$

The convergence results established in Proposition 3.3 and Proposition 3.4 are local in nature. Consequently, achieving the corresponding convergence rates requires a sufficiently accurate initial estimate. To ensure global convergence of the secant method, we incorporate the globalization strategy presented in Li et al. (2024). This strategy, summarized in Algorithm 3 and analyzed in Proposition 3.7, hybridizes the standard secant method with a bisection safeguard. The algorithm proceeds primarily via secant iterations. However, if a candidate iterate $\hat{\lambda}^{k+1}$ fails to sufficiently reduce the residual $|\varphi(\hat{\lambda}^{k+1}) - \varrho|$ —specifically, when the condition on line 6 is violated—the method automatically switches to a bisection step (lines 12–19) to generate a more robust subsequent iterate. With the condition on line 6 and the bisection steps, Algorithm 3 is globally linearly convergent. This adaptive mechanism guarantees global convergence of the sequence to a root while preserving the favorable local convergence properties of the secant method.

Proposition 3.7. *Denote λ^* as the solution of $\varphi(\lambda) = \varrho$ (Uni.Eq.). The Algorithm*

Algorithm 3 A globally convergent secant method for $\varphi(\lambda) = \varrho$ (Uni.Eq.)

Require: Initial points and bounds such that

$$0 \leq \lambda_{\text{lo}} \leq \lambda^{-1} < \lambda^0 \leq \lambda_{\text{up}} \leq \|\mathcal{A}^*b\|_2,$$

an iteration counter $i = 0$, and a parameter $\mu \in (0, 1)$.

```

1: for  $k = 0, 1, 2 \dots$  do
2:   Compute
      
$$\hat{\lambda}^{k+1} = \lambda^k - (\delta_\varphi(\lambda^k, \lambda^{k-1}))^{-1} (\varphi(\lambda^k) - \varrho). \quad (3.21)$$

3:   if  $\hat{\lambda}^{k+1} \in (\lambda_{\text{lo}}, \lambda_{\text{up}})$  then
4:     Compute  $X(\hat{\lambda}^{k+1})$  and  $\varphi(\hat{\lambda}^{k+1})$ .
5:     Set  $i = i + 1$ .
6:     if (i)  $i \geq 3$  and  $|\varphi(\hat{\lambda}^{k+1}) - \varrho| \leq \mu |\varphi(\lambda^{k-2}) - \varrho|$  or (ii)  $i < 3$  then
7:       Set  $\lambda^{k+1} = \hat{\lambda}^{k+1}, X(\lambda^{k+1}) = X(\hat{\lambda}^{k+1})$ ;
8:     else
9:       Go to Line 12.
10:    end if
11:  else
12:    if  $\varphi(\hat{\lambda}^{k+1}) > \varrho$  then
13:       $\lambda_{\text{up}} = \min\{\lambda_{\text{up}}, \hat{\lambda}^{k+1}\}$ ;
14:    else
15:       $\lambda_{\text{lo}} = \max\{\lambda_{\text{lo}}, \hat{\lambda}^{k+1}\}$ .
16:    end if
17:    Set  $\lambda^{k+1} = \frac{1}{2}(\lambda_{\text{up}} + \lambda_{\text{lo}})$ .
18:    Compute  $X(\lambda^{k+1})$  and  $\varphi(\lambda^{k+1})$ .
19:    Set  $i = 0$ .
20:  end if
21: end for

```

$\mathcal{J}$ is well-defined and sequence $\{\lambda^k\}$ and $\{X(\lambda^k)\}$ converge to solution of equation (Uni.Eq.) and program (CP(ϱ)). For large integer k , it holds that:

(i) $\{\lambda^k\}$ converges to λ^* 3-step Q -superlinearly, i.e.

$$|\lambda^{k+3} - \lambda^*| = o(|\lambda^k - \lambda^*|);$$

if we further assume $\partial\varphi(\lambda^*) = \{v^*\}$, $\{\lambda^k\}$ converges to λ^* superlinearly, i.e.

$$|\lambda^{k+1} - \lambda^*| = o(|\lambda^k - \lambda^*|);$$

(ii) if constraint nondegeneracy (3.19) holds with λ^* and $y(\lambda^*)$, $\{\lambda^k\}$ converges to λ^* superlinearly, i.e.

$$|\lambda^{k+1} - \lambda^*| = o(|\lambda^k - \lambda^*|);$$

if we further assume $\partial\varphi(\lambda^*) = \{v^*\}$, convergence rates in Proposition 3.4 (ii) hold.

Proof. These results follow from (Li et al., 2024, Theorem 5.5), Proposition 3.3, and Proposition 3.4. $\square$

Remark 3.3. Algorithm 3 presents a globally convergent approach. In practice, we utilize the bisection method to generate an initial point for the secant method; and it is observed that the secant method converges within 3 steps for most of our testing instances. Most of these three steps convergence demonstrate $(1 + \sqrt{5})/2$ - order superlinear convergent rate. Other numerical issues will be elaborated in Section 5.1.

Remark 3.4. In this subsection we establish a sufficient condition for the strong semismoothness of $\Pi_{\lambda Q}(b)$. An intriguing avenue for future research involves exploring the strong semismoothness of $\Pi_{\lambda Q}(b)$ in scenarios where nondegeneracy (3.19) is absent or at least weakened.

Chapter 4

A proximal generation method for the nuclear norm regularized least-squares problem (LS(λ)-P)

To solve the least-squares constrained nuclear norm minimization (CP(ϱ)) by the level set method, we need to solve a sequence of nuclear norm regularized least-squares problems (LS(λ)-P) with different λ . In nuclear norm regularized problems, we introduce a proximal generation method that effectively leverages the prior solution X^{k-1} to initialize the current subproblem and harnesses the low-rank nature of solutions to expedite the iteration process.

4.1 A proximal generation method with low rank decomposition for the nuclear norm regularized least-squares problem (LS(λ)-P)

For large scale problems, the factorization form is much more efficient than the original matrix. It is known that if the nuclear norm regularized least-squares problem (LS(λ)-P) has a rank- r solution, the original convex problem

$$\min_{X \in \mathbb{R}^{m \times n}} f_\lambda(X) := \lambda \|X\|_* + \frac{1}{2} \|\mathcal{A}X - b\|^2$$

is equivalent with following LR -decomposition form

$$\min_{L \in \mathbb{R}^{m \times r}, R \in \mathbb{R}^{n \times r}} g_\lambda(L, R) := \frac{\lambda}{2}(\|L\|^2 + \|R\|^2) + \frac{1}{2}\|\mathcal{A}(LR^T) - b\|^2. \quad (4.1)$$

The equivalence between (LS(λ)-P) and (4.1) is summarized in the following proposition, which could be proved by the similar approach in (Recht et al., 2010, Lemma 5.1).

Proposition 4.1. *Suppose the original convex problem (LS(λ)-P) has a solution X^* whose rank is not greater than r , the optimal function value of nonconvex problem (4.1) is same as (LS(λ)-P). Moreover, for any L, R s.t. $X = LR^T$, we have*

$$f_\lambda(X) \leq g_\lambda(L, R). \quad (4.2)$$

If matrix X has a SVD

$$X = U_X \Sigma_X V_X^T \text{ and } L = U_X \sqrt{\Sigma_X}, R = V_X \sqrt{\Sigma_X},$$

above inequality becomes an equality.

In our proximal generation method, we measure the solution quality by a subgradient residual (SGR) introduced in Toh and Yun (2010), which is calculated through the following procedures. This residual can serve as a stopping criterion of (LS(λ)-P), as discussed in (Du, 2015, Section 4.2.1). To ensure the paper is self-contained and to clarify the concept, we analyze the validity of this residual in the following proposition.

Given a candidate matrix X^k (or its decomposition $X^k = L^k R^{kT}$), we consider these mappings:

$$\text{PG}_\lambda(X^k) := \text{prox}_{(\lambda/L_{\mathcal{A}^* \mathcal{A}})\|\cdot\|_*} \left(X^k - \frac{1}{L_{\mathcal{A}^* \mathcal{A}}} \mathcal{A}^* (\mathcal{A}X^k - b) \right), \quad (4.3a)$$

$$\text{SGR}_\lambda(X^k) := L_{\mathcal{A}^* \mathcal{A}}(X^k - \text{PG}_\lambda(X^k)) + \mathcal{A}^* (\mathcal{A}(\text{PG}_\lambda(X^k) - X^k)), \quad (4.3b)$$

$$\text{and rSGR}_\lambda(X^k) := \text{SGR}_\lambda(X^k) / (L_{\mathcal{A}^* \mathcal{A}}(1 + \|\text{PG}_\lambda(X^k)\|)), \quad (4.3c)$$

where $L_{\mathcal{A}^*\mathcal{A}}$ is the Lipschitz constant of the gradient of $\frac{1}{2}\|\mathcal{A}X - b\|^2$, the largest eigenvalue of operator $\mathcal{A}^*\mathcal{A}$. Note that the low rank decomposition $X^k = U^k V^{kT}$ can enhance the efficiency of SGR's computation.

Proposition 4.2. *Let $\{X^k\}$ be a sequence that converges to a solution of (LS(λ)-P). It holds that*

- (i) *matrix X^* satisfy $X^* = \text{PG}_\lambda(X^*)$ if and only if X^* is a solution of (LS(λ)-P);*
- (ii) *$\text{SGR}_\lambda(X^k) \in \partial f_\lambda(\text{PG}_\lambda(X^k))$;*
- (iii) *$\lim_{k \rightarrow \infty} \|\text{SGR}_\lambda(X^k)\| = 0$.*

Proof. Results (i) and (ii) follow directly from the definition of proximal mapping. We focus on the third one.

Since X^k converges to a solution of (LS(λ)-P), the continuity of the proximal mapping implies that $\text{PG}_\lambda(X^k) - X^k$ converges to zero. By the definition of $\text{SGR}_\lambda(X^k)$ and the continuity of the mapping $\mathcal{A}^*\mathcal{A}$, we conclude that $\lim_{k \rightarrow \infty} \|\text{SGR}_\lambda(X^k)\| = 0$. This completes the proof. $\square$

Remark 4.1. *Proposition 4.2 could be generalized to more general optimization problems whose objective is a smooth function plus a closed convex function. The parameter $L_{\mathcal{A}^*\mathcal{A}}$ could be generalized as a positive definite operator.*

In practice, we consider the relative subgradient residual (4.3c). We terminate the algorithm if the following stopping criterion is satisfied

$$\|\text{rSGR}_\lambda(X^k)\| \leq \varepsilon. \quad (4.4)$$

In our algorithm, we frequently generate an LR -decomposition of a matrix X using its SVD $X = U_X \Sigma_X V_X^T$. We set $L = U_X \sqrt{\Sigma_X}$ and $R = V_X \sqrt{\Sigma_X}$. For simplicity, we denote this decomposition as

$$(U_X \sqrt{\Sigma_X}, V_X \sqrt{\Sigma_X}) = \text{LR-SVD}(X).$$

Algorithm 4 A proximal generation method for (LS(λ)-P)

Require: An initial solution $X^0 \in \mathbb{R}^{m \times n}$, tolerances $\varepsilon, \varepsilon_{L,R} > 0$, step length of proximal gradient $1/L_{\mathcal{A}^* \mathcal{A}}$ where $L_{\mathcal{A}^* \mathcal{A}}$ is the largest eigenvalue of operator $\mathcal{A}^* \mathcal{A}$, and a positive integer k_{check} .

1: Compute a step of proximal gradient:

$$X_+^0 = \text{prox}_{(\lambda/L_{\mathcal{A}^* \mathcal{A}})\|\cdot\|_*} \left(X^0 - \frac{1}{L_{\mathcal{A}^* \mathcal{A}}} \mathcal{A}^* (\mathcal{A}X^0 - b) \right). \quad (4.5)$$

2: Set $(L^1, R^1) = \text{LR-SVD}(X_+^0)$.

3: Set $\text{Res} = \text{rSGR}_\lambda(X^0)$.

4: $k = 1$.

5: **while** $\text{Res} > \varepsilon$ **do**

6: Minimize $g(L, R)$ alternatively:

$$\begin{aligned} L^{k+1} &= \underset{L}{\text{argmin}} g_\lambda(L, R^k); \\ R^{k+1} &= \underset{R}{\text{argmin}} g_\lambda(L^{k+1}, R). \end{aligned} \quad (4.6)$$

7: $k \leftarrow k + 1$.

8: **if** $\text{mod}(k, k_{\text{check}}) = 0$ or $\|\nabla g_\lambda(L^k, R^k)\| \leq \varepsilon_{L,R}$ **then**

9: Set $X^k = L^k R^{kT}$.

10: Compute a step of proximal gradient:

$$X_+^k = \text{prox}_{(\lambda/L_{\mathcal{A}^* \mathcal{A}})\|\cdot\|_*} \left(X^k - \frac{1}{L_{\mathcal{A}^* \mathcal{A}}} \mathcal{A}^* (\mathcal{A}X^k - b) \right). \quad (4.7)$$

11: Set $(L^{k+1}, R^{k+1}) = \text{LR-SVD}(X_+^k)$.

12: Set $\text{Res} = \text{rSGR}_\lambda(X^k)$.

13: $k \leftarrow k + 1$.

14: **end if**

15: **end while**

16: **Return** $X^* = X_+^{k-1} = L^k R^{kT}$.

Our proximal generation method, detailed in Algorithm 4, builds upon the essence of (Yuan et al., 2023, Algorithm 1) while addressing the challenges posed by the non-polyhedral nuclear norm structure and the nonconvexity of (4.1). The method combines several key innovations: First, it employs a proximal gradient step enhanced by partial SVD to construct a novel subspace $(L, R) \in \mathbb{R}^{m \times r} \times \mathbb{R}^{n \times r}$, which enables dynamic adjustment of the dimension r while simultaneously providing high-

quality initial points for the nonconvex optimization. The initial point X^0 is strategically chosen as either the solution to the previous regularized problem (i.e., $X^0 \in \arg \min_X f_{\lambda^{k-1}}(X)$ when the current parameter is λ^k) or, when unavailable, through multiple accelerated proximal gradient (APG) iterations from an arbitrary starting point.

The method’s computational efficiency stems from its effective exploitation of the low-rank solution structure through the nonconvex formulation (4.1). This formulation is particularly well-suited for large-scale problems, as it can be efficiently solved via alternating minimization (4.6). Each alternating subproblem in this framework exhibits strong convexity and reduces to solving two nonsingular symmetric linear systems, making the approach both theoretically sound and computationally tractable. By combining these elements – the subspace construction, careful initialization, and efficient alternating minimization – our method achieves significant improvements in solving the nuclear norm regularized least squares problem (LS(λ)-P), while maintaining global convergence guarantees through the incorporation of proximal gradient steps.

Remark 4.2. *The alternating minimization (4.6) is computationally efficient, with each convex subproblem reducible to well-conditioned linear systems. Actually, we can further improve the computational efficiency through a majorization technique that replaces complex objectives with simpler surrogates, significantly reducing per-iteration costs. Detailed construction and analysis of this majorization approach will be presented subsequently. This combination of tractable alternating optimization with strategic majorization yields both theoretical guarantees and practical efficiency, particularly beneficial for large-scale implementations.*

4.2 Convergence of the proximal generation method (Algorithm 4)

This section we analyze the convergence of the proximal generation method 4. Let k_i denote the indices at which the proximal gradient step (4.7) is executed. We have

$$f_\lambda(X_+^{k_{i+1}}) \leq f_\lambda(X^{k_{i+1}}) \leq g_\lambda(L^{k_{i+1}}, R^{k_{i+1}}) \leq g_\lambda(L^{k_i+1}, R^{k_i+1}) = f_\lambda(X_+^{k_i}). \quad (4.8)$$

The first inequality is from (Beck and Teboulle, 2009, Lemma 2.3). The second inequality and the last equality is from Proposition 4.1. And the third inequality is because of the alternating minimization (4.6). It is known that proximal gradient method for (LS(λ)-P) generates a sequence whose the function values converge to the optimal value with a complexity of $O(\frac{1}{k})$ (Beck and Teboulle, 2009). We can encapsulate the convergence of our proximal generation algorithm 4 in the following proposition.

Proposition 4.3. *Let $X_+^{k_i}$ be a sequence generated by proximal gradient steps in the proximal generation method (Algorithm 4), we have*

$$f_\lambda(X^{k_i}) - f_\lambda^* \leq \frac{L_{\mathcal{A}^* \mathcal{A}} \|X^0 - X^*\|^2}{2i}, \forall X^* \in \operatorname{argmin} f_\lambda(X);$$

and any limiting point of $\{X_+^{k_i}\}$ is a solution of (LS(λ)-P).

Proof. Consider the sequence

$$X^{k_0}, X_+^{k_0}, X^{k_1}, X_+^{k_1}, \dots, X^{k_i}, X_+^{k_i}, \dots \quad (4.9)$$

We can get the $O(\frac{1}{i})$ complexity of function value is from Beck and Teboulle (2009) and the nonincreasing property (4.8).

The existence of limiting point of $\{X_+^{k_i}\}$ is from the coerciveness of $f_\lambda(\cdot)$ and the nonincreasing property (4.8). Taking a subsequence if necessary, we know the

sequence (4.9) converges to a limiting point X^* and

$$X^{k_i} - X_+^{k_i} \rightarrow 0,$$

which means that the subgradient residual also converges to zero. Then the optimality of the limiting point, $0 \in \partial f_\lambda(X^*)$, comes from the upper semicontinuity of $\partial f_\lambda(\cdot)$ ((iv) of Proposition 2.2) and (ii) of Proposition 4.2. This completes the proof. $\square$

We emphasize that the complexity bound established in the preceding proposition represents a conservative theoretical estimate. In practical implementations, the alternating minimization framework is the substantial part in the solution process through its efficient exploitation of problem structure. Empirical observations demonstrate that the proximal gradient steps (4.7) typically converge to satisfactory solutions within just 2-5 iterations. This practical efficiency significantly outperforms the conservative theoretical bounds, highlighting the algorithm's computational advantages for real-world applications.

Remark 4.3. *In our proximal generation method, we choose not to use*

$$\|X - \text{prox}_{\lambda\|\cdot\|_*}(X - \mathcal{A}^*(\mathcal{A}X - b))\|$$

to measure the quality of current solution to avoid one extra SVD. The proximal gradient step in (4.3) serves not only to calculate the SGR as in (4.3b), but also to provide a warm start for the subsequent iteration of alternating minimization if the termination condition is not met.

Remark 4.4. *One can replicate the procedure outlined in (Yuan et al., 2023, Algorithm 1) by addressing a convex problem within a subspace, such as*

$$S' = \arg \min_{S \in \mathbb{R}^{r \times r}} \{\lambda^k \|S\|_* + \frac{1}{2} \|\mathcal{A}(U' S V'^T) - b\|^2\}$$

Algorithm 5 A proximal generation based level set method with secant iterations for $(\text{CP}(\varrho))$

Require: Initial points and bounds for λ : $0 \leq \lambda_{\text{lo}} \leq \lambda^{-1} < \lambda^0 \leq \lambda_{\text{up}} \leq \|\mathcal{A}^*b\|_2$, tolerances $\varepsilon_{\text{sec}} > 0$, initial point $X^0 = U_{X^0} \Sigma_{X^0} V_{X^0}^T$ and $k = 0$.

- 1: **while** $|\varphi(\lambda^k) - \varrho| > \varepsilon_{\text{sec}}$ **do**
- 2: Generate λ^{k+1} with Algorithm 3.
- 3: Compute $X^{k+1} = U_{X^{k+1}} \Sigma_{X^{k+1}} V_{X^{k+1}}^T$ by proximal generation algorithm 4 with initial point $X^k = U_{X^k} \Sigma_{X^k} V_{X^k}^T$.
- 4: $k \leftarrow k + 1$.
- 5: **end while**
- 6: **Return** $\lambda^* = \lambda^k$ and $X(\lambda^*)$.

and then derive new U'_+ and V'_+ using

$$U'_+ \Sigma_+ V'^T_+ = \text{prox}_{(\lambda)\|\cdot\|_*} \left(U' S' V'^T - \mathcal{A}^* \left(\mathcal{A} \left(U' S' V'^T \right) - b \right) \right).$$

However, in our experiments, we have observed that the alternating minimization (4.6) is more efficient than tackling these subspace-constrained convex subproblems, especially within the framework of the level set method.

By integrating the proximal generation strategy (Algorithm 4) with the secant method (Algorithm 3), we develop the complete proximal generation-based level set method with secant iterations presented in Algorithm 5. This unified framework achieves two key advantages: (1) superlinear convergence of the outer level set iterations through the secant method's rapid root-finding capabilities, and (2) significantly reduced computational cost per iteration via the proximal generation strategy's efficient low rank decomposition problems. The synergy between these components, combining global convergence guarantees with local simplicity, yields an exceptionally efficient solver for least squares constrained nuclear norm minimization problems.

Remark 4.5. In our implementation, we choose the initial points λ^{-1} and λ^0 close to the upper bound $\|\mathcal{A}^*b\|_2$. Although smaller initial values do not affect convergence

in theory, they tend to produce regularized subproblems with higher-rank solutions, incurring unnecessary computational cost. Starting near $\|\mathcal{A}^*b\|_2$ typically yields lower-rank iterates, allowing us to incrementally generate columns of the factor matrices U and V . This heuristic substantially reduces the overall computation.

The algorithm’s superior performance stems from three principal features: first, the careful balance between outer iteration count and subproblem difficulty; second, the adaptive exploitation of low-rank structures throughout the optimization process; and third, the proximal generation based initialization/update strategy that ensures stable convergence. These combined attributes make Algorithm 5 particularly effective for large-scale applications where both solution accuracy and computational efficiency are paramount. These architectural advantages translate directly to practical benefits, as demonstrated through comprehensive numerical experiments in the following section, where we detail implementation strategies including adaptive parameter selection, optimal stopping criteria, and comparative performance benchmarks against state-of-the-art alternatives.

Chapter 5

Numerical Experiment

In this section, we showcase the numerical performance of our proximal generation based level set method for least-squares constrained nuclear norm minimization problems. We test $(\text{CP}(\varrho))$ from low rank regressions and low rank matrix completions. In our experiments, we measure the accuracy of the obtained solution X and λ by

$$\eta := \max \left\{ \frac{|\|\mathcal{A}X - b\| - \varrho|}{\max\{1, \varrho\}}, \text{rSGR}_\lambda(X) \right\}. \quad (5.1)$$

We compare our proximal generation based level set method using secant iterations (GLS_{sec}) with the proximal generation based level set method using only bisection iterations (GLS_{bis}), as well as with SPGL1¹ (van den Berg and Friedlander, 2011) and ADMM (further elaborated in the subsequent subsection), all with a tolerance of 10^{-3} . For SPGL1 which has complicated stopping conditions, we just set `options.optTol` = 10^{-3} . And we choose the `SPQR-SPG-H` mode in SPGL1, which is the faster one according to (van den Berg and Friedlander, 2011, Section 8). We didn't change the internal parameter in SPGL1. All numerical experiments were implemented using Matlab R2023a (version 9.14) on Ubuntu 22.04.3 LTS, running on an Intel Core i9-10900 CPU operating at 2.80GHz and supported by 64GB of RAM.

¹ <https://friedlander.io/publications/2010-sparse-optimization-with-least-squares/>.

5.1 Numerical issues in the proximal generation based level set method with secant iterations (Algorithm 5)

In this subsection, we delve into numerical considerations concerning Algorithm 5. Some specific configurations will be elucidated with concrete applications in subsequent subsections.

5.1.1 Numerical issues in the secant method (Algorithm 3)

The secant step in Algorithm 2 or Algorithm 3 is based on local information. To enjoy the fast convergent rate in Proposition 3.3 and 3.4, we can implement several bisection steps to generate a good initial point for the secant method. The bisection step with λ^k is formulated as follow.

$$\begin{aligned}
& \text{Compute } X(\lambda^k), \varphi(\lambda^k). \\
& \text{if } \varphi(\lambda^k) > \varrho, \lambda_{\text{up}} = \lambda^k; \\
& \text{else: } \lambda_{\text{lo}} = \lambda^k. \\
& \lambda^{k+1} = \frac{1}{2}(\lambda_{\text{lo}} + \lambda_{\text{up}}) \\
& k \leftarrow k + 1.;
\end{aligned} \tag{5.2}$$

In our implementation, we transfer to the secant method (Algorithm 3) once

$$\frac{|\varphi(\lambda^k) - \varrho|}{\max\{1, \varrho\}} \leq 0.1.$$

Another issue is the line 3 in Algorithm 5. In practice, we only compute an inexact solution $X(\lambda^k)$ for the nuclear norm regularized problem (LS(λ)-P) with $\text{rSGR}_\lambda(X^k) \leq \varepsilon^k$, where the tolerance ε^k is decreasing. Inexact solutions and their value function may deteriorate the convergent rate of section iterations. Therefore,

if $\text{rSGR}_\lambda(X^k)$ stagnates, we generate next λ by the bisection method (5.2) with previous reliable bounds. The other technique involving inexactness to reduce computational effort is early termination, especially when λ is small. If we observe

$$\|\mathcal{A}X^{k_{i+1}} - b\| < \|\mathcal{A}X^{k_i} - b\| < \varrho, \quad (5.3)$$

where k_i and k_{i+1} are two consecutive iterations executing line 5 in Algorithm 4, we terminate the process of solving this nuclear norm regularized problem and proceed to the next secant or bisection iteration.

5.1.2 Numerical issues in the proximal generation strategy (Algorithm 4)

One tricky parameter in the proximal generation algorithm 4 is k_{check} . The proximal gradient step and matrix multiplication in line 5 of Algorithm 4 are computationally intensive, particularly for large-scale problems. In our implementation, we select a relatively larger k_{check} , such as 40 for the initial several λ to avoid unnecessary checking. As λ approaches λ^* , indicated by conditions like $\frac{|\varphi(\lambda^k) - \varrho|}{\max\{1, \varrho\}} \leq 10^{-2}$ or 5×10^{-3} , we switch to a smaller k_{check} like 4 even 1, to prevent unnecessary alternative minimization as described in (4.6).

Regarding the early termination strategy, in addition to the criterion outlined in (5.3), we also terminate the proximal generation process if the residual Res in Algorithm 4 stagnates. Furthermore, we impose a maximum iteration limit to ensure the termination of the proximal generation iteration.

We also discuss some specific numerical linear algebra operations in proximal generation Algorithm 4 here. For the proximal gradient steps in (4.5) and (4.7), we only implement partial SVD with PROPACK. The number of singular values desired is set as a slightly larger integer than the previous number of singular values. For the alternating minimization (4.6), we only need to solve two positive definite linear

systems,

$$0 = \lambda L + \mathcal{A}^*(\mathcal{A}(LR^{kT}) - b)R^k, \quad (5.4a)$$

$$0 = \lambda R + (\mathcal{A}^*(\mathcal{A}(L^{k+1}R^T) - b))^T L^{k+1}, \quad (5.4b)$$

where the L and R without superscript are the variables. We employ Cholesky decomposition to solve the linear system when the positive definite mapping can be expressed as a small-dimensional matrix. This scenario occurs for (5.4b) when $\mathcal{A}(\cdot)$ represents a matrix multiplication. Otherwise we adopt the quasi minimal residual method with the diagonal preconditioner (PSQMR). Furthermore, we may solve the majorization rather than original $g_\lambda(L, R)$. This approach is applicable to matrix completion problems, where we have $\text{Id} - \mathcal{A}^*\mathcal{A}$ is a positive semidefinite mapping. Then we can replace the Hessian mappings

$$\text{Hess}_L : \mathbb{R}^{m \times r} \rightarrow \mathbb{R}^{m \times r} : \text{Hess}_L d_L = \mathcal{A}^*(\mathcal{A}(d_L R^T))R + \lambda d_L;$$

$$\text{Hess}_R : \mathbb{R}^{n \times r} \rightarrow \mathbb{R}^{n \times r} : \text{Hess}_R d_R = (\mathcal{A}^*(\mathcal{A}(L d_R^T)))^T L + \lambda d_R$$

with

$$d_L R^T R + \lambda d_L \text{ and } d_R L^T L + \lambda d_R.$$

The explicit majorizations for matrix completion problems are as follows:

$$\begin{aligned} \hat{g}(L, R^k) &:= g(L^k, R^k) + \langle \nabla_L g(L^k, R^k), L - L^k \rangle \\ &\quad + \frac{1}{2} \langle L - L^k, \lambda(L - L^k) + (L - L^k)(R^k)^T R^k \rangle; \\ \hat{g}(L^k, R) &:= g(L^k, R^k) + \langle \nabla_R g(L^k, R^k), R - R^k \rangle \\ &\quad + \frac{1}{2} \langle R - R^k, \lambda(R - R^k) + (R - R^k)(L^k)^T L^k \rangle. \end{aligned} \quad (5.5)$$

5.2 ADMM algorithms for least-squares constrained nuclear norm minimization (CP(ϱ)) and nuclear norm regularized least square problem (LS(λ)-D)

In this subsection, we describe the implementation of ADMM methods for (CP(ϱ)).

The dual problem of (CP(ϱ)) is formulated as

$$\begin{aligned} \min_{s \in \mathbb{R}^p, T \in \mathbb{R}^{m \times n}} & \delta_{\|\cdot\|_2 \leq 1}(T) + \varrho \|s\| + \langle b, s \rangle \\ \text{s.t. } & \mathcal{A}^* s + T = 0, \end{aligned} \quad (5.6)$$

whose KKT conditions are

$$\begin{aligned} \mathcal{A}^* s + T &= 0; \\ X &\in \partial \delta_{\|\cdot\|_2 \leq 1}(T); \\ \mathcal{A}X - b &\in \partial(\varrho \|s\|). \end{aligned}$$

Augmented Lagrangian function corresponding to (5.6) is

$$\begin{aligned} L^\sigma(s, T; X) &= \delta_{\|\cdot\|_2 \leq 1}(T) + \varrho \|s\| + \langle b, s \rangle - \langle X, \mathcal{A}^* s + T \rangle \\ &\quad + \frac{\sigma}{2} \|\mathcal{A}^* s + T\|^2, \quad (s, T, X) \in \mathbb{R}^p \times \mathbb{R}^{m \times n} \times \mathbb{R}^{m \times n}. \end{aligned}$$

If $\arg\min_s L^\sigma(s, T; X)$ is uncomplicated, such as when $\mathcal{A}\mathcal{A}^*$ acts as an identity mapping in matrix completion problems, we can apply ADMM directly. Given an initial point $(\bar{s}^0, \bar{T}^0, \bar{X}^0)$, we generate new iterates as

$$T^{k+1} = \arg\min_T L^\sigma(\bar{s}^k, T; \bar{X}^k); \quad (5.7a)$$

$$X^{k+1} = \bar{X}^k - \tau\sigma(\mathcal{A}^* \bar{s}^k + T^{k+1}); \quad (5.7b)$$

$$s^{k+1} = \arg\min_s L^\sigma(s, T^{k+1}; X^{k+1}); \quad (5.7c)$$

$$(\bar{s}^{k+1}, \bar{T}^{k+1}, \bar{X}^{k+1}) = (1 - \rho)(\bar{s}^k, \bar{T}^k, \bar{X}^k) + \rho(s^{k+1}, T^{k+1}, X^{k+1}). \quad (5.7d)$$

Otherwise, we consider the following equivalent form of (5.6)

$$\begin{aligned}
& \min_{r, s \in \mathbb{R}^p, T \in \mathbb{R}^{m \times n}} \delta_{\|\cdot\|_2 \leq 1}(T) + \varrho \|r\| + \langle b, s \rangle \\
& \text{s.t. } \mathcal{A}^* s + T = 0, \\
& r - s = 0,
\end{aligned} \tag{5.8}$$

whose KKT is

$$\begin{aligned}
& \mathcal{A}^* s + T = 0; \\
& r - s = 0; \\
& X \in \partial \delta_{\|\cdot\|_2 \leq 1}(T); \\
& w \in \partial (\varrho \|r\|); \\
& b - \mathcal{A}X + w = 0.
\end{aligned}$$

Then we can apply the ADMM with the augmented Lagrangian function of (5.8)

$$\begin{aligned}
L^\sigma(r, s, T; X, w) &= \delta_{\|\cdot\|_2 \leq 1}(T) + \varrho \|r\| + \langle b, s \rangle - \langle X, \mathcal{A}^* s + T \rangle - \langle w, r - s \rangle \\
&+ \frac{\sigma}{2} \|\mathcal{A}^* s + T\|^2 + \frac{\sigma}{2} \|r - s\|^2 \\
&(r, s, T, X, w) \in \mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R}^{m \times n} \times \mathbb{R}^{m \times n} \times \mathbb{R}^p.
\end{aligned}$$

Given the initial point $(\bar{r}^0, \bar{s}^0, \bar{T}^0, \bar{X}^0, \bar{w}^0)$, the ADMM iterates as

$$r^{k+1} = \underset{r}{\operatorname{argmin}} L^\sigma(r, \bar{s}^k, \bar{T}^k; \bar{X}^k, \bar{w}^k) \tag{5.9a}$$

$$T^{k+1} = \underset{T}{\operatorname{argmin}} L^\sigma(\bar{r}^k, \bar{s}^k, T; \bar{X}^k, \bar{w}^k) \tag{5.9b}$$

$$X^{k+1} = \bar{X}^k - \tau \sigma (\mathcal{A}^* \bar{s}^k + T^{k+1}) \tag{5.9c}$$

$$w^{k+1} = \bar{w}^k - \tau \sigma (r^{k+1} - \bar{s}^k) \tag{5.9d}$$

$$s^{k+1} = \underset{s}{\operatorname{argmin}} L^\sigma(r^{k+1}, s, T^{k+1}; X^{k+1}, w^{k+1}) \tag{5.9e}$$

$$\begin{aligned}
(\bar{r}^{k+1}, \bar{s}^{k+1}, \bar{T}^{k+1}, \bar{X}^{k+1}, \bar{w}^{k+1}) &= (1 - \rho)(\bar{r}^k, \bar{s}^k, \bar{T}^k, \bar{X}^k, \bar{w}^k) \\
&+ \rho(r^{k+1}, s^{k+1}, T^{k+1}, X^{k+1}, w^{k+1})
\end{aligned} \tag{5.9f}$$

To keep the stopping criteria consistent with (4.4) and (5.1), we measure the optimality of ADMM methods by

$$\frac{|\|\mathcal{A}X - b\| - \varrho|}{\max\{1, \varrho\}} \text{ and } \text{rSGR}_\lambda(X), \quad (5.10)$$

where the $\text{PG}_\lambda(X)$ in (4.3b) and (4.3c) is replaced by

$$\text{prox}_{\frac{\varrho}{\|\cdot\|_{L_{\mathcal{A}^* \mathcal{A}}}}} \left(X - \frac{1}{L_{\mathcal{A}^* \mathcal{A}}} \mathcal{A}^* (\mathcal{A}X - b) \right)$$

or

$$\text{prox}_{\frac{\varrho}{\|\cdot\|_{L_{\mathcal{A}^* \mathcal{A}}}}} \left(X - \frac{1}{L_{\mathcal{A}^* \mathcal{A}}} \mathcal{A}^* (\mathcal{A}X - b) \right)$$

respectively.

We also apply ADMM to the nuclear norm regularized problem (LS(λ)-D) with the optimal λ returned from the level set method. The augmented Lagrangian function of (LS(λ)-D) is

$$L^\sigma(y, Z; X) = -\langle b, y \rangle + \frac{1}{2} \|y\|^2 + \delta_{\|\cdot\|_2 \leq \lambda}(Z) + \langle X, \mathcal{A}^* y - Z \rangle + \frac{\sigma}{2} \|\mathcal{A}^* y - Z\|^2.$$

The specific ADMM iterations for solving (LS(λ)-D) are straightforward. Regarding stopping criteria, we employ the second residual in (5.1).

Remark 5.1. *Recent advancements have introduced accelerated ADMM-type methods with promising theoretical guarantees (Sun et al., 2024; Chen et al., 2024). However, in our empirical evaluations, these accelerated variants did not demonstrate significant performance improvements. We attribute this observation to the presence of non-polyhedral components in our problem formulations, which necessitate an expensive singular value decomposition (SVD) at each ADMM iteration, potentially offsetting the benefits of acceleration.*

5.3 Matrix regression problems

Low rank matrix regression is one important application of $(\text{CP}(\varrho))$. We collect some data mainly from bioinformatics. In these examples, linear map $\mathcal{A}$ is an explicit matrix and $\mathcal{A}X$ is matrix multiplication with a vectorization (realign the matrix's columns as a vector). As in Huang et al. (2010) and Li et al. (2018a), we expand the matrix $\mathcal{A}$ by using polynomial basis functions over their columns. Instances denoted by names ending with ' ek ', where ' e ' is a character and ' k ' is a number, are expanded using a k -order polynomial. For instance, 'yeaste3' undergoes expansion through a third-order polynomial, while 'B_C11' remains unaltered in this manner. Moreover, we also expand the rows in some instances. We summarize the statistics of our test data in Table 5.1. The breast cancer dataset referenced in Chin et al. (2006) can be accessed through Witten et al. (2009) at <https://cran.r-project.org/web/packages/PMA/index.html>. This dataset is organized into distinct instances based on chromosomal divisions. The yeast dataset from Spellman et al. (1998), as cited in Chun and Keleş (2010), is available via <https://cran.r-project.org/web/packages/spls/index.html>. Additional data from the study by Qin et al. (2021) can be found in <https://qinlab.sysu.edu.cn/GS0>. Note that the ϱ in $(\text{CP}(\varrho))$ is set by $\varrho = c\|b\|$ where coefficient c is listed in Table 5.1.

Table 5.1: Statistics of test data in low rank regressions

Prob idx	Name	Source	m	n	p	$\ b\ $	$c = \frac{\rho}{\ b\ }$
1	B_C1to5	Chin et al. (2006)	601	5930	527770	5.50+3	0.08
2	B_C4	Chin et al. (2006)	167	715	63635	1.71+3	0.08
3	B_C6to10	Chin et al. (2006)	609	4084	363476	3.86+3	0.08
4	B_C7	Chin et al. (2006)	161	923	82147	2.00+3	0.08
5	B_C11	Chin et al. (2006)	179	1122	99858	2.25+3	0.08
6	B_C11to20	Chin et al. (2006)	822	7778	692242	5.58+3	0.08
7	B_C20plus	Chin et al. (2006)	173	1247	110983	2.29+3	0.08
8	B_C7e2	Chin et al. (2006)	13203	2767	246263	3.44+3	0.08
9	B_C20pluse2	Chin et al. (2006)	15225	3739	332771	3.91+3	0.08
10	bt	Qin et al. (2021)	939	4000	980000	3.61+0	0.15
11	srf_o3	Qin et al. (2021)	2042	3199	10809421	1.02+0	0.6
12	yeaste2	Spellman et al. (1998)	5778	919	166339	1.86+2	0.08
13	yeaste3	Spellman et al. (1998)	44362	1837	332497	2.63+2	0.15
14	B_CWhole	Chin et al. (2006)	2149	19672	1750808	9.13+3	0.08
15	test1_e3	Qin et al. (2021)	4556	2991	149550	2.63+1	0.08
16	test1_e4	Qin et al. (2021)	31961	2991	149550	2.63+1	0.08
17	B_C1to5e2	Chin et al. (2006)	5711	5930	527770	5.50+3	0.08
18	B_C6to10e2	Chin et al. (2006)	19733	4084	363476	3.86+3	0.08
19	B_C11to20e2	Chin et al. (2006)	25748	7778	692242	5.58+3	0.08
20	bte2	Qin et al. (2021)	25058	4000	980000	3.61+0	0.35
21	B_C11e2	Chin et al. (2006)	16290	5606	498934	4.98+3	0.08

Table 5.2: The performance of proximal generation based level set method with secant iterations (GLS_{sec}), SPGL1(SPG), and ADMM in solving the least-squares constrained nuclear norm minimization ($\text{CP}(\varrho)$) with the $\rho = c\|b\|$ in Table 5.1

idx	rank	time(s)			residual		
		GLS_{sec}	SPG	ADMM	GLS_{sec}	SPG	ADMM
1	11	4.36+0	3.01+2	1.26+1	8.89-4	9.89-4	9.68-4
2	25	7.02-1	7.70+0	8.89-1	1.23-4	1.31-4	1.79-4
3	13	3.85+0	2.49+2	5.88+0	3.97-5	8.63-4	7.85-4
4	17	4.80-1	7.88+0	7.78-1	8.59-4	1.04-4	9.83-4
5	16	6.40-1	1.42+1	7.88-1	2.20-4	8.43-4	9.06-4
6	11	4.44+0	3.24+2	3.87+1	5.93-4	8.63-4	9.93-4
7	14	2.63-1	1.52+1	7.72-1	7.19-4	2.15-4	7.49-4
8	3	3.76+1	2.74+3	2.04+2	3.27-4	8.88-4	2.56-4
9	5	2.60+1	>10800	5.73+2	7.99-4	6.59-1	1.20-4
10	200	1.09+2	3.82+2	9.72+2	7.28-4	4.19-4	5.20-4
11	10	1.08+1	2.70+2	2.53+3	5.73-4	7.56-4	9.99-4
12	17	1.55+1	7.16+2	1.18+1	1.95-4	4.45-4	4.59-4
13	17	3.71+2	>10800	4.25+2	2.27-4	5.69-1	7.26-4
14	10	2.54+1	6.80+3	5.04+2	9.86-4	9.29-4	9.99-4
15	50	9.54+1	>10800	7.36+2	2.79-4	1.24-2	2.09-4
16	50	6.76+2	>10800	5.38+3	2.65-4	8.40-1	1.64-4
17	9	4.42+1	3.95+3	2.72+2	6.75-4	7.50-4	3.43-4
18	13	1.20+2	>10800	9.80+2	8.92-4	1.84-1	1.79-4
19	12	2.01+2	>10800	2.95+3	4.52-4	7.85-1	5.05-4
20	103	1.13+3	>10800	>10800	5.24-4	5.02-1	2.17-3
21	5	1.10+2	>10800	5.57+2	4.88-5	5.11-1	9.82-4

We conduct a comparative analysis of our proximal generation based level set method with secant iteration (GLS_{sec}) against SPGL1 and ADMM for solving least-squares constrained problems. Furthermore, we report the performance of APG and ADMM on nuclear norm regularized problems. In our ADMM implementation for solving ($\text{CP}(\varrho)$), we set $\tau = 1.618$ in (5.9d), exclude the relaxation step (5.9f), and adjust σ based on the primal-dual feasibility criterion similar to the approach in Yang et al. (2021). This tailored configuration has yielded improved performance based on our preliminary testing. The initial bounds for λ are set as $\lambda_{\text{lo}} = 0$ and $\lambda_{\text{up}} = \|\mathcal{A}^*b\|_2$, with an initial value of $\lambda^1 = 10^{-2}\|\mathcal{A}^*b\|_2$. On alternating minimization (4.6) (5.4),

we solve (5.4a) by PSQMR and (5.4b) by Cholesky decomposition. Additionally, we perform 20 iterations of the Accelerated Proximal Gradient (APG) method using λ^1 to generate an initial point, which is then used across all the algorithms we evaluate. The maximum allowed running time for each algorithm is set to 10,800 seconds.

The test results are presented in Table 5.2. The second column represents the rank of X computed by our level set method with secant iterations. The residual, denoted as η in (5.1) for the level set method and ADMM when solving $(\text{CP}(\varrho))$, is also presented. For SPGL1, the residual is characterized by the **RelError** in its log, which is actually the first term in (5.1). In cases where SPGL1 or ADMM are unable to solve $(\text{CP}(\varrho))$ within the time limit, we simply report the residual of their solution from the final iteration.

The results presented in Table 5.2 demonstrate the superiority of our proximal generation based level set method with secant iterations (GLS_{sec}). In comparison with other methods, both GLS_{sec} successfully solve all instances, whereas SPGL1 fails to solve many instances within the time limit. For these unsolved cases, the solutions returned by SPGL1 at the maximum time are significantly below the desired accuracy. For instances that SPGL1 can solve, GLS_{sec} is substantially faster, with speedups reaching nearly 300 times. Although ADMM also fails to solve all instances, it is more efficient than SPGL1 and even slightly outperforms GLS_{sec} in Problem 12. For most of problems, GLS_{sec} is more efficient than ADMM with speedups exceeding 200 times in some cases.

The comparison between the secant method and the bisection method is detailed in Table 5.3. The second and sixth columns display the solution λ of the univariate nonlinear equation (Uni.Eq.) as obtained by the secant and bisection methods, respectively. The third column lists the number of iterations for both the bisection and secant methods in GLS_{sec} , while the seventh column shows the iterations for the pure bisection method. Compared with the bisection variant (GLS_{bis}), GLS_{sec}

Table 5.3: The performance of proximal generation based level set method with secant iterations (GLS_{sec}) and with bisection iterations (GLS_{bis})

idx	GLS _{sec}				GLS _{bis}			
	λ	iter	time(s)	residual	λ	iter	time(s)	residual
1	3.216+2	3:2	4.363+0	8.893-4	3.205+2	9	8.323+0	6.605-4
2	2.062+1	3:3	7.016-1	1.226-4	2.069+1	9	1.528+0	8.242-4
3	2.379+2	3:3	3.846+0	3.972-5	2.381+2	9	5.929+0	1.571-4
4	3.158+1	3:2	4.803-1	8.587-4	3.153+1	10	1.031+0	2.944-4
5	5.164+1	3:3	6.403-1	2.198-4	5.170+1	11	1.065+0	5.106-5
6	4.729+2	2:1	4.435+0	5.925-4	4.886+2	9	1.678+1	7.470-4
7	5.980+1	3:1	2.634-1	7.193-4	6.067+1	7	6.215-1	9.378-4
8	7.569+2	7:2	3.758+1	3.274-4	7.563+2	12	5.177+1	4.301-4
9	1.087+3	2:1	2.601+1	7.993-4	1.087+3	9	6.424+1	4.096-4
10	1.472-3	4:3	1.087+2	7.276-4	1.467-3	8	1.211+2	4.527-4
11	1.256-3	2:1	1.081+1	5.726-4	1.669-3	4	2.051+1	6.659-4
12	2.895+1	3:1	1.547+1	1.950-4	2.896+1	7	2.145+1	5.221-4
13	1.220+2	7:1	3.713+2	2.268-4	1.219+2	12	4.493+2	3.469-4
14	1.283+3	2:1	2.543+1	9.858-4	1.383+3	9	1.230+2	5.632-4
15	6.092-2	7:2	9.543+1	2.792-4	6.097-2	15	1.588+2	6.340-4
16	6.097-2	7:2	6.760+2	2.653-4	6.097-2	15	1.081+3	2.362-4
17	5.048+2	2:3	4.415+1	6.746-4	4.324+2	3	1.907+1	2.333-5
18	5.409+2	2:3	1.196+2	8.917-4	5.382+2	8	2.151+2	5.412-5
19	9.876+2	2:2	2.011+2	4.521-4	9.938+2	7	4.395+2	9.809-4
20	7.133-3	11:4	1.126+3	5.242-4	7.154-3	17	1.362+3	2.840-4
21	1.229+3	3:2	1.103+2	4.883-5	1.225+3	8	1.380+2	7.543-4

requires fewer iterations and achieves a speedup of approximately 2 to 5 times, highlighting the effectiveness of our secant approach.

Besides, we report the performance of APG and ADMM on nuclear norm regularized problems with the optimal λ returned from the level set method with secant iterations. The APG code we adopt is from Toh and Yun (2010), <https://blog.nus.edu.sg/mattohkc/software/npls/>. For ADMM we adopt the same setting with the ADMM for the least squares constrained problem (CP(ϱ)). The results are demonstrated in Table 5.4. Note that we also list other rest results for clarity. For many instances, our proximal generation based level set method is often faster than solving just one regularized problem.

Table 5.4: The performance of proximal generation based level set method with secant iterations (GLS_s) and bisection iterations (GLS_b), SPGL1(SPG), and ADMM(ADM) in solving the least square constrained nuclear norm minimization (CP(ρ)) with the $\rho = c\|b\|$ in Table 5.1; and the performance of ADMM (ADM _{λ^*}) and APG (APG _{λ^*}) in solving the nuclear norm regularized least square problem (LS(λ)-P) with the λ returned by GLS_{sec}.

	idx	rank	$\lambda_{\text{sec}} \mid \lambda_{\text{bis}}$		(bis:sec) bis	time(s)				residual					
						GLS _s GLS _b SPG ADM	APG _{λ^*} ADM _{λ^*}	GLS _s GLS _b SPG ADM	APG _{λ^*} ADM _{λ^*}						
∞	1	11	3.216+2	3.205+2	(3:2) 9	4.36+0 8.32+0 3.01+2 1.26+1	4.28+0 4.25+0	8.89-4 6.61-4 9.89-4 9.68-4	9.62-4 2.42-4						
	2	25	2.062+1	2.069+1	(3:3) 9	7.02-1 1.53+0 7.70+0 8.89-1	2.81+0 1.21-1	1.23-4 8.24-4 1.31-4 1.79-4	5.71-4 5.80-4						
	3	13	2.379+2	2.381+2	(3:3) 9	3.85+0 5.93+0 2.49+2 5.88+0	3.43+0 2.19+0	3.97-5 1.57-4 8.63-4 7.85-4	8.26-4 2.16-4						
	4	17	3.158+1	3.153+1	(3:2) 10	4.80-1 1.03+0 7.88+0 7.78-1	2.10+0 1.32-1	8.59-4 2.94-4 1.04-4 9.83-4	6.49-4 7.07-4						
	5	16	5.164+1	5.170+1	(3:3) 11	6.40-1 1.07+0 1.42+1 7.88-1	1.88+0 3.04-1	2.20-4 5.11-5 8.43-4 9.06-4	5.84-4 2.08-4						
	6	11	4.729+2	4.886+2	(2:1) 9	4.44+0 1.68+1 3.24+2 3.87+1	7.75+0 9.18+0	5.93-4 7.47-4 8.63-4 9.93-4	9.90-4 2.80-4						
	7	14	5.980+1	6.067+1	(3:1) 7	2.63-1 6.22-1 1.52+1 7.72-1	1.87+0 3.56-1	7.19-4 9.38-4 2.15-4 7.49-4	4.49-4 1.47-4						
	8	3	7.569+2	7.563+2	(7:2) 12	3.76+1 5.18+1 2.74+3 2.04+2	1.86+1 3.50+1	3.27-4 4.30-4 8.88-4 2.56-4	9.85-4 6.45-4						
	9	5	1.087+3	1.087+3	(2:1) 9	2.60+1 6.42+1 >10800 5.73+2	6.97+1 6.77+1	7.99-4 4.10-4 6.59-1 1.20-4	2.98-4 9.32-4						
	10	200	1.472-3	1.467-3	(4:3) 8	1.09+2 1.21+2 3.82+2 9.72+2	1.39+2 3.15+1	7.28-4 4.53-4 4.19-4 5.20-4	9.65-4 9.06-4						
	11	10	1.256-3	1.669-3	(2:1) 4	1.08+1 2.05+1 2.70+2 2.53+3	3.74+1 4.27+1	5.73-4 6.66-4 7.56-4 9.99-4	9.34-4 9.80-4						
	12	17	2.895+1	2.896+1	(3:1) 7	1.55+1 2.15+1 7.16+2 1.18+1	1.02+1 1.08+1	1.95-4 5.22-4 4.45-4 4.59-4	7.38-4 2.88-5						
	13	17	1.220+2	1.219+2	(7:1) 12	3.71+2 4.49+2 >10800 4.25+2	1.27+2 2.10+2	2.27-4 3.47-4 5.69-1 7.26-4	8.89-4 2.40-5						
	14	10	1.283+3	1.383+3	(2:1) 9	2.54+1 1.23+2 6.80+3 5.04+2	3.93+1 7.74+1	9.86-4 5.63-4 9.29-4 9.99-4	9.89-4 3.66-4						
	15	50	6.092-2	6.097-2	(7:2) 15	9.54+1 1.59+2 >10800 7.36+2	1.87+2 3.74+1	2.79-4 6.34-4 1.24-2 2.09-4	2.88-4 3.31-5						
	16	50	6.097-2	6.097-2	(7:2) 15	6.76+2 1.08+3 >10800 5.38+3	1.12+3 2.66+2	2.65-4 2.36-4 8.40-1 1.64-4	2.87-4 3.30-5						
	17	9	5.048+2	4.324+2	(2:3) 3	4.42+1 1.91+1 3.95+3 2.72+2	2.83+1 7.73+1	6.75-4 2.33-5 7.50-4 3.43-4	9.69-4 6.22-4						
	18	13	5.409+2	5.382+2	(2:3) 8	1.20+2 2.15+2 >10800 9.80+2	8.14+1 1.68+2	8.92-4 5.41-5 1.84-1 1.79-4	9.45-4 2.47-4						
	19	12	9.876+2	9.938+2	(2:2) 7	2.01+2 4.40+2 >10800 2.95+3	2.18+2 6.70+2	4.52-4 9.81-4 7.85-1 5.05-4	7.42-4 3.10-4						
	20	103	7.133-3	7.154-3	(11:4) 17	1.13+3 1.36+3 >10800 >10800	7.28+2 1.55+2	5.24-4 2.84-4 5.02-1 2.17-3	9.89-4 6.16-5						
	21	5	1.229+3	1.225+3	(3:2) 8	1.10+2 1.38+2 >10800 5.57+2	6.37+1 1.23+2	4.88-5 7.54-4 5.11-1 9.82-4	9.69-4 2.59-4						

5.4 Matrix completion problems

We also test our method on the classic matrix completion problems. We use the MovieLens datasets (Harper and Konstan, 2015) from <https://grouplens.org/datasets/movielens/> and the Jester joke datasets (Goldberg et al., 2001) from <https://eigentaste.berkeley.edu/dataset/>. We extend the original data in Jester datasets by add some comments by averaging some random users' values with some perturbation. We also generate some synthetic data with the same way in Toh and Yun (2010). We summarize the statistics of these data in Table 5.5. The ϱ in $(CP(\varrho))$ is set by $\varrho = c\|b\|$ where coefficient c is in the last column of 5.5.

Table 5.5: Statistics of test data in low rank matrix completion

Prob idx	Name	Source	m	n	p	$\ b\ $	$c = \frac{\rho}{\ b\ }$
1	u100K	Harper and Konstan (2015)	943	1682	49918	8.27+2	0.2
2	uleastsmall	Harper and Konstan (2015)	610	9724	53246	8.39+2	0.25
3	u1M	Harper and Konstan (2015)	6040	3706	498742	2.65+3	0.25
4	u20m	Harper and Konstan (2015)	7120	14026	525002	2.67+3	0.25
5	u10M	Harper and Konstan (2015)	71567	10677	4983232	8.19+3	0.2
6	u25m	Harper and Konstan (2015)	162541	59047	12472784	1.30+4	0.25
7	ulatest	Harper and Konstan (2015)	330975	83239	17198680	1.54+4	0.25
8	jester1ex	Goldberg et al. (2001)	24983	10100	249830	3.35+3	0.3
9	jester2ex	Goldberg et al. (2001)	23500	10100	235000	3.23+3	0.3
10	jester3ex	Goldberg et al. (2001)	24938	9926	265164	4.27+3	0.3
11	jester4ex	Goldberg et al. (2001)	54905	6653	553353	5.91+3	0.3
12	jester5ex	Goldberg et al. (2001)	7699	9011	88920	2.33+3	0.2
13	Rand1	Toh and Yun (2010)	7000	8000	2992002	1.22+4	0.2
14	Rand2	Toh and Yun (2010)	7000	10000	3385373	1.30+4	0.2
15	Rand3	Toh and Yun (2010)	7500	12000	4280292	1.54+4	0.2
16	Rand4	Toh and Yun (2010)	7500	10000	4182480	1.59+4	0.2
17	Rand5	Toh and Yun (2010)	9000	14000	3444147	1.31+4	0.2
18	Rand6	Toh and Yun (2010)	10000	12000	3947772	1.54+4	0.25
19	Rand7	Toh and Yun (2010)	10000	12500	4037941	1.56+4	0.25
20	Rand8	Toh and Yun (2010)	20000	15000	5245442	1.62+4	0.25
21	Rand9	Toh and Yun (2010)	30000	31000	10972303	2.57+4	0.25
22	Rand10	Toh and Yun (2010)	50000	8000	10433300	2.50+4	0.25

Most settings of algorithms for matrix completion are same as the previous subsection on matrix regression. We only show some changes. For our proximal generation based level set methods GLS_{sec} and GLS_{bis} , we set the initial value as $\lambda^1 = 0.1 \times \|\mathcal{A}^*b\|_2$ for (semi)real instances and $\lambda^1 = 0.3 \times \|\mathcal{A}^*b\|_2$ for synthetic instances. On alternating minimization (4.6), we adopt the majorization (5.5). Accordingly we use a larger k_{check} for line 4 in Algorithm 4. On ADMM for the least-squares constrained problem $(\text{CP}(\varrho))$, we set $\tau = 1$ in (5.7b), set $\rho = 1.8$ in (5.7d), and adjust σ by the primal-dual feasibility criterion. This configuration performs better in our preliminary test.

The results of matrix completion problems are presented in Table 5.6. Our GLS_{sec} method continues to demonstrate its superiority in this context. When evaluating SPGL1 and ADMM for solving $(\text{CP}(\varrho))$, numerous instances exceed the time limit (indicated as '>10800' in the table) or are terminated due to insufficient memory (marked as 'oom' in the table), despite our desktop having 64GB of RAM. For instances that SPGL1 can solve, GLS_{sec} is 10 to 100 times faster. Although ADMM outperforms GLS_{sec} on Problem 1, the smallest instance in our test, GLS_{sec} is 15 to 40 times faster for other instances that ADMM can solve.

The comparison between the secant method and the bisection method is detailed in Table 5.7. The secant method GLS_{sec} is approximately twice as fast as the bisection variant GLS_{bis} for most instances, and it requires significantly fewer iterations.

We also test the APG and ADMM on the nuclear norm regularized problems with λ from the secant method. The setting of APG and ADMM is same as the previous section. The results are demonstrated in Table 5.8. Compared to APG and ADMM for single regularized problems, GLS_{sec} applied to a sequence of regularized problems is faster than ADMM for a single regularized problem. While APG appears efficient for some synthetic instances, it just meets the 10^{-3} tolerance level, indicating that APG is not a suitable option for level set subproblems, even for these instances.

Table 5.6: The performance of proximal generation based level set method with secant iterations (GLS_{sec}), SPGL1(SPG), and ADMM in solving the least-squares constrained nuclear norm minimization ($\text{CP}(\varrho)$) with the $\rho = c\|b\|$ in Table 5.5.

idx	rank	time(s)			residual		
		$\text{GLS}_{\text{sec}} : \text{SPG} : \text{ADMM}$			$\text{GLS}_{\text{sec}} : \text{SPG} : \text{ADMM}$		
1	60	6.88+0	5.19+1	4.45+0	2.33-4	9.86-4	9.54-4
2	43	1.05+1	2.02+2	1.63+2	5.54-4	4.34-4	9.61-4
3	32	4.70+1	4.02+3	7.16+2	2.68-4	5.30-4	2.10-4
4	91	2.96+2	>10800	8.17+3	4.04-4	5.85-2	5.61-4
5	303	6.14+3	oom	oom	4.16-5	nan	nan
6	139	5.63+3	oom	oom	8.47-4	nan	nan
7	215	1.18+4	oom	oom	2.03-4	nan	nan
8	150	5.15+2	>10800	>10800	8.15-4	5.14-1	2.76+0
9	153	4.37+2	>10800	>10800	1.88-4	5.14-1	2.21+4
10	131	3.90+2	>10800	>10800	8.05-4	5.54-1	7.25-2
11	65	7.53+2	oom	>10800	7.10-4	nan	2.28+4
12	84	7.99+1	>10800	2.94+3	8.10-4	5.13-1	6.53-4
13	50	2.26+2	3.96+3	6.08+3	4.68-4	4.02-4	5.89-4
14	50	2.59+2	7.16+3	6.78+3	1.43-4	8.60-4	7.91-4
15	55	4.11+2	9.90+3	9.48+3	5.38-4	3.63-1	4.09-4
16	60	3.63+2	3.41+3	8.34+3	7.02-4	9.99-4	3.16-4
17	50	4.08+2	>10800	>10800	6.24-4	3.16-2	1.71-3
18	60	5.38+2	>10800	>10800	9.76-4	1.56-1	9.50-3
19	60	5.57+2	>10800	>10800	9.73-4	9.57-2	1.15-2
20	50	1.07+3	>10800	>10800	6.69-4	6.71-1	1.13+5
21	60	5.00+3	oom	oom	6.10-6	nan	nan
22	60	2.64+3	>10800	>10800	7.43-5	2.68-1	1.56+5

Table 5.7: The performance of proximal generation based level set method with secant iterations (GLS_{sec}) and with bisection iterations (GLS_{bis})

idx	GLS_{sec}				GLS_{bis}			
	λ	iter	time(s)	residual	λ	iter	time(s)	residual
1	1.016+1	5:1	6.88+0	2.33-4	1.016+1	9	1.21+1	2.28-4
2	1.314+1	2:3	1.05+1	5.54-4	1.315+1	10	2.29+1	7.06-5
3	3.015+1	3:3	4.70+1	2.68-4	3.014+1	10	9.54+1	9.30-4
4	2.306+1	9:5	2.96+2	4.04-4	2.299+1	17	2.85+2	6.56-4
5	3.857+1	7:3	6.14+3	4.16-5	3.855+1	9	6.43+3	6.77-4
6	7.624+1	5:1	5.63+3	8.47-4	7.693+1	9	1.17+4	7.53-5
7	7.741+1	4:2	1.18+4	2.03-4	7.860+1	7	2.26+4	8.73-4
8	1.539+1	7:0	5.15+2	8.15-4	1.552+1	11	7.09+2	3.12-4
9	1.496+1	6:0	4.37+2	1.88-4	1.496+1	6	4.32+2	1.88-4
10	2.259+1	3:1	3.90+2	8.05-4	2.383+1	20	1.05+3	2.77-2
11	3.632+1	7:5	7.53+2	7.10-4	3.673+1	20	1.36+3	6.57-4
12	1.389+1	5:2	7.99+1	8.10-4	1.389+1	7	8.54+1	4.86-4
13	7.252+1	2:1	2.26+2	4.68-4	7.257+1	10	4.50+2	5.47-4
14	7.346+1	2:1	2.59+2	1.43-4	7.344+1	10	5.42+2	2.21-4
15	8.196+1	2:1	4.11+2	5.38-4	8.194+1	10	8.98+2	8.13-4
16	8.775+1	2:2	3.63+2	7.02-4	8.778+1	12	8.02+2	6.78-5
17	5.455+1	2:3	4.08+2	6.24-4	5.440+1	13	8.59+2	6.48-4
18	8.246+1	4:2	5.38+2	9.76-4	8.229+1	8	7.08+2	1.33-4
19	8.268+1	4:2	5.57+2	9.73-4	8.283+1	9	7.94+2	2.88-4
20	6.902+1	4:1	1.07+3	6.69-4	6.939+1	8	1.59+3	1.81-5
21	8.239+1	6:2	5.00+3	6.10-6	8.235+1	10	6.23+3	3.50-4
22	1.194+2	4:1	2.64+3	7.43-5	1.194+2	10	4.59+3	8.15-4

Table 5.8: The performance of proximal generation based level set method with secant iterations (GLS_s) and bisection iterations (GLS_b), SPGL1(SPG), and ADMM(ADM) in solving the least square constrained nuclear norm minimization (CP(ρ)) with the $\rho = c\|b\|$ in Table 5.5; and the performance of ADMM (ADM _{λ^*}) and APG (APG _{λ^*}) in solving the nuclear norm regularized least square problem (LS(λ)-P) with the λ returned by GLS_{sec}.

	idx	rank	$\lambda_{\text{sec}} \mid \lambda_{\text{bis}}$		(bis:sec) bis	time(s)				residual							
						GLS _s	GLS _b	SPG	ADM	APG $_{\lambda^*}$	ADM $_{\lambda^*}$	GLS _s	GLS _b	SPG	ADM	APG $_{\lambda^*}$	ADM $_{\lambda^*}$
68	1	60	1.016+1	1.016+1	(5:1) 9	6.88+0	1.21+1	5.19+1	4.45+0	5.74+0	4.58+0	2.33-4	2.28-4	9.86-4	9.54-4	9.58-4	9.50-4
	2	43	1.314+1	1.315+1	(2:3) 10	1.05+1	2.29+1	2.02+2	1.63+2	1.06+1	8.25+1	5.54-4	7.06-5	4.34-4	9.61-4	9.98-4	5.67-4
	3	32	3.015+1	3.014+1	(3:3) 10	4.70+1	9.54+1	4.02+3	7.16+2	1.61+1	2.58+2	2.68-4	9.30-4	5.30-4	2.10-4	9.57-4	8.85-4
	4	91	2.306+1	2.299+1	(9:5) 17	2.96+2	2.85+2	>10800	8.17+3	7.83+1	2.46+3	4.04-4	6.56-4	5.85-2	5.61-4	9.90-4	6.18-4
	5	303	3.857+1	3.855+1	(7:3) 9	6.14+3	6.43+3	oom	oom	8.09+2	oom	4.16-5	6.77-4	nan	nan	9.78-4	nan
	6	139	7.624+1	7.693+1	(5:1) 9	5.63+3	1.17+4	oom	oom	1.87+3	oom	8.47-4	7.53-5	nan	nan	9.81-4	nan
	7	215	7.741+1	7.860+1	(4:2) 7	1.18+4	2.26+4	oom	oom	4.72+3	oom	2.03-4	8.73-4	nan	nan	9.95-4	nan
	8	150	1.539+1	1.552+1	(6:1) 11	5.15+2	7.09+2	>10800	>10800	5.05+2	7.37+3	8.15-4	3.12-4	5.14-1	2.76+0	9.88-4	9.67-4
	9	153	1.496+1	1.496+1	(6:0) 6	4.37+2	4.32+2	>10800	>10800	4.56+2	6.88+3	1.88-4	1.88-4	5.14-1	2.21+4	9.92-4	9.59-4
	10	131	2.259+1	2.383+1	(3:1) 20	3.90+2	1.05+3	>10800	>10800	4.09+2	1.10+4	8.05-4	2.77-2	5.54-1	7.25-2	9.94-4	9.38-4
	11	65	3.632+1	3.673+1	(7:5) 20	7.53+2	1.36+3	oom	>10800	4.29+2	>10800	7.10-4	6.57-4	nan	2.28+4	9.97-4	1.52-3
	12	84	1.389+1	1.389+1	(5:2) 7	7.99+1	8.54+1	>10800	2.94+3	8.91+1	1.42+3	8.10-4	4.86-4	5.13-1	6.53-4	9.94-4	9.87-4
	13	50	7.252+1	7.257+1	(2:1) 10	2.26+2	4.50+2	3.96+3	6.08+3	3.01+1	7.65+2	4.68-4	5.47-4	4.02-4	5.89-4	8.56-4	9.84-4
	14	50	7.346+1	7.344+1	(2:1) 10	2.59+2	5.42+2	7.16+3	6.78+3	3.36+1	1.18+3	1.43-4	2.21-4	8.60-4	7.91-4	7.83-4	2.32-4
	15	55	8.196+1	8.194+1	(2:1) 10	4.11+2	8.98+2	9.90+3	9.48+3	4.67+1	1.56+3	5.38-4	8.13-4	3.63-1	4.09-4	9.94-4	2.58-4
	16	60	8.775+1	8.778+1	(2:2) 12	3.63+2	8.02+2	3.41+3	8.34+3	4.99+1	1.32+3	7.02-4	6.78-5	9.99-4	3.16-4	9.22-4	1.68-4
	17	50	5.455+1	5.440+1	(2:3) 13	4.08+2	8.59+2	>10800	>10800	2.07+2	2.95+3	6.24-4	6.48-4	3.16-2	1.71-3	9.78-4	8.88-4
	18	60	8.246+1	8.229+1	(4:2) 8	5.38+2	7.08+2	>10800	>10800	7.36+1	2.43+3	9.76-4	1.33-4	1.56-1	9.50-3	8.39-4	3.01-4
	19	60	8.268+1	8.283+1	(4:2) 9	5.57+2	7.94+2	>10800	>10800	6.13+1	2.47+3	9.73-4	2.88-4	9.57-2	1.15-2	9.04-4	3.15-4
	20	50	6.902+1	6.939+1	(4:1) 8	1.07+3	1.59+3	>10800	>10800	6.00+1	9.55+3	6.69-4	1.81-5	6.71-1	1.13+5	8.13-4	7.79-4
	21	60	8.239+1	8.235+1	(6:2) 10	5.00+3	6.23+3	oom	oom	1.28+3	oom	6.10-6	3.50-4	nan	nan	9.51-4	nan
	22	60	1.194+2	1.194+2	(4:1) 10	2.64+3	4.59+3	>10800	>10800	2.98+2	7.46+3	7.43-5	8.15-4	2.68-1	1.56+5	9.75-4	3.92-4

We also compare our method with LoRADS (Han et al., 2024), a state-of-the-art SDP solver for matrix completion problems. Although LoRADS is designed for linear equality constraints, we use the optimal solution from GLS_{sec} to construct the equality constrained version to enable comparison. Following their large-scale matrix completion setup, we use initial rank $\log m$, heuristic factor 2.5, phase 2 tolerance 10^{-3} , and test four phase 1 tolerances (10^{-3} , 5×10^{-3} , 10^{-2} and 10^{-1}) with a 3600s time limit. For fair comparison, we include APG initialization time in GLS_{sec}'s total runtime since LoRADS cannot leverage this warm-start. Note that LoRADS deals with the SDP form, whose primal problem is $\min_{X \in \mathbb{S}_+} \langle C, X \rangle$ s.t. $\mathcal{A}X = b$ and dual problem is $\max_{\lambda} b^T \lambda$ s.t. $C - \mathcal{A}^* \lambda \in \mathbb{S}_+$. For evaluation, we compute the following primal feasibility, dual feasibility and relative gap:

$$\begin{aligned} & \frac{\|\mathcal{A}(X) - b\|_2}{1 + \|b\|_2}, \\ & \frac{|\min \{0, \sigma_{\min}(C - \mathcal{A}^*(\lambda))\}|}{1 + \|C\|_2}, \\ & \frac{\langle C, X \rangle - \lambda^T b}{1 + |\langle C, X \rangle| + |\lambda^T b|}. \end{aligned}$$

Table 5.9 presents the best LoRADS results across different settings. The numerical results show that on our test problems, LoRADS encounters several challenges: (1) it cannot solve some larger-scale instances, (2) requires substantially more computation time for solvable problems, and (3) obtains solutions with relatively higher residuals within the time limit. Our method demonstrates consistent performance across all problem scales, achieving both computational efficiency and superior solution quality.

Table 5.9: The performance of proximal generation based level set method with secant iterations (GLS_{sec}) and LoRADS

idx	GLS _{sec}		LoRADS		
	time	residual	time	gap	prim : dual
3	4.97+1	2.68-4	3.73+3	2.75-3	2.13-3 : NA
5	6.21+3	4.16-5		out of memory	
4	2.99+2	4.04-4	3.93+3	1.43-1	3.39-2 : NA
6	5.83+3	8.47-4		out of memory	
7	1.21+4	2.03-4		out of memory	
8	5.23+2	8.15-4	2.81+3	4.49-5	3.10-4 : 6.58-4
9	4.46+2	1.88-4	2.56+3	1.92-4	3.70-4 : 6.31-4
10	3.96+2	8.05-4	3.22+3	2.56-4	2.75-4 : 2.76-3
11	7.61+2	7.10-4		out of memory	
12	8.16+1	8.10-4	9.74+2	3.20-5	2.00-4 : 8.31-3
13	2.87+2	4.68-4	3.87+3	9.90-1	1.44+0 : NA
15	5.02+2	5.38-4	4.30+3	9.60-1	1.29+0 : NA
17	4.67+2	6.24-4	4.11+3	9.60-1	1.29+0 : NA
19	6.21+2	9.73-4	4.79+3	9.71-1	1.45+0 : NA
20	1.21+3	6.69-4	4.94+3	9.42-1	1.71+0 : NA
21	5.33+3	6.10-6		out of memory	
22	3.00+3	7.43-5		out of memory	

5.5 Numerical performance of secant iterations

The efficiency of secant iterations is demonstrated through comparison with the bisection variant. Table 5.10 presents some numerical results from several contiguous secant iterations, where

$$\eta_{\text{LS}}(X^k) = \frac{||\mathcal{A}X^k - b|| - \varrho}{\max\{1, \varrho\}}.$$

The results for 'B.C6to10' and 'u10m' exhibit superlinear convergence with a Q-order of $(1 + \sqrt{5})/2$, as stated in Proposition 3.4. Based on the data in the third and fifth columns of Table 5.10, we assert that our secant iterations achieve 3-step quadratic convergence. Note that high precision is not required for the regularized subproblems. We anticipate that the residual in the last column could be reduced to less than 10^{-4} with additional alternating minimizations in the corresponding

Table 5.10: Numerical performance of secant iterations

B_C4	iter(k)	4	5	6
	$\eta_{LS}(X^k)$	1.886-2	1.121-3	1.226-4
B_C6to10	iter(k)	4	6	6
	$\eta_{LS}(X^k)$	1.042-2	1.522-3	3.972-5
bt	iter(k)	5	6	7
	$\eta_{LS}(X^k)$	2.385-2	1.699-2	7.276-4
B_C1to5e2	iter(k)	3	4	5
	$\eta_{LS}(X^k)$	1.966-2	6.077-3	6.746-4
B_C6to10e2	iter(k)	3	4	5
	$\eta_{LS}(X^k)$	2.648-2	3.416-3	8.917-4
uleastsmall	iter(k)	3	4	5
	$\eta_{LS}(X^k)$	2.545-2	7.833-3	5.536-4
u10m	iter(k)	8	9	10
	$\eta_{LS}(X^k)$	1.865-2	3.367-3	4.159-5
u1m	iter(k)	4	5	6
	$\eta_{LS}(X^k)$	1.701-2	2.148-3	2.678-4

subproblem.

Chapter 6

Conclusions and future work

This chapter draws conclusions on the thesis, and points out some possible research directions related to the work done in this thesis.

In this thesis, we designed a proximal generation based level set method with secant iterations for the least square constrained nuclear norm minimization (CP(ϱ)). The level set method with secant iterations is the outer iteration. We investigate some properties related with the convergence of the secant method. For any closed convex set C with $0 < \lambda < \Upsilon(b \mid C) < +\infty$, we established the nonsingularity of the Clarke generalized Jacobian of the projection norm function $\psi(\lambda) = \|\Pi_{\lambda C}(b)\|$, specifically $\partial\psi(\lambda) \subset \mathbb{R}_{++}$. The value function $\varphi(\lambda)$ in level set method serves as an example of $\psi(\lambda)$. This nonsingularity enabled us to achieve superlinear convergence of the secant method. Under the condition of dual constraint nondegeneracy (3.19), the secant method can attain a convergence rate of $(1 + \sqrt{5})/2$. For the subproblems within the level set method, specifically the nuclear norm regularized problem (LS(λ)-P), we developed a proximal generation strategy that significantly reduces computational cost while preserve the global convergence. Our proximal generation-based level set method with secant iterations demonstrated efficient numerical performance compared to other state-of-the-art algorithms in extensive numerical experiments.

For future research, we plan to investigate the strong semismoothness of the value

function $\varphi(\lambda)$ or the projection $y(\lambda) = \Pi_{\lambda Q}(b)$ without relying on nondegeneracy. On general low rank matrix optimization problems, we aim to develop novel algorithms for decomposed matrix optimizations that are more general than (4.1) and customize algorithms for related problems in Section 1.1.4.

Bibliography

- Abdolali, M. and Gillis, N. (2021), “Beyond linear subspace clustering: A comparative study of nonlinear manifold clustering algorithms,” *Computer Science Review*, 42, 100435.
- Aravkin, A., Kumar, R., Mansour, H., Recht, B., and Herrmann, F. J. (2014), “Fast methods for denoising matrix completion formulations, with applications to robust seismic data interpolation,” *SIAM Journal on Scientific Computing*, 36, S237–S266.
- Aravkin, A. Y., Burke, J. V., Drusvyatskiy, D., Friedlander, M. P., and Roy, S. (2019), “Level-set methods for convex optimization,” *Mathematical Programming*, 174, 359–390.
- Beck, A. and Teboulle, M. (2009), “A fast iterative shrinkage-thresholding algorithm for linear inverse problems,” *SIAM journal on imaging sciences*, 2, 183–202.
- Bi, S., Pan, S., and Sun, D. (2020), “A multi-stage convex relaxation approach to noisy structured low-rank matrix recovery,” *Mathematical Programming Computation*, 12, 569–602.
- Bolte, J., Daniilidis, A., and Lewis, A. (2009), “Tame functions are semismooth,” *Mathematical Programming*, 117, 5–19.
- Bonnans, J. F. and Shapiro, A. (2013), *Perturbation analysis of optimization problems*, Springer Science & Business Media.
- Burer, S. and Monteiro, R. D. (2003), “A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization,” *Mathematical programming*, 95, 329–357.
- Burer, S. and Monteiro, R. D. (2005), “Local minima and convergence in low-rank semidefinite programming,” *Mathematical programming*, 103, 427–444.
- Cai, J.-F., Candès, E. J., and Shen, Z. (2010), “A singular value thresholding algorithm for matrix completion,” *SIAM Journal on optimization*, 20, 1956–1982.

- Chen, J. and Yang, J. (2013), “Robust subspace segmentation via low-rank representation,” *IEEE transactions on cybernetics*, 44, 1432–1445.
- Chen, K., Sun, D., Yuan, Y., Zhang, G., and Zhao, X. (2024), “HPR-LP: An implementation of an HPR method for solving linear programming,” *arXiv preprint arXiv:2408.12179*.
- Chen, L., Chen, R., Sun, D., and Zhang, L. (2025), “Characterizations of the Aubin property of the solution mapping for nonlinear semidefinite programming: L. Chen et al.” *Mathematical Programming*, pp. 1–32.
- Chen, Y., Yi, X., and Caramanis, C. (2014), “A convex formulation for mixed regression with two components: Minimax optimal rates,” in *Conference on Learning Theory*, pp. 560–604, PMLR.
- Chen, Y., Chi, Y., and Goldsmith, A. J. (2015), “Exact and stable covariance estimation from quadratic sampling via convex programming,” *IEEE Transactions on Information Theory*, 61, 4034–4059.
- Chen, Y., Yi, X., and Caramanis, C. (2017), “Convex and nonconvex formulations for mixed regression with two components: Minimax optimal rates,” *IEEE Transactions on Information Theory*, 64, 1738–1766.
- Chen, Z. and Wang, S. (2022), “A review on matrix completion for recommender systems,” *Knowledge and Information Systems*, 64, 1–34.
- Chin, K., DeVries, S., Fridlyand, J., Spellman, P. T., Roydasgupta, R., Kuo, W.-L., Lapuk, A., Neve, R. M., Qian, Z., Ryder, T., et al. (2006), “Genomic and transcriptional aberrations linked to breast cancer pathophysiologies,” *Cancer cell*, 10, 529–541.
- Chun, H. and Keleş, S. (2010), “Sparse partial least squares regression for simultaneous dimension reduction and variable selection,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72, 3–25.
- Clarke, F. (1976), “On the inverse function theorem,” *Pacific Journal of Mathematics*, 64, 97–102.
- Clarke, F. H. (1990), *Optimization and nonsmooth analysis*, SIAM.
- Cui, Y. and Sun, D. (2018), “A complete characterization of the robust isolated calmness of nuclear norm regularized convex optimization problems,” *Journal of Computational Mathematics*, pp. 441–458.
- Cui, Y., Hoheisel, T., NGHIA, T. T., and SUN, D. (2024), “Lipschitz stability of least-squares problems regularized by functions with C2-cone reducible conjugates,” *arXiv preprint arXiv:2409.13118*.

- Ding, C. (2012), “An introduction to a class of matrix optimization problems,” *Singapore: National University of Singapore*.
- Ding, C. (2017), “Variational analysis of the Ky Fan k -norm,” *Set-Valued and Variational Analysis*, 25, 265–296.
- Ding, C., Sun, D., and Ye, J. J. (2014a), “First order optimality conditions for mathematical programs with semidefinite cone complementarity constraints,” *Mathematical Programming*, 147, 539–579.
- Ding, C., Sun, D., and Toh, K.-C. (2014b), “An introduction to a class of matrix cone programming,” *Mathematical Programming*, 144, 141–179.
- Ding, C., Sun, D., and Zhang, L. (2017), “Characterization of the robust isolated calmness for a class of conic programming problems,” *SIAM Journal on Optimization*, 27, 67–90.
- Dontchev, A. L. and Rockafellar, R. T. (1996), “Characterizations of strong regularity for variational inequalities over polyhedral convex sets,” *SIAM Journal on Optimization*, 6, 1087–1105.
- Drusvyatskiy, D., Krislock, N., Voronin, Y.-L., and Wolkowicz, H. (2017), “Noisy Euclidean distance realization: robust facial reduction and the Pareto frontier,” *SIAM Journal on Optimization*, 27, 2301–2331.
- Du, M. (2015), “AN INEXACT ALTERNATING DIRECTION METHOD OF MULTIPLIERS FOR CONVEX COMPOSITE CONIC PROGRAMMING WITH NONLINEAR CONSTRAINTS,” Ph.D. thesis, PhD’s thesis, Department of Mathematics, National University of Singapore.
- Facchinei, F. and Pang, J.-S. (2003), “Finite-dimensional variational inequalities and complementarity problems,” *New York: Springer-Verlag*.
- Fan, J., Gong, W., and Zhu, Z. (2019), “Generalized high-dimensional trace regression via nuclear norm regularization,” *Journal of econometrics*, 212, 177–202.
- Fazel, M., Hindi, H., and Boyd, S. P. (2001), “A rank minimization heuristic with application to minimum order system approximation,” in *Proceedings of the 2001 American control conference.(Cat. No. 01CH37148)*, vol. 6, pp. 4734–4739, IEEE.
- Fazel, M., Pong, T. K., Sun, D., and Tseng, P. (2013), “Hankel matrix rank minimization with applications to system identification and realization,” *SIAM Journal on Matrix Analysis and Applications*, 34, 946–977.
- Fischer, A. (1997), “Solution of monotone complementarity problems with locally Lipschitzian functions,” *Mathematical Programming*, 76, 513–532.

- Ghaoui, L. E., Viallon, V., and Rabbani, T. (2010), “Safe feature elimination for the lasso and sparse supervised learning problems,” *arXiv preprint arXiv:1009.4219*.
- Goldberg, K., Roeder, T., Gupta, D., and Perkins, C. (2001), “Eigentaste: A constant time collaborative filtering algorithm,” *information retrieval*, 4, 133–151.
- Guo, S.-Y., Zhang, L.-W., and Hao, S.-L. (2015), “On Some Aspects of Perturbation Analysis for Matrix Cone Optimization Induced by Spectral Norm,” *Journal of the Operations Research Society of China*, 3, 275–296.
- Han, Q., Li, C., Lin, Z., Chen, C., Deng, Q., Ge, D., Liu, H., and Ye, Y. (2024), “A Low-Rank ADMM Splitting Approach for Semidefinite Programming,” *arXiv preprint arXiv:2403.09133*.
- Harper, F. M. and Konstan, J. A. (2015), “The movielens datasets: History and context,” *Acm transactions on interactive intelligent systems (tiis)*, 5, 1–19.
- Hsieh, C.-J. and Olsen, P. (2014), “Nuclear norm minimization via active subspace selection,” in *International Conference on Machine Learning*, pp. 575–583, PMLR.
- Huang, L., Jia, J., Yu, B., Chun, B.-G., Maniatis, P., and Naik, M. (2010), “Predicting execution time of computer programs using sparse polynomial regression,” *Advances in neural information processing systems*, 23.
- Ji, S. and Ye, J. (2009), “An accelerated gradient method for trace norm minimization,” in *Proceedings of the 26th annual international conference on machine learning*, pp. 457–464.
- Jiang, K., Sun, D., and Toh, K.-C. (2013), “Solving nuclear norm regularized and semidefinite matrix least squares problems with linear equality constraints,” *Discrete Geometry and Optimization*, pp. 133–162.
- Jiang, K., Sun, D., and Toh, K.-C. (2014), “A partial proximal point algorithm for nuclear norm regularized matrix least squares problems,” *Mathematical Programming Computation*, 6, 281–325.
- Kong, D., An, B., Zhang, J., and Zhu, H. (2020), “L2RM: Low-rank linear regression models for high-dimensional matrix responses,” *Journal of the American Statistical Association*.
- Kummer, B. (1992), “Newton’s method based on generalized derivatives for nonsmooth functions: convergence analysis,” in *Advances in Optimization: Proceedings of the 6th French-German Colloquium on Optimization Held at Lambrecht, FRG, June 2–8, 1991*, pp. 171–194, Springer.

- Lee, C.-p., Liang, L., Tang, T., and Toh, K.-C. (2022), “Accelerating nuclear-norm regularized low-rank matrix optimization through Burer-Monteiro decomposition,” *arXiv preprint arXiv:2204.14067*.
- Li, Q., Jiang, B., and Sun, D. (2023), “MARS: A second-order reduction algorithm for high-dimensional sparse precision matrices estimation,” *Journal of Machine Learning Research*, 24, 1–44.
- Li, Q., Sun, D., and Yuan, Y. (2024), “An Efficient Sieving-Based Secant Method for Sparse Optimization Problems with Least-Squares Constraints,” *SIAM Journal on Optimization*, 34, 2038–2066.
- Li, X., Sun, D., and Toh, K.-C. (2018a), “A highly efficient semismooth Newton augmented Lagrangian method for solving Lasso problems,” *SIAM Journal on Optimization*, 28, 433–458.
- Li, X., Sun, D., and Toh, K.-C. (2018b), “On efficiently solving the subproblems of a level-set method for fused lasso problems,” *SIAM Journal on Optimization*, 28, 1842–1866.
- Liu, G., Lin, Z., and Yu, Y. (2010), “Robust subspace segmentation by low-rank representation,” in *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 663–670.
- Liu, J., Chen, Y., Zhang, J., and Xu, Z. (2014), “Enhancing low-rank subspace clustering by manifold regularization,” *IEEE Transactions on Image Processing*, 23, 4022–4030.
- Liu, Y.-J., Sun, D., and Toh, K.-C. (2012), “An implementable proximal point algorithmic framework for nuclear norm minimization,” *Mathematical programming*, 133, 399–436.
- Luo, L., Xie, Y., Zhang, Z., and Li, W.-J. (2015), “Support matrix machines,” in *International conference on machine learning*, pp. 938–947, PMLR.
- Ma, S., Goldfarb, D., and Chen, L. (2011), “Fixed point and Bregman iterative methods for matrix rank minimization,” *Mathematical Programming*, 128, 321–353.
- Meng, F., Sun, D., and Zhao, G. (2005), “Semismoothness of solutions to generalized equations and the Moreau-Yosida regularization,” *Mathematical programming*, 104, 561–581.
- Miao, W. (2013), “Matrix completion models with fixed basis coefficients and rank regularized problems with hard constraints,” Ph.D. thesis, National University of Singapore, Dissertation for Ph. D. Degree. Republic of Singapore.

- Mifflin, R. (1977), “Semismooth and semiconvex functions in constrained optimization,” *SIAM Journal on Control and Optimization*, 15, 959–972.
- Monteiro, R. D., Sujanani, A., and Cifuentes, D. (2024), “A low-rank augmented Lagrangian method for large-scale semidefinite programming based on a hybrid convex-nonconvex approach,” *arXiv preprint arXiv:2401.12490*.
- Mordukhovich, B. S., Nghia, T. T., and Rockafellar, R. T. (2015), “Full stability in finite-dimensional optimization,” *Mathematics of Operations Research*, 40, 226–252.
- Pang, J.-S., Sun, D., and Sun, J. (2003), “Semismooth homeomorphisms and strong stability of semidefinite and Lorentz complementarity problems,” *Mathematics of Operations Research*, 28, 39–63.
- Potra, F. (1989), “On Q-order and R-order of convergence,” *Journal of Optimization Theory and Applications*, 63, 415–431.
- Potra, F. A., Qi, L., and Sun, D. (1998), “Secant methods for semismooth equations,” *Numerische Mathematik*, 80, 305–324.
- Qin, J., Hu, Y., Yao, J.-C., Leung, R. W. T., Zhou, Y., Qin, Y., and Wang, J. (2021), “Cell fate conversion prediction by group sparse optimization method utilizing single-cell and bulk OMICs data,” *Briefings in bioinformatics*, 22, bbab311.
- Qu, W., Xiu, X., Chen, H., and Kong, L. (2023), “A survey on high-dimensional subspace clustering,” *Mathematics*, 11, 436.
- Recht, B., Fazel, M., and Parrilo, P. A. (2010), “Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization,” *SIAM review*, 52, 471–501.
- Robinson, S. M. (1980), “Strongly regular generalized equations,” *Mathematics of Operations Research*, 5, 43–62.
- Rockafellar, R. T. (1970), *Convex Analysis*, Princeton University Press.
- Rockafellar, R. T. and Wets, R. J.-B. (2009), *Variational analysis*, vol. 317, Springer Science & Business Media.
- Shapiro, A. (1990), “On concepts of directional differentiability,” *Journal of optimization theory and applications*, 66, 477–487.
- Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998), “Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization,” *Molecular biology of the cell*, 9, 3273–3297.

- Sun, D. (2006), “The strong second-order sufficient condition and constraint nondegeneracy in nonlinear semidefinite programming and their implications,” *Mathematics of Operations Research*, 31, 761–776.
- Sun, D. and Sun, J. (2002), “Semismooth matrix-valued functions,” *Mathematics of Operations Research*, 27, 150–169.
- Sun, D., Yuan, Y., Zhang, G., and Zhao, X. (2024), “Accelerating preconditioned ADMM via degenerate proximal point mappings,” *arXiv preprint arXiv:2403.18618*.
- Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J., and Tibshirani, R. J. (2012), “Strong rules for discarding predictors in lasso-type problems,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 74, 245–266.
- Toh, K.-C. and Yun, S. (2010), “An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems,” *Pacific Journal of optimization*, 6, 15.
- van den Berg, E. and Friedlander, M. P. (2009), “Probing the Pareto frontier for basis pursuit solutions,” *Siam journal on scientific computing*, 31, 890–912.
- van den Berg, E. and Friedlander, M. P. (2011), “Sparse optimization with least-squares constraints,” *SIAM Journal on Optimization*, 21, 1201–1229.
- Wang, S., Ding, C., Zhang, Y., and Zhao, X. (2023), “Strong variational sufficiency for nonlinear semidefinite programming and its implications,” *SIAM Journal on Optimization*, 33, 2988–3011.
- Witten, D. M., Tibshirani, R., and Hastie, T. (2009), “A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis,” *Biostatistics*, 10, 515–534.
- Wu, B. (2014), “High-dimensional analysis on matrix decomposition with application to correlation matrix estimation in factor models,” Ph.D. thesis, National University of Singapore, Dissertation for Ph. D. Degree. Republic of Singapore.
- Wu, C., Cui, Y., Li, D., and Sun, D. (2023), “Convex and nonconvex risk-based linear regression at scale,” *INFORMS Journal on Computing*, 35, 797–816.
- Yang, L., Li, J., Sun, D., and Toh, K.-C. (2021), “A fast globally linearly convergent algorithm for the computation of Wasserstein barycenters,” *Journal of Machine Learning Research*, 22, 1–37.

- Yuan, Y., Chang, T.-H., Sun, D., and Toh, K.-C. (2022), “A dimension reduction technique for large-scale structured sparse optimization problems with application to convex clustering,” *SIAM Journal on Optimization*, 32, 2294–2318.
- Yuan, Y., Lin, M., Sun, D., and Toh, K.-C. (2023), “Adaptive sieving: a dimension reduction technique for sparse optimization problems,” *arXiv preprint arXiv:2306.17369*.
- Zeng, Y., Yang, T., and Breheny, P. (2021), “Hybrid safe–strong rules for efficient optimization in lasso-type problems,” *Computational statistics & data analysis*, 153, 107063.
- Zhang, L., Guo, S., Wu, J., and Hao, S. (2013), “Nonsingularity in matrix conic optimization induced by spectral norm via a smoothing metric projector,” *Mathematical Methods of Operations Research*, 78, 373–404.
- Zhong, P. and Xu, Y. (2023), “Subspace screening rule for multi-label estimator with sparsity-inducing regularization,” *Neurocomputing*, 527, 143–154.
- Zhou, Q. and Zhao, Q. (2015), “Safe subspace screening for nuclear norm regularized least squares problems,” in *International Conference on Machine Learning*, pp. 1103–1112, PMLR.