A Computational and Statistical Study of Convex and Nonconvex Optimization with Applications to Structured Source Demixing and Matrix Factorization Problems

A DISSERTATION
SUBMITTED TO THE FACULTY
OF THE UNIVERSITY OF MINNESOTA
BY

Mojtaba Kadkhodaie Elyaderani

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Professor Jarvis Haupt, Advisor

September, 2017

Acknowledgements

First and foremost, I would like to sincerely thank my doctoral advisor Prof. Jarvis Haupt for his continuous encouragement and guidance during this research. Despite his busy schedule, Jarvis has always been extremely open to discussions. He has given me the freedom and support to carve out and pursue my favorite research problems under his mentorship. His caring character has always persuaded me to seek his advice on challenges that I have encountered during the years of working with him. His academic competency, along with his optimism, kindness, and modesty has turned him into my favorite professor at the graduate school.

I should also thank Prof. Nikos Sidiropoulos, Prof. Mingyi Hong, and Prof. Stefano Gonella for agreeing to serve on my committee. I am honored to know Prof. Sidiropoulos, since my early years in graduate school. I have learned a lot from his amazing lectures, novel ideas, and exceptionally kind personality. Also, I am very thankful to Prof. Hong, who was a great mentor and friend of mine during my first years at the University of Minnesota. His openness to academic discussions has encouraged me to frequently bring my questions to him. Finally, over the past three years, I have had the privilege to be part of an interdisciplinary collaboration with Prof. Gonella. His knowledge and approachable character made this collaboration very fruitful for me.

I would also like to thank my professors who put in their heart and soul in teaching and sincerely sharing their knowledge and expertise. I would especially like to thank Prof. Zhi-Quan (Tom) Luo, Prof. Arindam Banerjee, Prof. Jarvis Haupt, Prof. Georgios B. Giannakis, and Prof. Ravi Janardan, whose teachings and guidance have helped me develop a good understanding of the basics required for research in various topics in Electrical and Computer Engineering and for their exclusive feedback on my projects, research works, and presentations. In particular, I am very thankful to my master's

degree advisor, Prof. Zhi-Quan (Tom) Luo. His lectures and supervision offered me a unique opportunity to learn the fundamentals required for conducting research in optimization and signal processing.

I have also had the honor of collaborating with the members of the Artificial Intelligence lab at 3M Company as an intern. In particular, I should thank the lab manager Dr. Robert D. Lorentz for his endless support and my mentor, Dr. Guruprasad Somasundaram, for providing me with continuous guidance during the internship.

I express my sincere gratitude to my peers Swayambhoo Jain, Akshay Soni, Alexander Gutierrez, Sirisha Rambhatla, Xingguo Li, Abhinav V. Sambasivan, Di Xiao, and Scott Sievert for their patience and support. The material in this thesis has greatly benefited from frequent discussions with them. I would also like to thank the former members of the OSPAC group: Mingyi Hong, Meisam Razaviyayn, Maziar Sanjabi, Ruoyu Sun, Andy Tseng, and Wei-Cheng Liao. I feel honored to have worked with them, who are now experts in their respective domains.

Living as a graduate student has never been without challenges. I was fortunate to have great friends during that time who helped me weather all kind of obstacles. I would like to thank Meisam Razaviyayn, Maziar Sanjabi, Morteza Mardani, Ali Ghoreyshi, Ehsan Ebrahimzadeh, Ali Ashtari, Davood Taherinia, Mehdi Behroozi, Amin Tadayon, Mohsen Mahmoodi, Mahdi Ahmadi, Mamad Nasiri, Mehdi Lamee, Armin Zare, Maral Mousavi, Pardees Azodanloo, Fatemeh Sheikholeslami, Sara Parhizgari, Behnaz and Farnaz Forootaninia, Mehran Elyasi, Saeed Hashemi, Hassan Najafi, Arash Mahnoon, Amir Ghasemi, Milad Shaddelan, Siavash Ghavami, Alireza Sadeghi, Navid Reyhanian, Swayambhoo Jain, Alexander Gutierrez, Sirisha Rambhatla, Xingguo Li, Di Xiao, Abhinav Sambhasivan, John Spaulding, Philip Friesen, and finally my wonderful piano teacher Donna Swanson.

Last but not least, I would like to thank my family for always being by my side. A special feeling of gratitude to my loving parents Ali and Fakhri, who always encouraged me and provided me the required support to pursue my dreams and aspirations, and to my lovely siblings, Mohsen and Mahsa, who are the most precious gifts in my life.

Mojtaba K. Elyaderani, Minneapolis, MN, August 2017.

Dedication

To my parents ...

Abstract

Modern machine learning problems that emerge from real-world applications typically involve estimating high dimensional model parameters, whose number may be of the same order as or even significantly larger than the number of measurements. In such high dimensional settings, statistically-consistent estimation of true underlying models via classical approaches is often impossible, due to the lack of identifiability. A recent solution to this issue is through incorporating regularization functions into estimation procedures to promote intrinsic low-complexity structure of the underlying models. Statistical studies have established successful recovery of model parameters via structure-exploiting regularized estimators and computational efforts have examined efficient numerical procedures to accurately solve the associated optimization problems.

In this dissertation, we study the statistical and computational aspects of some regularized estimators that are successful in reconstructing high dimensional models. The investigated estimation frameworks are motivated by their applications in different areas of engineering, such as structural health monitoring and recommendation systems. In particular, the group Lasso recovery guarantees provided in Chapter 2 will bring insight into the application of this estimator for localizing material defects in the context of a structural diagnostics problem.

Chapter 3 describes the convergence study of an accelerated variant of the well-known alternating direction method of multipliers (ADMM) for minimizing strongly convex functions. The analysis is followed by several experimental evidence into the algorithm's applicability to a ranking problem.

Finally, Chapter 4 presents a local convergence analysis of regularized factorization-based estimators for reconstructing low-rank matrices. Interestingly, the analysis of this chapter reveals the interplay between statistical and computational aspects of such (non-convex) estimators. Therefore, it can be useful in a wide variety of problems that involve low-rank matrix estimation.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Machine learning problems are typically studied from two fundamental aspects. One is the statistical aspect, which is concerned with finding estimators to accurately recover the true underlying model of the acquired data. The other is the computational aspect, which analyses numerical procedures for computing statistically-consistent estimators.

From the statistical perspective, a major challenge in modern high dimensional problems is the so called *curse of dimensionality*, which refers to the condition where the number of unknown parameters (to be estimated) exceeds the number of acquired data points. A relatively recent and widely studied solution to this problem is via incorporating extra knowledge about the structure of the true model into the estimation procedure. In fact, by leveraging an appropriate *regularization* function that promotes inherent *low-complexity* structure of the true underlying model, it has been shown that successful recovery of the true model is possible in high dimensional settings [1].

On the other hand, thinking of the computational aspects of modern machine learning problems reveals that any numerical algorithm for a recent real world problem has to be able to handle massive amounts of data. This is partly due to the emergence of automated data acquisition systems, which have provided us with datasets of huge scales. Therefore, scalable numerical procedures that are able to provide accurate solutions, without demanding significant computational and memory resources are most favorable. As a result, an important body of recent work in machine learning and optimization, studies algorithms that only exploit *first-order* knowledge of the optimization problems to compute statistically-consistent estimators.

Efforts have been made recently to jointly investigate the statistical and computational aspects of learning problems [2]. Such efforts are in contrast with classical works, which were generally studying the two aspects in isolation. By merging the two, the recent studies are able to develop stronger analyses and avoid certain classical assumptions (examples to appear) that are typically invalid in modern applications. From the statistical point of view, such modern studies consider limitations of high dimensionality. From the computational perspective, it has been shown that classical notions (such as strong convexity and smoothness, that were classically crucial to ensure fast rates of convergence) can be relaxed and replaced by their less-stringent counterparts, which can be verified via a statistical examination of the ground-truth model.

The analyses that are presented in the chapters of this thesis exhibit the flavors of the above described studies. As will be discussed shortly, most chapters investigate the statistical and/or numerical aspects of certain inference approaches, which are inspired by recent real-world applications.

1.1 Estimation of Group-Sparse Models in Structural Diagnostics

As mentioned earlier, to mitigate the effect of the curse of dimensionality on statistical learning, many recent studies leverage regularization as a means to promote the underlying low-complexity structure of the true model. Widely-studied examples of such structures are sparsity [3–5], group sparsity [6–8], and low-rank models [9, 10]. An important class of estimators (also called M-estimators) for such low-complexity parameters involve minimizing properly-regularized loss functions, where the loss function is formed such that its minimization ensures the consistency of the learned model with acquired data. Examples of those M-estimators are the Lasso [11], group Lasso [6], and nuclear norm minimization methods [9], which are well-understood estimators of sparse, group sparse, and low rank models, respectively.

The flavors of statistical guarantees for these estimators are generally different and to some extent dependent on the application of interest. For instance, different measures of proximity, between the true model and the one inferred, might be exploited when providing estimation guarantees. One metric, that is particularly useful for model

selection in the context of sparse (or group sparse) inference, measures the mismatch between the *supports* (or group supports) of the estimated and true models, i.e. it checks whether the two models happen to be non-zero on the same entries (or groups) [12].

Chapter 2 of the thesis is focused on studying the conditions for exact group support recovery of signals (another term for model parameters) which exhibit group sparse structures via the group Lasso estimator. The use of this metric to evaluate the performance of group Lasso is motivated by its application in a problem that arises in the field of structural health monitoring, where exactly recovering the location of material defects amounts to correctly estimating the group support of the associated model parameter.

When modeling the structural health monitoring problem of this chapter, we seek to *demix* (or decompose) the acquired measurements, of an acoustic wavefield propagating through a physical medium, into two inherently-different components. One component is responsible for the smooth behavior of the healthy bulk of the medium, while the other one models the locally-sparse response of defected regions. The inclusion of the structural knowledge about the two components into the estimation procedure is crucial, since the resulting demixing problem involves more number of unknowns than measurements and is in general ill-posed. The underlying structure that is exploited here is the group sparsity of the components under appropriate transformations. The final M-estimator then takes the form of minimizing a loss function, which ensures the fit of the model to acquired measurements, and group-sparsity promoting regularizers that encourage the low-complexity structure of the two components.

1.2 An Accelerated Alternating Direction Method of Multipliers

The problem of demixing two intrinsically-different components of wavefield measurements that arises in Chapter 2 is just an instance of modern estimation problems that emerge out of studying real-world applications. In general, *superposition* models, where two or more intrinsically-different components are employed when modeling measurements appear in many machine learning problems [13]. To overcome the lack of identifiability of such models, structural assumptions about the model components are made and appropriate regularization functions are utilized to form statistical estimators.

However, computing such estimators often requires utilizing numerical algorithms to solve associated optimization problems. Efficient algorithms are those that are capable of handling such complex superposition models and are at the same time scalable and accurate. An important cexample of such algorithms is the so-called alternating direction method of multipliers (ADMM), which is useful to minimize objective functions that are composed of multiple terms depending on different subsets (also called "blocks") of variables, while the blocks of variables are coupled via linear constraints. In the context of superposition models, different terms of the objective function are the structure-imposing regularization functions, each of which depends on a different subsets of model parameters and the problem constraint is included to ensure the model's consistency with measurements.

Chapter 3 is a study on the iteration complexity of an accelerated variant of ADMM. The acceleration is important here since, while being a powerful framework for solving optimization problems, ADMM is still a first-order numerical procedure and can be slow when seeking high-accuracy solutions to large-scale problems. The acceleration method that is utilized in this chapter is inspired by the classical work of Nesterov [14] and is proved to yield an improvement in terms of convergence rates.

A main drawback of the study in Chapter 3 is the dependence of its underlying analysis on strong structural assumptions, which are fairly common in some classical works of optimization. In particular, the analysis relies on the strong convexity of the functions forming the objective. Since this assumption is not realized in many practical scenarios, studies are carried out that attempt to avoid the assumption when analyzing the convergence of ADMM. Two examples of such studies are [15, 16]. We would like to note that, in addition to those studies, the analysis of Chapter 4 of this thesis circumvents this issue by using milder variants of this assumption (although in the context of analyzing different algorithms).

1.3 Estimation of Low-Rank Models via Regularized Factorization

We mentioned low-rank matrices as an example of low-complexity models, whose statistical inference can be facilitated via incorporating this knowledge of their structure.

Over the past decade, studies have been conducted on convex M-estimators, which make use of the so-called nuclear norm to imply the underlying low-rank structure of the true model. However, such (typically convex) estimators suffer from having high computational and memory requirements, especially in recent high dimensional applications.

An alternative line of work has been developed, which involves representing the low-rank matrix as the product of two matrices (also called factors), which possess smaller dimensions and their common dimension is indicated by the rank of the true underlying matrix. Even though such a representation usually leads to non-convex optimization problems, recent studies advocate its use through their experiments and are even able to provide guarantees for their successful job in reconstructing the true model [17, 18].

The material of this chapter follows up on this thread of works and proposes a general regularized variant of factorization-based estimators. By jointly studying the computational and statistical aspects of the proposed non-convex inference approach, we are able to develop convergence results that reveal the interplay between the statistical and computational aspects of the estimator. In our study we will rely on modified versions of the assumptions that classical studies are built upon, namely universal strong convexity and smoothness assumptions will be replaced by their restricted counterparts.

1.4 Results

The topics investigated in this dissertation have resulted in several publications. In particular, the study of the support recovery guarantees for group Lasso and its application in structural health monitoring has resulted in a number of conference publications [19–22], one publication in the structural health monitoring journal [23], and an under-review submission to the IEEE Transactions on Signal Processing [24]. The convergence study of the accelerated alternating direction method of multipliers has resulted in one conference publication [25] in the proceedings of the 21st international conference on knowledge discovery and data mining (KDD). Part of the results of Chapter 4 on the estimation of low-rank matrices via regularized factorization are submitted for revision to the conference on neural information processing systems.

Chapter 2

Estimation of Group Sparse Signals for Structural Diagnostics

2.1 Introduction

In recent years, the recovery of structured signals from noisy linear measurements has been an active area of research in the fields of signal processing, high-dimensional statistics, and machine learning [1, 26–28].

Suppose an unknown signal $\boldsymbol{\beta}^* \in \mathbb{R}^p$ is observed via the noisy linear measurements

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{w}, \tag{2.1}$$

where $\boldsymbol{y} \in \mathbb{R}^n$ is the vector of observations, $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is the dictionary matrix, and $\boldsymbol{w} \in \mathbb{R}^n$ describes noise and/or model inaccuracies. Many contemporary works assume $n < p$, in which case it is (in general) impossible to recover general $\boldsymbol{\beta}^*$ from the measurements. However, exploiting the fact that in many applications the signal of interest exhibits a low-dimensional structure opens the opportunity for using contemporary inference approaches from high dimensional statistics and compressed sensing. The low dimensional structure may be exhibited in different forms; for example, the signal of interest $\boldsymbol{\beta}^*$ might be entry-wise "sparse," i.e. it may have only a few non-zero entries, it might be sparse under an appropriate transformation, or it might be "group-wise" sparse, meaning that given a partition of its entries into groups, only a few groups may be non-zero. Remarkable results such as those in [29, 30] illustrate that, when the signal

of interest is sparse and the dictionary $\boldsymbol{X}$ satisfies certain structural conditions, one can accurately infer $\boldsymbol{\beta}^*$ by solving the so-called Lasso problem [11] even when the number of non-zero entries of $\boldsymbol{\beta}^*$ is nearly proportional to the number of measurements.

When the signal of interest is group-sparse, the group Lasso estimator [6],

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{2}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \sum_{g=1}^{G} \lambda_g ||\boldsymbol{\beta}_{\mathcal{I}_g}||_2, \tag{2.2}$$

can be used to infer the signal. In the formulation of interest here, $\boldsymbol{\beta}$ is expressed in terms of a given partition of its entries into G non-overlapping blocks (or groups)

$$\boldsymbol{\beta} = \left[(\boldsymbol{\beta}_{\mathcal{I}_1})^T (\boldsymbol{\beta}_{\mathcal{I}_2})^T \cdots (\boldsymbol{\beta}_{\mathcal{I}_G})^T\right]^T, \tag{2.3}$$

where $\boldsymbol{\beta}_{\mathcal{I}_g} \in \mathbb{R}^{d_g}$ represents the g-th constituent block of $\boldsymbol{\beta}$ with $\mathcal{I}_g$ denoting the subset of entries of $\boldsymbol{\beta}$ that belong to the g-th block, d_g denotes the cardinality of the g-th block, the $\lambda_g > 0$ are regularization parameters, and $\|\cdot\|_2$ denotes the Euclidean norm. This estimator exploits the extra knowledge about the natural grouping of the signal entries, and when this structure is present, its performance can exceed that of the standard Lasso estimator (which amounts to the case where each element of $\boldsymbol{\beta}$ is a singleton group) [6, 7, 31].

The existing studies that provide statistical guarantees for the group Lasso problem, when the measurements are generated according to (2.1), are diverse in terms of their statistical signal generation assumptions and the requirements they prescribe for successful recovery. In terms of the statistical model assumptions, a large body of work is focused on the case where the measurement matrix $\boldsymbol{X}$ is random [7], e.g., generated according to a Gaussian distribution [32,33]. Another line of work studies the asymptotic behavior of this recovery procedure when the number of measurements and unknown parameters are allowed to tend to infinity [34–36]. In terms of requirements for successful recovery, various conditions have been proposed so far, including the group RIP condition of [37] and the restricted group eigenvalue condition of [28,38]. Verifying such conditions for structured measurement matrices can be computationally prohibitive [39]; therefore, some existing works do not base their analyses on those requirements and instead use the concept of block coherence [40], which is computable in polynomial time. The recent effort [41] analyzes group-sparse estimation using structured dictionary matrices, with a sole focus on providing regression error guarantees.

Here, our investigation is motivated by an application in structural health monitoring, where we model our acquired data via the noisy linear model (2.1) for a signal that is assumed group-sparse in a fixed, structured dictionary [19, 20, 23]. In this context, and in contrast to the existing works discussed above, we seek finite sample, group-level *support recovery* guarantees for the group Lasso procedure, in order to pinpoint locations of material defects. We describe our motivating application in detail below.

2.1.1 Anomaly Detection for Structural Health Monitoring

In the past few decades, the need to improve the reliability of structural components and reduce their life-management costs has motivated the development of numerous structural diagnostics and structural health monitoring methodologies. Dynamics-based methods include popular techniques based on guided waves that are generated and received by transmitter-receiver pairs distributed over the structure, with detection processes that follow pulse-echo principles [42]. Namely, signatures of wave reflection are captured along each transmitter-receiver path, enabling the triangulation of the position of defects using data from multiple transducer pairs. Within this paradigm, numerous works have examined estimators of damage location likelihood from sparsely positioned sensors (see, e.g., [43–46]). However, these methods can suffer when ideality assumptions on the medium are relaxed (common in the context of damage formation and aging materials).

Recently, a powerful new class of diagnostic methodologies has emerged, leveraging the availability of laser-based sensing systems [47, 48]. Through the use of a *Scanning Laser Doppler Vibrometer* (SLDV) it is possible to perform non-contact measurements at a large number of points on a *scanning grid* defined on the surface of an object under test, thus providing full spatial reconstruction of the material's surface dynamic response (e.g., to an induced acoustic excitation). Dedicated image processing techniques have been developed which utilize laser acquired data and meet desired anomaly identification and visualization criteria (see, e.g., [49–52]).

Laser-based methods facilitate diagnostic methods in which the inference is performed directly on a data-rich, spatially reconstructed response. Central to this view is the notion that, from a data standpoint, a wavefield is a data cube, slices of which represent snapshots (or frames) of the dynamic response at different temporal instants.

The task of locating anomalies in a physical medium, then, can be recast as a problem of identifying atypical patterns in the observed data structures. Such efforts have recently been among the essential themes in machine learning and computer vision [53].

2.1.2 Approach

In this chapter, we utilize and expand notions from the sparsity-based source separation literature [54–57] and group Lasso inference to analyze the damage localization problem. The key observation underlying our approach is that SLDV measurements of a material subjected to narrowband acoustic excitation, acquired in the vicinity of the anomalous regions, exhibit different spatiotemporal behavior than do those acquired in the bulk of the material. We therefore attempt to decompose the acquired wavefield data into two components, one of which is a spatially-localized component arising near the defected areas while the other one is a generally smooth component in the pristine bulk of the structure; Fig. 2.1 illustrates one nominal measurement frame as well as its constituent components. This facilitates a *baseline-free*, agnostic inference approach whereby the locations of the defects in a material may be accurately estimated without a priori characterization of (a pristine version of) the medium. This feature distinguishes our method from the recent efforts in [58, 59] which also exploit group sparse inference techniques in the context of Lamb wave-based structural health monitoring but follow pitch-catch principles and require knowledge of the propagation model over the structure.

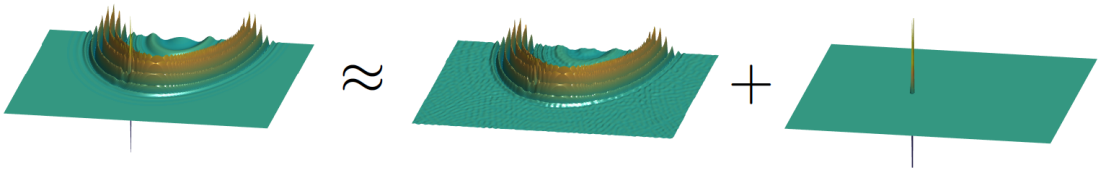


Figure 2.1: A snapshot of the wavefield measurement and its structurally-distinct components; the generally smooth component, which is characteristic of the undamaged bulk of the structure, and the spatially-localized component, which is zero except in the vicinity of the anomaly.

In order to separate the two structurally-distinct components of each measurement frame, we assume that (upon vectorizing the measurement snapshots) each component can be efficiently expressed as the product of an appropriate dictionary or basis matrix and a coefficient vector. The dictionaries should be chosen such that they capture the structural characteristics of their respective components. In the context of dictionary-based signal representation, this translates to choosing dictionaries that enable the characterization of the respective component in terms of the superposition of a few their columns. Since defects are generally spatially-localized, an appropriate dictionary for the defects is the identity matrix (i.e., the discrete Dirac basis), which comprises columns that are zero at every location except for one. Likewise, the *Discrete Cosine Transform* (DCT) matrix is onesuitable basis for the smooth component of the response from the undamaged regions. In this sense, our model is reminiscent of the basis pairs utilized in the initial works on Basis Pursuit [54, 60].

To further facilitate the task of detecting anomalies, we notice that the effect of anomalies will change the wavefield characteristics at several pixel locations adjacent to the defect. In other words, one can expect that anomalies manifest themselves as *spatially-contiguous* pixel blocks of the overall anomaly vector. Therefore, we propose to define a spatial grouping over the domain of the defect component and make use of a spatial block-sparsity-promoting technique over the anomalous component of the measurement decomposition. Imposing the spatial block-sparsity condition is justified by the fact that the bulk of a medium is undamaged and therefore most of the spatial blocks of the anomalous component should be zero blocks. In addition, since the effect of anomalies is usually persistent across multiple consecutive measurement frames (i.e., across time), we propose to extend the spatial grouping to a *spatiotemporal* one. This can be accomplished by partitioning the entries of multiple anomaly vectors, corresponding to multiple consecutive frames, into blocks which comprise spatially *and* temporally adjacent pixels.

In this setting, defect localization may be achieved by identifying the locations of the (nominally few) nonzero spatiotemporal groups describing the anomalous response, called the *group-level support* of the signal, in their respective dictionaries. Here, we analyze the performance of the group Lasso optimization for this support recovery task.

We consider two approaches to analyze the group support recovery of the group

Lasso. In the first (baseline) case we make no specific assumptions on the generative model of the signal except that it be group-sparse as described above. In the second we impose an (arguably natural) generative probabilistic model on the signal. In each case, we identify sufficient conditions under which the group Lasso succeeds in identifying the group-level support of the unknown signal. Motivated by the application outlined above, our specific focus here is on the *number* of recoverable nonzero groups (relative to the total number of groups), as well as the functional relationship between the group sparsity of the signal and the signal-to-noise ratio (SNR) – quantified in terms of the ratio between the Euclidean norms of the nonzero groups' coefficients and the additive noise variance – for which group Lasso support recovery provably succeeds.

Our results for the baseline setting are somewhat analogous to existing results analyzing support recovery for the group Lasso (e.g., in [35]), and are provided here largely to facilitate comparison with our results in the second setting, which improve upon the number of recoverable groups (relative to the total number of groups). As in [35], our analyses are based on an application of the primal-dual witness construction approach used in [61], under a predefined coefficient group structure, and for our second (stronger) result, under the specified generative signal model.

2.1.3 Notation and Organization

Throughout the chapter, bold-face lowercase and uppercase letters will be used to denote vectors and matrices, respectively. For a vector $\boldsymbol{v}$, we use $\|\boldsymbol{v}\|_2$ to denote its Euclidean norm and for a matrix $\boldsymbol{V}$, its spectral and Frobenius norms are denoted by $\|\boldsymbol{V}\|_{2\to 2}$ and $\|\boldsymbol{V}\|_F$, respectively. Moreover, the sum of the absolute values of the entries of a matrix $\boldsymbol{V}$ (or a vector $\boldsymbol{v}$) are denoted by $\|\boldsymbol{V}\|_1$ (or $\|\boldsymbol{v}\|_1$) and the maximum absolute value of entries is represented by $\|\boldsymbol{V}\|_\infty$ (or $\|\boldsymbol{v}\|_\infty$).

We use $[m]$ as the shorthand for the set $\{1, 2, \cdots, m\}$, for any integer m. If n denotes the length of $\boldsymbol{\beta}$ and the number of columns of $\boldsymbol{X}$, then for the index set $\mathcal{I}_g \subset [n]$, $\boldsymbol{\beta}_{\mathcal{I}_g}$ will denote the group of entries of $\boldsymbol{\beta}$ whose indices belong to this set and $\boldsymbol{X}_{\mathcal{I}_g}$ will denote the submatrix comprised of columns of $\boldsymbol{X}$ indexed by $\mathcal{I}_g$. For a column-wise block partitioned matrix $\boldsymbol{M} = [\boldsymbol{M}_{\mathcal{I}_1}\, \boldsymbol{M}_{\mathcal{I}_2} \cdots \boldsymbol{M}_{\mathcal{I}_G}]$ the norm $\|\boldsymbol{M}\|_{B,1}$ is defined as

$$\|\boldsymbol{M}\|_{B,1} := \max_{g\in[G]} \|\boldsymbol{M}_{\mathcal{I}_g}\|_{2\to 2}.$$

Throughout the chapter, we will use different notions of support defined as follows:

- $\mathcal{S}(\boldsymbol{\beta}) := \{j \in [n] : \beta_j \neq 0\}$ will be the support of $\boldsymbol{\beta} \in \mathbb{R}^n$.
- $\mathcal{G}(\boldsymbol{\beta}) := \{g \in [G] : \boldsymbol{\beta}_{\mathcal{I}_g} \neq \mathbf{0}\}$ will denote the set that contains the indices of the nonzero groups of $\boldsymbol{\beta}$, where G is the total number of groups.
- $\mathcal{S}_{\mathcal{G}}(\boldsymbol{\beta}) := \cup_{g \in \mathcal{G}(\boldsymbol{\beta})} \mathcal{I}_g$. In words, $\mathcal{S}_{\mathcal{G}}(\boldsymbol{\beta})$ will denote the set that contains all indices comprising groups that are nonzero (even if there are zero elements at those particular indices). Note that $\mathcal{S}(\boldsymbol{\beta}) \subseteq \mathcal{S}_{\mathcal{G}}(\boldsymbol{\beta})$.

We let

$$d_{\min} := \min_{g \in [G]} d_g \quad \text{and} \quad d_{\max} := \max_{g \in [G]} d_g$$

denote the minimum and maximum group sizes, respectively, and

$$d_{\mathcal{G}(\boldsymbol{\beta})} := \sum_{g \in \mathcal{G}(\boldsymbol{\beta})} d_g$$

be the total number of entries in the group-level support $\mathcal{G}(\boldsymbol{\beta})$ of $\boldsymbol{\beta}$. Similarly, we define

$$\lambda_{\min} := \min_{g \in [G]} \lambda_g \quad \text{and} \quad \lambda_{\max} := \max_{g \in [G]} \lambda_g$$

to be the minimum and maximum regularization constants, respectively, and let $\boldsymbol{\lambda}_{\mathcal{G}(\boldsymbol{\beta})}$ be the $|\mathcal{G}(\boldsymbol{\beta})|$-dimensional vector whose entries are the regularization parameters corresponding to the groups in $\mathcal{G}(\boldsymbol{\beta}^*)$. In order to clarify notation, we will use $\mathcal{G}^*$, $\mathcal{S}^*_{\mathcal{G}}$, and $d^*_{\mathcal{G}}$ as abbreviations for $\mathcal{G}(\boldsymbol{\beta}^*)$, $\mathcal{S}_{\mathcal{G}}(\boldsymbol{\beta}^*)$, and $d_{\mathcal{G}(\boldsymbol{\beta}^*)}$, respectively.

The rest of the chapter is organized as follows. We provide our main recovery results in Section 2.2, and discuss their implications in the context of our motivating application in Section 2.3. We validate our theoretical results experimentally in Section 2.4, where we evaluate the efficacy of the group Lasso for support recovery on both synthetic data (adhering to our generative signal model) as well as in an FEM (finite element method) simulation of our structural anomaly detection problem. Section A.1 outlines the main steps of the primal-dual witness construction approach, which is used for proving our main recovery result, and how we instantiate this framework under our statistical assumptions. Section 2.5 provides a few brief concluding comments and discussion of some future directions. Intermediate analytical results are relegated to the supplementary material.

2.2 Main Theoretical Results

Our main theoretical contribution here comes in the form of a new support recovery guarantee for the group Lasso estimator under a random signal model. As alluded above, we assume measurements acquired according to the linear model (2.1), and examine the performance of the group lasso estimator (2.2) under the assumption that the unknown $\boldsymbol{\beta}^*$ can be parsimoniously expressed in terms of a given partition of its entries into blocks, as in (2.3). We first present a baseline result applicable to deterministic signal models, then proceed to formulating our main result. [1]

In both settings, our recovery guarantees are expressed in terms of the inter-block and intra-block coherence parameters [40, 41] of the dictionary $\boldsymbol{X}$ which are defined with respect to a given column-wise block partition of $\boldsymbol{X}$.

Definition 2.2.1. *For any dictionary* $\boldsymbol{X} = [\boldsymbol{X}_{\mathcal{I}_1} \boldsymbol{X}_{\mathcal{I}_2} \cdots \boldsymbol{X}_{\mathcal{I}_G}]$ *with blocks* $\boldsymbol{X}_{\mathcal{I}_g} \in \mathbb{R}^{n \times d_g}$ *and whose columns all have unit Euclidean norm, the* inter-block coherence constant $\mu_B(\boldsymbol{X})$ *is defined as*

$$\mu_B(\boldsymbol{X}) := \max_{1 \leq g \neq g' \leq G} \|\boldsymbol{X}_{\mathcal{I}_g}^T \boldsymbol{X}_{\mathcal{I}_{g'}}\|_{2 \to 2}, \tag{2.4}$$

and the intra-block coherence parameter $\mu_I(\boldsymbol{X})$ *is defined as*

$$\mu_I(\boldsymbol{X}) := \max_{g \in [G]} \|\boldsymbol{X}_{\mathcal{I}_g}^T \boldsymbol{X}_{\mathcal{I}_g} - \boldsymbol{I}_{d_g \times d_g}\|_{2 \to 2}. \tag{2.5}$$

Here, $\mu_B(\boldsymbol{X})$ measures similarity between the blocks of $\boldsymbol{X}$ and reduces to the standard coherence parameter when the groups over the dictionary columns are singletons. Further, $\mu_I(\boldsymbol{X})$ measures the deviation of the blocks $\{\boldsymbol{X}_{\mathcal{I}_g}\}_{g \in [G]}$ from orthonormal ones.

2.2.1 Baseline Result

Our first theoretical result can be stated as follows; its proof is structurally similar to that of our next main result (though simpler), and is provided in the supplementary material, for completeness.

Theorem 2.2.1. *Consider the linear measurement model* (2.1) *with* $\boldsymbol{w} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_{n \times n})$. *Assume that*

[1] The material of this section and the following one are reprints from [20, 22, 24].

1. $|\mathcal{G}(\boldsymbol{\beta}^*)| \leq \min\left\{\frac{0.5-\mu_I(\boldsymbol{X})}{\mu_B(\boldsymbol{X})}+1, \sqrt{\frac{d_{\min}}{d_{\max}}}\cdot\frac{1}{4\mu_B(\boldsymbol{X})}\right\}$
2. $\|\boldsymbol{\beta}^*_{\mathcal{I}_g}\|_2 \geq 5\sigma(1+\epsilon)\left(\sqrt{d_g}+\sqrt{d^*_{\mathcal{G}}}\right),\ \forall g \in \mathcal{G}^*$
3. $\lambda_g = 4\sigma(1+\epsilon)\sqrt{d_g},\ \ \forall g \in [G]$

all hold for some

$$\epsilon \geq \sqrt{\frac{(1+\mu_I(\boldsymbol{X}))\,\log(pG)}{d_{\min}}}.$$

Then, the following hold simultaneously, with probability at least $1-6\,p^{-2\log 2}$:

- *the solution* $\widehat{\boldsymbol{\beta}}$ *of problem* (2.2) *will have the same group-level support as* $\boldsymbol{\beta}^*$*; that is,* $\mathcal{G}(\boldsymbol{\beta}^*) = \mathcal{G}(\widehat{\boldsymbol{\beta}})$*, and*
- $\left\|\widehat{\boldsymbol{\beta}}_{\mathcal{I}_g} - \boldsymbol{\beta}^*_{\mathcal{I}_g}\right\|_2 \leq 5\sigma(1+\epsilon)\left(\sqrt{d_g}+\sqrt{d^*_{\mathcal{G}}}\right),\quad \forall g \in \mathcal{G}^*.$

Remark 2.2.1. *As alluded above, this result is reminiscent of a main result of [35], though those results are asymptotic in nature, and the analogous SNR condition there was specified in terms of* $\|\boldsymbol{\beta}^*_{\mathcal{I}_g}\|_\infty$ *rather than the group Euclidean norm* $\|\boldsymbol{\beta}^*_{\mathcal{I}_g}\|_2$ *as here.*

The above theorem provides conditions under which the recovery of the true group-level support is achievable via the group Lasso. Note in particular that the first condition, which limits the group-sparsity level, relates the number of nonzero groups $|\mathcal{G}(\boldsymbol{\beta}^*)|$ to the inverse of the block coherence constant $\mu_B(\boldsymbol{X})$. This condition leads to sub-optimal scaling between the number of measurements and the number of nonzero groups, as will become clear in the context of our next main result, as well as in the next section, in the context of the material anomaly detection.

2.2.2 Strengthened Result

To strengthen the baseline result, we impose some mild statistical assumptions on the generation of the coefficient vector $\boldsymbol{\beta}^*$. Specifically, similar to [41], we assume the group-sparse vector $\boldsymbol{\beta}^* \in \mathbb{R}^p$ in (2.3) is randomly generated according to the assumptions outlined below:

M_1) The block support $\mathcal{G}^*$ of $\boldsymbol{\beta}^*$comprises $s := |\mathcal{G}^*|$ non-zero blocks, whose indices are selected uniformly at random from all subsets of $[G]$ of size s.

M_2) The non-zero entries of $\boldsymbol{\beta}^*$ are equally likely to be positive or negative:

$$\mathbb{E}[\text{sign}(\boldsymbol{\beta}^*_j)] = 0, \text{ for } j \in [p].$$

M_3) The non-zero blocks of $\boldsymbol{\beta}^*$ have statistically independent "directions." Specifically, it is assumed that

$$\Pr\left(\bigcap_{g\in\mathcal{G}^*} \frac{\boldsymbol{\beta}^*_{\mathcal{I}_g}}{\|\boldsymbol{\beta}^*_{\mathcal{I}_g}\|_F} \in \mathcal{A}_g\right) = \prod_{g\in\mathcal{G}^*_i} \Pr\left(\frac{\boldsymbol{\beta}^*_{\mathcal{I}_g}}{\|\boldsymbol{\beta}^*_{\mathcal{I}_g}\|_F} \in \mathcal{A}_g\right),$$

where for each g, $\mathcal{A}_g \subset \mathbb{S}^{d_g-1}$ with $\mathbb{S}^{d_g-1}$ representing the unit sphere in $\mathbb{R}^{d_g}$.

Utilizing this model, we obtain the following theorem, proved in Section A.1.

Theorem 2.2.2. *Consider the measurement model* (2.1) *with* $\boldsymbol{w} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_{n\times n})$. *If*

1. $\mu_I(\boldsymbol{X}) \le c_0$ *and* $\mu_B(\boldsymbol{X}) \le \sqrt{\frac{d_{\min}}{d^2_{\max}}}\,\frac{c_1}{\log p}$,
2. $|\mathcal{G}(\boldsymbol{\beta}^*)| \le \min\left\{\frac{c_2\, G}{\|\boldsymbol{X}\|^2_{2\to 2}\log p}, \frac{d_{\min}}{d^2_{\max}}\,\frac{c_2'\,\mu_B^{-2}(\boldsymbol{X})}{\log p}\right\}$,
3. $\|\boldsymbol{\beta}^*_{\mathcal{I}_g}\|_2 \ge 10\sigma(1+\epsilon)(\sqrt{d^*_{\mathcal{G}}} + \sqrt{d_g})\,\max\left\{1, \sqrt{\frac{s}{d_{\max}\log p}}\right\},\ \forall g \in \mathcal{G}(\boldsymbol{\beta}^*)$
4. $\lambda_g = 4\sigma(1+\epsilon)\sqrt{d_g},\ \forall g \in [G]$,

all hold for some positive constants c_0, $c_1 \le 0.001$, $c_2 \le \left[\sqrt{9 + \frac{1}{2}\left(\frac{1}{4} - 3c_0 - 48c_1\right)} - 3\right]^2$, $c_2' = \min\{c_2, 0.0001\}$, *and some*

$$\epsilon \ge \sqrt{\frac{(1+\mu_I(\boldsymbol{X}))\,\log(p\,G)}{d_{\min}}}.$$

Then the following hold simultaneously, with probability at least $1 - 12\,p^{-2\log 2}$:

- *the solution* $\widehat{\boldsymbol{\beta}}$ *of* (2.2) *is unique and has the same group-level support as* $\boldsymbol{\beta}^*$*; that is,* $\mathcal{G}(\boldsymbol{\beta}^*) = \mathcal{G}(\widehat{\boldsymbol{\beta}})$, *and*
- $\left\|\widehat{\boldsymbol{\beta}}_{\mathcal{I}_g} - \boldsymbol{\beta}^*_{\mathcal{I}_g}\right\|_2 \le 5\sigma(1+\epsilon)\left(\sqrt{d_g} + \sqrt{d^*_{\mathcal{G}}}\right),\ \forall g \in \mathcal{G}(\boldsymbol{\beta}^*)$.

Remark 2.2.2. *As required by the first theorem assumption, the support recovery guarantee relies on the well-conditioning of the dictionary $\boldsymbol{X}$. We measure the well-conditioning in terms of block coherence constants $\mu_I(\boldsymbol{X})$ and $\mu_B(\boldsymbol{X})$ of the dictionary. Fortunately, both constants can be computed in polynomial time for a given column-wise partitioned dictionary (unlike other quantities such as restricted isometry constant, which are widely used in proving similar recovery guarantees; but can be NP-hard to compute [39]). Regarding the material anomaly detection framework, this first assumption will impose very mild conditions on the problem parameters, as will be seen in the following sections.*

Remark 2.2.3. *The second condition specifies the requirement on the maximum number of allowable non-zero groups in the group-level support of $\boldsymbol{\beta}^*$ that can be recovered. Unlike the earlier recovery result, the condition provided here is less stringent since the block coherence parameter appears in the upper-bound in the form of $\mu_B^{-2}(\boldsymbol{X})$, which is a significant improvement over similar results, e.g. in [40, 62], that require $|\mathcal{G}^*|$ be bounded by functions of $\mu_B^{-1}(\boldsymbol{X})$.*

Remark 2.2.4. *The third assumption here (like the second assumption in our baseline result) is on the strength of the non-zero groups, which requires their magnitudes to be above a certain threshold depending on the noise variance σ. More discussions on this assumption, and its implications in our motivating application are provided in Section 2.4.*

Remark 2.2.5. *Finally, we note that our choices of the universal constants c_0, c_1, c_2, c_2' are not optimized here.*

2.3 Theoretical Results in the Context of Structural Diagnostics

As mentioned earlier, one of our goals is to quantify the performance of the group Lasso for laser-enabled anomaly localization in a structural health monitoring application. In this section we apply our main results from the previous section to that problem.

Assume that one vectorized snapshot of wavefield measurements, captured at time instant $t \in [T]$, is denoted by the vector $\boldsymbol{y}(t) \in \mathbb{R}^N$, where the integer N denotes the

total number of acquired measurements. In the case where the physical structure is a two-dimensional medium, every snapshot of measurements will be a two-dimensional image with N denoting the total number of pixels of the image. Moreover, assume that the matrix $\boldsymbol{Y} = [\boldsymbol{y}(1)\, \boldsymbol{y}(2) \cdots \boldsymbol{y}(T)] \in \mathbb{R}^{N \times T}$ stores all the measurement vectors for time instants 1 to T.

As discussed in the introduction, we aim to separate the spatially smooth component of wavefield measurements, which captures the response of the pristine bulk of the medium, from the spatially-localized component, which arises due to the presence of internal material defect(s). To perform the separation, we first assume that both components can be represented in terms of appropriate dictionaries, which capture structural characteristics of their respective components. To make the idea more formal, let $\boldsymbol{X}_{(1)} \in \mathbb{R}^{N \times p_1}$ and $\boldsymbol{X}_{(2)} \in \mathbb{R}^{N \times p_2}$ represent the dictionaries that appropriately represent the spatially-smooth and sparse components, respectively. Examples of the appropriate choices for $\boldsymbol{X}_{(1)}$ and $\boldsymbol{X}_{(2)}$, as alluded earlier, are the two-dimensional discrete cosine transform (DCT) and identity matrices, respectively (with $p_1 = p_2 = N$).

Given the knowledge of appropriate dictionaries, we assume the measurement matrix is generated by the following underlying model

$$\boldsymbol{Y} = \boldsymbol{X}_{(1)}\boldsymbol{B}^*_{(1)} + \boldsymbol{X}_{(2)}\boldsymbol{B}^*_{(2)} + \boldsymbol{W}, \tag{2.6}$$

where $\boldsymbol{B}^*_{(1)} \in \mathbb{R}^{N \times T}$ and $\boldsymbol{B}^*_{(2)} \in \mathbb{R}^{N \times T}$ denote the corresponding coefficient matrices and $\boldsymbol{W} \in \mathbb{R}^{N \times T}$ represents noise and model ambiguities. In this model the first term $\boldsymbol{X}_{(1)}\boldsymbol{B}^*_{(1)}$ stands for the smooth component of measurements generated by the pristine bulk of the medium and $\boldsymbol{X}_{(2)}\boldsymbol{B}^*_{(2)}$ models the defect component. Given the above model the problem of anomaly detection reduces to finding the support of the defect component $\boldsymbol{X}_{(2)}\boldsymbol{B}^*_{(2)}$ (or simply $\boldsymbol{B}^*_{(2)}$ when $\boldsymbol{X}_{(2)} = \boldsymbol{I}_{N \times N}$).

A practical assumption that improves the performance of the anomaly detection procedure is that defects manifest themselves in the form of spatially-contiguous groups of pixels. Therefore, given a *spatial* partition of the measurement domain into groups of $D \geq 1$ adjacent pixels, one can expect the pixels within a group to be corrupted once a defect is present in that region. In the measurement model expressed by (2.6), with $\boldsymbol{X}_{(2)} = \boldsymbol{I}_{N \times N}$, this implies that each column of $\boldsymbol{B}^*_{(2)}$ can be partitioned into $G_2 := p_2/D = N/D$ groups of size D, where the entries within a group are adjacent

pixels in the two-dimensional representation of the measurements. Furthermore, since the effect of anomalies changes the wavefield characteristics across multiple consecutive frames, it makes sense to define a more general *spatiotemporal* grouping over $\boldsymbol{B}^*_{(2)}$. Then the imposed grouping will partition the coefficients in $\boldsymbol{B}^*_{(2)}$ into G_2 sub-matrices of size $D \times T$, where the entries of a sub-matrix are spatiotemporally adjacent. On the other hand, a *temporal* grouping can be applied to the entries of the coefficient matrix $\boldsymbol{B}^*_{(1)}$ corresponding to the smooth component, with the idea that the same frequencies (i.e. the same columns of the DCT dictionary) should appear in the decomposition of consecutive frames. Doing so, $\boldsymbol{B}^*_{(1)}$ can be partitioned into $G_1 := p_1$ sub-matrices of dimensions $1 \times T$. To enable the recovery of $\boldsymbol{B}^*_{(1)}$ and $\boldsymbol{B}^*_{(2)}$ from the measurements in (2.6), we assume both coefficient matrices are block-sparse with respect to the groupings described, i.e. only a few groups in the partition of every coefficient matrix are non-zero.

Given these assumptions we propose to estimate the true coefficient matrices $\boldsymbol{B}^*_{(1)}$ and $\boldsymbol{B}^*_{(2)}$ by $\widehat{\boldsymbol{B}}_{(1)}$ and $\widehat{\boldsymbol{B}}_{(2)}$, which are solutions of the following optimization problem

$$\min_{\boldsymbol{B}_{(1)},\boldsymbol{B}_{(2)}\in\mathbb{R}^{N\times T}} \Bigg\{ \frac{1}{2} \left\| \boldsymbol{Y} - \boldsymbol{X}_{(1)}\boldsymbol{B}_{(1)} - \boldsymbol{X}_{(2)}\boldsymbol{B}_{(2)} \right\|_F^2 + \lambda_1 \sum_{g_1\in[G_1]} \left\| (\boldsymbol{B}_{(1)})_{\mathcal{I}_{g_1}} \right\|_F + \lambda_2 \sum_{g_2\in[G_2]} \left\| (\boldsymbol{B}_{(2)})_{\mathcal{I}_{g_2}} \right\|_F \Bigg\}, \tag{2.7}$$

where λ_1 and λ_2 are positive scalars, and g_1 and g_2 index the blocks of $\boldsymbol{B}_{(1)}$ and $\boldsymbol{B}_{(2)}$, respectively, which are formed according to the grouping techniques described above. In this formulation, minimizing the first term will ensure that the model fits the measurements; while minimizing the last two terms guarantee that the two components comprise a small number of atoms from the corresponding dictionaries. In particular, minimizing the third term promotes the group sparsity of the recovered anomaly component with respect to the specified spatiotemporal grouping.

To enable the application of the theoretical results developed in the previous section, we adopt a vectorized representation of the measurement model (2.6). Specifically, we choose $\boldsymbol{y} \in \mathbb{R}^n$ to denote the vector of measurements acquired by stacking all the T columns of $\boldsymbol{Y}$ in one vector (therefore obtaining a measurement vector of length $n := NT$). Upon vectorizing the entire measurement model (2.6), the new representation becomes

$$\boldsymbol{y} = \widetilde{\boldsymbol{X}}_{(1)}\boldsymbol{\beta}^*_{(1)} + \widetilde{\boldsymbol{X}}_{(2)}\boldsymbol{\beta}^*_{(2)} + \boldsymbol{w}, \tag{2.8}$$

where $\boldsymbol{y} = \text{vec}\,(\boldsymbol{Y}) \in \mathbb{R}^n$, $\boldsymbol{\beta}^*_{(1)} = \text{vec}(\boldsymbol{B}^*_{(1)}) \in \mathbb{R}^n$, $\boldsymbol{\beta}^*_{(2)} = \text{vec}(\boldsymbol{B}^*_{(2)}) \in \mathbb{R}^n$, $\boldsymbol{w} = \text{vec}(\boldsymbol{W}) \in \mathbb{R}^n$ are vectors, with the vectorization operator $\text{vec}(\cdot)$ stacking the columns of the argument matrix into a single-column vector, and $\widetilde{\boldsymbol{X}}_{(1)}$ and $\widetilde{\boldsymbol{X}}_{(2)}$ are Kronecker-structured dictionaries given as $\widetilde{\boldsymbol{X}}_{(i)} = \boldsymbol{I}_{T\times T} \otimes \boldsymbol{X}_{(i)}$, for $i = 1, 2$. Notice that after the vectorization, the previously-discussed partitions over the entries of $\boldsymbol{B}^*_{(1)}$ and $\boldsymbol{B}^*_{(2)}$ result in non-canonical groups, which are either of size T (for the groups over the smooth component) or of size DT (for the groups over the second spatially-sparse component). Using vector notation, the problem (2.7) can be recast as

$$\min_{\boldsymbol{\beta}_{(1)},\boldsymbol{\beta}_{(2)}\in\mathbb{R}^n} \Bigg\{ \frac{1}{2} \left\| \boldsymbol{y} - \widetilde{\boldsymbol{X}}_{(1)}\boldsymbol{\beta}_{(1)} - \widetilde{\boldsymbol{X}}_{(2)}\boldsymbol{\beta}_{(2)} \right\|_2^2 + \lambda_1 \sum_{g_1\in[G_1]} \left\| (\boldsymbol{\beta}_{(1)})_{\mathcal{I}_{g_1}} \right\|_2 + \lambda_2 \sum_{g_2\in[G_2]} \left\| (\boldsymbol{\beta}_{(2)})_{\mathcal{I}_{g_2}} \right\|_2 \Bigg\}. \tag{2.9}$$

We may write the model (2.8) in terms of the overall dictionary $\boldsymbol{X} := \left[\widetilde{\boldsymbol{X}}_{(1)} \,|\, \widetilde{\boldsymbol{X}}_{(2)}\right] \in \mathbb{R}^{n\times p}$, with $p := 2n$, and the coefficient vector $(\boldsymbol{\beta}^*)^T := [(\boldsymbol{\beta}^*_{(1)})^T \,|\, (\boldsymbol{\beta}^*_{(2)})^T] \in \mathbb{R}^p$ as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{w}, \tag{2.10}$$

which is the linear measurement model discussed in the previous section. The implications of Theorem 2.2.1 for the anomaly detection scenario are outlined below.

Corollary 2.3.1. *Consider the linear measurement model* (2.6) *with* $\boldsymbol{X}_{(1)}$ *and* $\boldsymbol{X}_{(2)}$ *specialized to the 2D-DCT and identity matrices of size* $N \times N$*, respectively, and the entries of* $\boldsymbol{W}$ *independently drawn from the Gaussian distribution* $\mathcal{N}(0, \sigma^2)$*. Moreover, suppose* $\boldsymbol{B}^*_{(1)}$ *and* $\boldsymbol{B}^*_{(2)}$ *have* s_1 *and* s_2 *non-zero groups, respectively, which are arbitrarily drawn from the partitions defined over the entries of these matrices. If*

1. $s = s_1 + s_2 \leq \frac{\sqrt{N}}{8D}$
2. $\min_{g_1\in\mathcal{G}^*_1} \left\| (\boldsymbol{B}^*_{(1)})_{\mathcal{I}_{g_1}} \right\|_F \geq 5\sigma\sqrt{T}(1+\epsilon)\left(1+\sqrt{s_1+s_2D}\right)$
3. $\min_{g_2\in\mathcal{G}^*_2} \left\| (\boldsymbol{B}^*_{(2)})_{\mathcal{I}_{g_2}} \right\|_F \geq 5\sigma\sqrt{T}(1+\epsilon)\left(\sqrt{D}+\sqrt{s_1+s_2D}\right)$
4. $\lambda_1 = 4\sigma\,(1+\epsilon)\,\sqrt{T}$ *and* $\lambda_2 = 4\sigma\,(1+\epsilon)\,\sqrt{TD}$

all hold for some $\epsilon \geq \sqrt{\frac{2\log(2NT)}{T}}$*, then the group-level support of* $\widehat{\boldsymbol{B}}_{(1)}$ *and* $\widehat{\boldsymbol{B}}_{(2)}$ *will exactly match those of* $\boldsymbol{B}^*_{(1)}$ *and* $\boldsymbol{B}^*_{(2)}$*, respectively, with probability at least* $1-6\,(2NT)^{-2\log 2}$.

As this result (whose proof appears in the supplementary material) asserts, in order for us to be able to exactly recover the group-level supports of $\boldsymbol{B}^*_{(1)}$ and $\boldsymbol{B}^*_{(2)}$, the dependence of the number of non-zero groups s on N is no larger than $\sqrt{N}$. This kind of result, derived using only coherence-based arguments, is sometimes known as the "square-root bottleneck;" see, e.g., [29].

Next we summarize the implications of our main theoretical result, Theorem 2.2.2, for the anomaly detection scenario described above. As the theorem states under the statistical assumptions M_1, M_2, and M_3, and some extra conditions on the number of anomalies and their severity, exact detection of anomalous groups is possible.

Corollary 2.3.2. *Consider the linear measurement model* (2.6) *with* $\boldsymbol{X}_{(1)}$ *and* $\boldsymbol{X}_{(2)}$ *specialized to the two-dimensional DCT and identity matrices of size* $N \times N$*, respectively, and the entries of* $\boldsymbol{W}$ *drawn independently from* $\mathcal{N}(0, \sigma^2)$*. Moreover, suppose* $\boldsymbol{B}^* := [\boldsymbol{B}^*_{(1)} \,|\, \boldsymbol{B}^*_{(2)}] \in \mathbb{R}^{N \times 2T}$ *has* s *randomly-selected non-zero groups selected according to the statistical assumptions* M_1*,* M_2*, and* M_3*. If*

1. $\sqrt{N} \geq \frac{2\log(2NT)}{c_1}\sqrt{D^3\,T}$
2. $s \leq \frac{c_2 N}{TD^3\log(2NT)}$
3. $\forall g \in \mathcal{G}_1^* : \left\|(\boldsymbol{B}^*_{(1)})_{\mathcal{I}_g}\right\|_F \geq 10(1+\epsilon)\sigma\sqrt{T}\left(1+\sqrt{s_1+s_2D}\right)\cdot\max\left\{1, \sqrt{\frac{s}{TD\,\log(2NT)}}\right\}$
4. $\forall g \in \mathcal{G}_2^* : \left\|(\boldsymbol{B}^*_{(2)})_{\mathcal{I}_g}\right\|_F \geq 10(1+\epsilon)\sigma\sqrt{T}\left(\sqrt{D}+\sqrt{s_1+s_2D}\right)\cdot\max\left\{1, \sqrt{\frac{s}{TD\,\log(2NT)}}\right\}$
5. $\lambda_1 = 4\sigma(1+\epsilon)\sqrt{T}$ *and* $\lambda_2 = 4\sigma(1+\epsilon)\sqrt{DT}$

all hold for $c_1 \leq 0.001$*,* $c_2 \leq 0.0001$*, and*

$$\epsilon \geq \sqrt{\frac{2\log(2NT)}{T}},$$

where s_1 *and* s_2 *denote the number of nonzero groups selected in* $\boldsymbol{B}^*_{(1)}$ *and* $\boldsymbol{B}^*_{(2)}$ *respectively, then the group-level support of* $\widehat{\boldsymbol{B}}$ *will exactly match that of* $\boldsymbol{B}^*$ *with probability at least* $1 - 12\,(2NT)^{-2\log 2}$*.*

The above result, whose proof is provided in the supplementary material, is a direct consequence of Theorem 2.2.2 of the previous section. As the theorem asserts, the group-level support of $\boldsymbol{B}^*$ should be drawn uniformly at random from among the $\binom{G}{s}$ different

subsets of $[G]$ comprised of s elements. The randomness of the (group-level) support of $\boldsymbol{B}^*$ enables us to bring in tools from the concentration of random variables theory to prove the sufficient conditions of the Theorem. The second condition provides an upper bound on how many anomalous groups can be detected by the convex demixing procedure in (2.7), and the third gives lower bounds for the strength of the non-zero groups in order for them to be detectable using the group Lasso approach.

2.4 Numerical Experiments

In this section we test the ability of the group Lasso formulation (2.2) in recovering the non-zero coefficients $\boldsymbol{\beta}^*$ for dictionary-based representation of the measurements. The first set of experiments that are presented here are carried out using synthetically generated measurements. These experiments will be followed by experiments on real-world data. [2]

2.4.1 Phase Transition Diagram

In this sub-section, we use synthetically generated data to study the relationship between the group-sparsity level of the unknown coefficient vector and the strength of non-zero groups that guarantees successful recovery. The inspiration for this investigation comes from the conditions 3 and 4 of Corollary 2.3.2 (and similar conditions in Corollary 2.3.1), which present lower bounds on $\left\|(\boldsymbol{B}^*_{(1)})_{\mathcal{I}_g}\right\|_2$ and $\left\|(\boldsymbol{B}^*_{(2)})_{\mathcal{I}_g}\right\|_2$ as part of the sufficient conditions for having exact support recovery.

Operating under the measurement model assumptions introduced in section 2.3, we generate measurements according to Equation (2.6). More specifically, we generate $T = 8$ frames of measurements, each of dimensions 100×100, therefore $N = 10^4$ in (2.6). To generate each frame we choose $\boldsymbol{X}_{(1)}$ to be the $N \times N$ 2D-DCT matrix, and set $\boldsymbol{X}_{(2)}$ to be the $N \times N$ identity matrix (to explain the sizes of the dictionaries we would like to note that $\boldsymbol{X}_{(1)}$ and $\boldsymbol{X}_{(2)}$ operate on vectorized images). Once $\boldsymbol{X}_{(1)}$ and $\boldsymbol{X}_{(2)}$ are selected, it remains to generate $\boldsymbol{B}^*_{(1)} \in \mathbb{R}^{N\times T}$, $\boldsymbol{B}^*_{(2)} \in \mathbb{R}^{N\times T}$ and $\boldsymbol{W} \in \mathbb{R}^{N\times T}$ in order to make the measurement vectors as according to (2.6).

[2] The material in this section is partly published in [19, 21, 23].

Inspired by the spatial contiguity assumption of anomalies, we assume each column of $\boldsymbol{B}^*_{(2)}$, which corresponds to a vectorized 100×100 image, is partitioned into groups of size $D = d^2$, where each group corresponds to a $d \times d$ spatially-contiguous block in the original image representation of the column. Here we report the results for $d = 2$ (therefore $D = 4$). Also by the assumption of the temporal persistency of anomalies, we extend the grouping across all the frames resulting in the entries of $\boldsymbol{B}^*_{(2)}$ be partitioned into groups of size $d^2 \times T$. Doing so the total number of blocks over the support of $\boldsymbol{B}^*_{(2)}$ will become $G_2 = (N/d)^2$. For the coefficient matrix $\boldsymbol{B}^*_{(1)}$ corresponding to the spatially-smooth component, we assume no spatial grouping structure over its columns; therefore each of its $G_1 = N = 10^4$ rows will comprise a group. Next, in order to give values to $\boldsymbol{B}^* = \left[\boldsymbol{B}^*_{(1)}\ \boldsymbol{B}^*_{(2)}\right]$ we first choose $s = s_1 + s_2$ out of the entire $G = G_1 + G_2$ blocks uniformly at random (for s ranging from 1 to 800) and then set the selected entries to i.i.d. standard Gaussian values. Finally, the noise matrix $\boldsymbol{W}$ is set to have i.i.d. entries generated according to $\mathcal{N}(0, \alpha^{-2})$, where α can be thought as the parameter which defines the signal to noise ratio and is varied from 0 to 80.

For each choice of the (s_2, α) pair, we generate $MC = 100$ different realizations and test the performance of the proposed algorithm in recovering the coefficients. The numerical algorithm that we have adopted for solving the corresponding optimization problem (2.7) is alternatively minimizing the objective with respect to two coefficient matrices $\boldsymbol{B}_{(1)}$ and $\boldsymbol{B}_{(2)}$. The algorithm is detailed in Algorithm 1.

Algorithm 1 Alternating Minimization

Initialize $\boldsymbol{B}_{(1)} \leftarrow \boldsymbol{0}$ and $\boldsymbol{B}_{(2)} \leftarrow \boldsymbol{0}$

repeat

$\boldsymbol{R}_{(1)} \leftarrow \boldsymbol{X}^T_{(1)} \left(\boldsymbol{Y} - \boldsymbol{X}_{(2)}\boldsymbol{B}_{(2)}\right)$

$(\boldsymbol{B}_{(1)})_{\mathcal{I}_{g_1}} \leftarrow \left(1 - \frac{\lambda_1}{\|(\boldsymbol{R}_{(1)})_{\mathcal{I}_{g_1}}\|_F}\right)_+ (\boldsymbol{R}_{(1)})_{\mathcal{I}_{g_1}}, \forall g_1 \in \mathcal{G}_1$

$\boldsymbol{R}_{(2)} \leftarrow \boldsymbol{X}^T_{(2)} \left(\boldsymbol{Y} - \boldsymbol{X}_{(1)}\boldsymbol{B}_{(1)}\right)$

$(\boldsymbol{B}_{(2)})_{\mathcal{I}_{g_2}} \leftarrow \left(1 - \frac{\lambda_2}{\|(\boldsymbol{R}_{(2)})_{\mathcal{I}_{g_2}}\|_F}\right)_+ (\boldsymbol{R}_{(2)})_{\mathcal{I}_{g_2}}, \forall g_2 \in \mathcal{G}_2$

until convergence

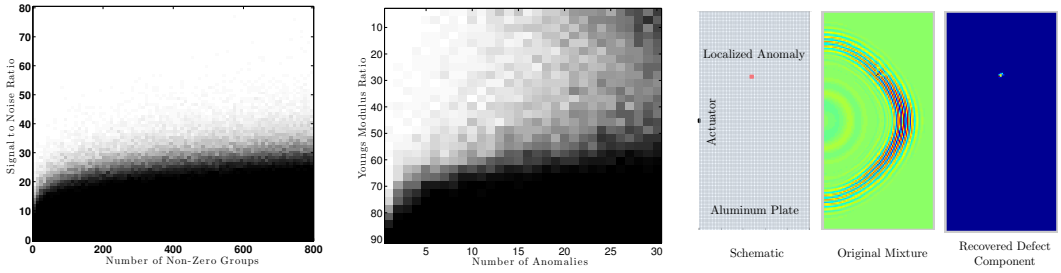
Since by the specific grouping defined over the entries of $\boldsymbol{B}^*_{(1)}$ and $\boldsymbol{B}^*_{(2)}$ only two distinct group sizes exist, the regularization parameters are set to either $\lambda_1 = \frac{5}{\alpha}\sqrt{T}$ for all the groups defined over the support of $\boldsymbol{B}^*_{(1)}$ or to $\lambda_2 = \frac{5}{\alpha}\sqrt{T\,d^2}$ for all the groups over the support of $\boldsymbol{B}^*_{(2)}$. The probability of success is then simply defined as the ratio of the number of realizations for which the successful recovery of the group-level support of both $\boldsymbol{B}^*_{(1)}$ and $\boldsymbol{B}^*_{(2)}$ occurs to the total number of trails MC. To avoid errors due to numerical inaccuracies, we declare the groups of the recovered coefficient matrices as being non-zero if their norms exceed a precision constant $\epsilon_p = 10^{-6}$ times the norms of their corresponding groups in the ground-truth coefficient matrices.

The left-hand side panel of Fig. 2.2 shows the phase transition diagram for the described set up. According to the diagram, as the number of active non-zero groups increases, to enable successful group-level support recovery one needs to increase the strength of the active groups as well. Also, when the group sparsity level goes above almost a hundred, the edge of the black region becomes almost a straight line, which agrees with the sufficient conditions of our Corollary 2.3.2. More precisely, notice that conditions 3 and 4 of this Theorem require $\|\boldsymbol{B}^*_{\mathcal{I}_g}\|_F = \Omega(s)$ for every non-zero group g, whenever $s \geq TD\log(2NT)$ in order to enable successful group support recovery.

2.4.2 Finite Element Simulations

We also use synthetic wavefield measurements generated by finite element simulations to study the relationship between the number of defects, the severity of them, and the ability of the proposed group Lasso estimator in successful defect recovery. To do so, we model an aluminum plate, with dimensions 100 cm $\times$ 50 cm and thickness 5 mm, which is probed by a flexural wavefield induced by an actuator located in the middle of the left edge of the domain. Localized anomalies are introduced by reducing the Young's modulus constant of the material of a 1.5 cm $\times$ 1.5 cm region to simulate a soft inclusion. The actuator is set to generate $N_c = 5$ bursts of a narrow-band sine wave at the frequency $f_c = 10^5$. We then record 100 (two milliseconds apart) snapshots of the nodal displacements, over a grid with 160 $\times$ 80 nodes, and store them as columns of a measurements matrix $\boldsymbol{Y}$. Given the grid size and the number of frames, the measurement matrix $\boldsymbol{Y}$ ends up having dimensions $N \times 100$, where $N = 160 \times 80 = 12800$. Fig. 2.2 (a) shows the schematic of the simulated plate, a wavefield snapshot,

and the recovered defect component for that snapshot.



(a) (b) (c)

Figure 2.2: Panel (a) from left to right shows the phase transition diagram for the experiment with synthetically generated Gaussian data. The vertical axis denotes the value of signal to noise ratio varied through the scalar α^{-1}. Panel (b) shows the resulting phase transition diagram from finite element experiments. The vertical axis denotes the ratio of the Young's modulus constant of defects to the bulk of the medium. The smaller this ratio is, the more severe the anomalies would be. After performing our decomposition, a strong mismatch would result in larger magnitudes of non-zero coefficients corresponding to the anomalous component and therefore higher signal to noise ratio. Panel (c) shows the schematic of aluminum plate with actuator and soft inclusion, a snapshot of wavefield, and the recovered defect for that snapshot.

Similar to the previous sub-section, we would like to generate a phase transition diagram for the successful recovery rate of our procedure, with the horizontal and vertical axes indicating the number of defects and their severity level, respectively. In this experiment, we vary the number of anomalies between one and thirty. Having the number of defects fixed, we then place them at randomly selected locations over the surface of the simulated structure. To change defects' severity we use a scalar parameter $\eta \in (0, 1)$, which yields the Young's modulus constant of defected regions once multiplied by the Young's modulus constant of the healthy bulk of the structure. On the vertical axis of the phase transition diagram the defect severity is changed by raising η to different integer powers i, where i takes values between one and thirty. Intuitively speaking, as the integer power i increases the defect severity increases as well, since the Young's modulus constant of defected regions become a smaller fraction of that corresponding

to the healthy regions of the structure, which in turn makes the recovery easier. In the current experiment we set $\eta = 0.9$. We solve the group Lasso problem (2.7) for five consecutive frames, i.e. $T = 5$, and adopt a partitioning of the defect component coefficient vectors into spatial groups of size four pixels. The regularization parameters were experimentally tuned to $\lambda_1 = 0.005$ and $\lambda_2 = 0.12$ for the groups over the smooth and sparse components, respectively. We repeat the experiment 50 times for every specialization of the number of defects and their severity level. Fig 2.2 (b) shows the phase transition diagram for this experiment. Interestingly, the overall trend of the phase transition diagram resembles the diagram of the former sub-section. In particular, we again observe an almost-linear transition edge for the medium range of sparsity values. In fact, by increasing the mismatch between the Young's modulus constant of defects and the rest of the medium, local displacements at the place of anomalies increase. The displacements are effectively captured by the sparse coefficient matrix of our decomposition model and therefore contribute to stronger coefficient values in this matrix.

Finally, we would like to note modifying the Young's modulus is but one principled approach to adjust the strength of an anomaly in a physical setting. Properly speaking, by adjusting this parameter, we are varying the contrast in elastic properties (acoustic mismatch). By extension, we can also model partial holes (see [23], which reports our experiments with partial holes).

2.4.3 Synthetic Experiments

We put the method to the test against synthetic data obtained via finite element (FE) simulations. For our benchmark problem, we model a thin Aluminum plate probed by a flexural wavefield induced by an actuator located in the middle of the left edge of the domain. A localized anomaly is introduced in the domain by reducing (by two orders of magnitude) the Young's modulus and density of the material inside a small region, to simulate a soft inclusion (or a partial hole), as schematically shown in panel (a) of Fig. 2.3. We record and vectorize the nodal displacements and we arrange them as columns the response data matrix $\boldsymbol{X}$; panel (b) of Fig. 2.3 shows the original mixture, i.e., a slice frame of the propagating wavefield; the smooth and sparse components obtained though the demixing of the response data matrix $\boldsymbol{X}$ according to Eq. 2.7 are

shown in panels (c) and (d), respectively.

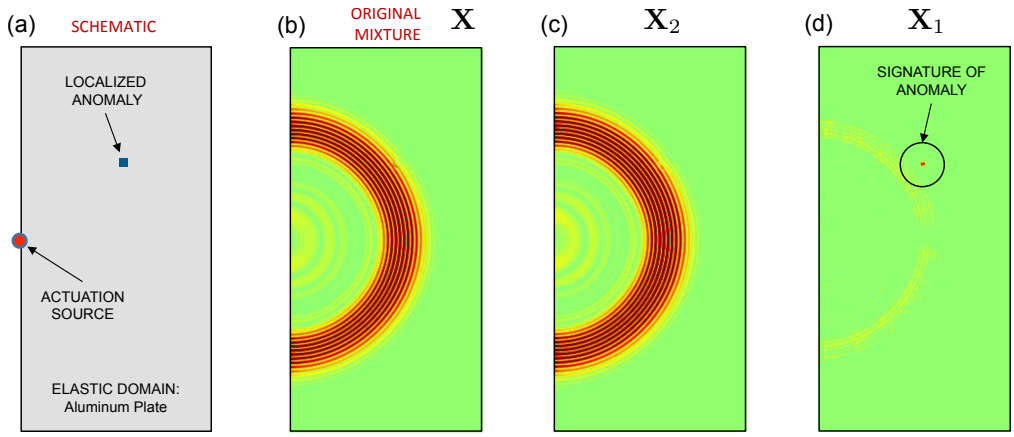
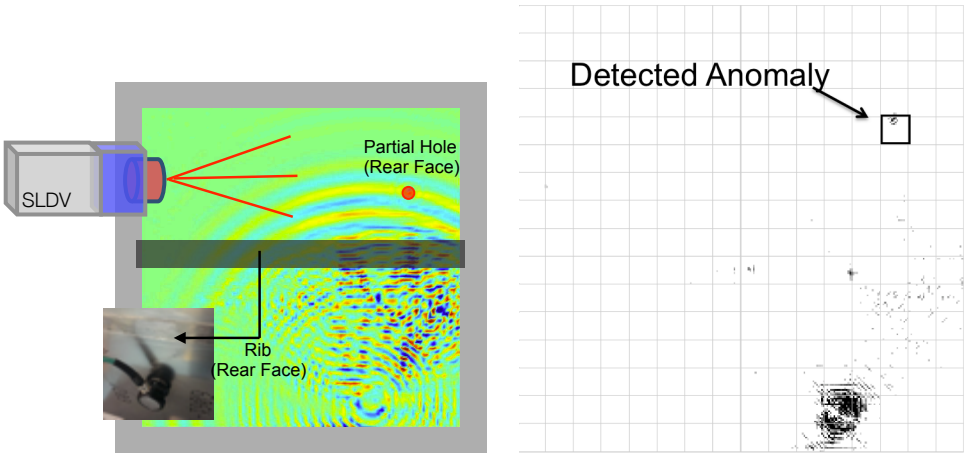


Figure 2.3: Results of anomaly detection and triangulation in a simulated defected plate. (a) Schematic of Aluminum plate with soft inclusion. (b) Snapshot of wavefield. (c) Smooth component $\boldsymbol{X}_2\boldsymbol{B}_2$ capturing the bulk response. (d) Sparse component $\boldsymbol{X}_1\boldsymbol{B}_1$ correctly pinpointing the anomaly.

2.4.4 Real-world Data Experiments

We examine the efficacy of the method by testing it against experimentally acquired data. We consider an Aluminum plate with dimensions 61 $\times$ 61 cm and thickness 2.54 mm excited with a piezoelectric transducer generating a 5-cycle burst with carrier frequency $f_c = 200\,\text{kHz}$. To test the robustness of the method against benign structural heterogeneity, we introduce a stiffening rib glued to the back face of the plate between the defect and the transducer. The wavefield is reconstructed from surface velocity data using the Polytec PSV-400-3D SLDV. A defect is introduced by drilling a partial hole on the plate surface which is left unscanned by the SLDV. Panel (a) in Fig. 2.4 shows the strong reflections that occur as the propagating wavefield impinges on the rib. In addition to the reflections, we observe a cascade of noisy features in the neighborhood of the rib, arguably a byproduct of the non-ideal (possibly nonlinear) contact conditions between the rib and the plate. The performance of our algorithm is shown in panel (b) in Fig. 2.4, where it can be seen that the proposed method allows successful triangulation of the anomaly despite the pronounced level of wavefield distortion.



(a) (b)

Figure 2.4: Detection of defect on the rear surface of Aluminum plate. Panel (a) shows the superimposed wavefield snapshot and the schematic of plate with stiffening rib affixed to the rear surface between excitation source and defect. Panel (b) shows successful localization of the anomaly despite the pronounced level of wavefield distortion due to the strong scattering from the rib.

2.5 Conclusion

In this chapter we examined recovery of group-sparse signals from low-dimensional noisy linear measurements using the group Lasso estimation procedure, motivated by a defect localization application in non-destructive evaluation. We established practically relevant group-level support recovery guarantees for non-asymptotic regimes in terms of the block coherence parameter, and validated our analytical results via simulation on both synthetic data, as well as simulated data generated according to a realistic model for our motivating defect localization application.

Chapter 3

An Accelerated Alternating Direction Method of Multipliers

As we mentioned in the introduction chapter, the recent advances in many areas of machine learning has led to formulating many arising problems into an optimization formulation. [1] Therefore, the proposed methodologies in these areas, require solving an optimization problem in their core and their applicability is dependent on solving such problems as fast and efficiently as possible. The proposed algorithms for solving such optimization problems should remain efficient as the size of the problem grows. This scalability criterion would cross out many traditional optimization methods such as interior point methods [63] and Newton method which are based on complicated iterations (due to requiring the second order information and matrix inversion). An alternative approach is to use first order methods that have lower cost per iteration, but show slower convergence. Unfortunately, general first order methods [64–66] require projections to the problem solution set. Such projections, even for solution sets that are defined by simple linear constraints, can be intractable in high dimensions. Therefore the applicability of such methods is limited.

Our focus in this chapter is on an alternative procedure that helps us deal with linear equality constraints. The algorithm that we will discuss is called Alternating Direction Method of Multiplier (ADMM) which has a long history in literature. It

[1] All the findings reported in this chapter are published in [25].

was first proposed in [67] and has recently gained lots of attention [68, 69] due to its simplicity and wide range of problems that it covers. Specifically, ADMM is designed for solving convex optimization problems of the form

$$\min_{\boldsymbol{x},\boldsymbol{y}} f_1(\boldsymbol{x}) + f_2(\boldsymbol{y}) \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{y} = \boldsymbol{c} \tag{3.1}$$

where $\boldsymbol{x} \in \mathbb{R}^{n_1}$, $\boldsymbol{y} \in \mathbb{R}^{n_2}$ are the optimization variables, $\boldsymbol{A} \in \mathbb{R}^{m\times n_1}$, $\boldsymbol{B} \in \mathbb{R}^{m\times n_2}$ are linear operators, $\boldsymbol{c} \in \mathbb{R}^m$ is a vector of data, and finally f_1 and f_2 are closed convex functions. As the formulation suggests, the objective is separable across the variables, while the constraints are coupling the variables. Such coupling linear equality constraints are not easy to deal with in general.

ADMM [68,69] is an iterative method that uses a Gauss Seidel type update to solve (3.1). Given a penalty parameter $\tau > 0$, ADMM minimizes the augmented Lagrangian

$$L(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\lambda}) = f_1(\boldsymbol{x}) + f_2(\boldsymbol{y}) - \langle \boldsymbol{\lambda}, \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{y} - \boldsymbol{c}\rangle + \frac{\tau}{2}\|\boldsymbol{c} - \boldsymbol{A}\boldsymbol{x} - \boldsymbol{B}\boldsymbol{y}\|^2$$

with respect to $\boldsymbol{x}$ and $\boldsymbol{y}$ alternatively and then updates the dual variable $\boldsymbol{\lambda} \in \mathbb{R}^m$. Steps of ADMM are summarized in Algorithm 2.

One of the main advantages of ADMM framework to other methods is its flexibility towards parallel computation [68]. As a result, in many applications it might be favorable to cast an unconstrained optimization problem into a constrained form (by introducing new variables) and solve the resulting constrained formulation by ADMM in a parallel fashion (for examples of this type of reformulation, see [68]).

Note that the effectiveness of ADMM depends on the simplicity of its updates for $\boldsymbol{x}$ and $\boldsymbol{y}$ in Algorithm 2. There are other variations of ADMM that consider inexact updates for $\boldsymbol{x}$ and $\boldsymbol{y}$ in order to make the algorithm more tractable in practice (for such inexact variants of ADMM see [68, 69]).

Algorithm 2 ADMM

Input: $\boldsymbol{y}_0 \in \mathbb{R}^{n_2}$, $\boldsymbol{\lambda}_0 \in \mathbb{R}^m$, $\tau > 0$

Initialize: $k = 0$

repeat

$\boldsymbol{x}_{k+1} = \text{argmin}_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{y}_k, \boldsymbol{\lambda}_k)$

$\boldsymbol{y}_{k+1} = \text{argmin}_{\boldsymbol{y}} L(\boldsymbol{x}_{k+1}, \boldsymbol{y}, \boldsymbol{\lambda}_k)$

$\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k - \tau(\boldsymbol{A}\boldsymbol{x}_{k+1} + \boldsymbol{B}\boldsymbol{y}_{k+1} - \boldsymbol{c})$

$k = k + 1$

until Convergence

Algorithm 3 Accelerated ADMM (A2DM2)

Input: $\boldsymbol{y}_0 = \widehat{\boldsymbol{y}}_0 \in \mathbb{R}^{n_2}$, $\boldsymbol{\lambda}_0 = \widehat{\boldsymbol{\lambda}}_0 \in \mathbb{R}^m$, $\tau > 0$, $a_0 = 1$

Initialize: $k = 0$

repeat

$\boldsymbol{x}_k = \text{argmin}_{\boldsymbol{x}} L(\boldsymbol{x}, \widehat{\boldsymbol{y}}_k, \widehat{\boldsymbol{\lambda}}_k)$

$\boldsymbol{y}_k = \text{argmin}_{\boldsymbol{y}} L(\boldsymbol{x}_k, \boldsymbol{y}, \widehat{\boldsymbol{\lambda}}_k)$

$\boldsymbol{\lambda}_k = \widehat{\boldsymbol{\lambda}}_k - \tau(\boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}\boldsymbol{y}_k - \boldsymbol{c})$

$a_{k+1} = \frac{1+\sqrt{1+4a_k^2}}{2}$

$\widehat{\boldsymbol{\lambda}}_{k+1} = \boldsymbol{\lambda}_k + \frac{a_k - 1}{a_{k+1}}(\boldsymbol{\lambda}_k - \boldsymbol{\lambda}_{k-1})$

$\widehat{\boldsymbol{y}}_{k+1} = \text{argmin}_{\boldsymbol{y}} f_2(\boldsymbol{y}) + \langle \widehat{\boldsymbol{\lambda}}_{k+1}, -\boldsymbol{B}\boldsymbol{y} \rangle$

$k = k + 1$

until Convergence

The iteration complexity of ADMM has been extensively studied in the literature (see [69] and the references therein). It is shown that the algorithm has $\mathcal{O}(1/k)$ convergence rate [68, 69] under some mild conditions on the problem. Recently, some variants of this algorithm were studied which exhibit faster convergence rates while requiring only a little change in the computational effort of each iteration [70, 71]. The acceleration methods considered in these works are of the form first proposed by Nesterov [72] for gradient descent algorithms. Nesterov's accelerated gradient descent scheme in [72]

was initially designed for solving unconstrained smooth convex problems and was shown to provide a $\mathcal{O}(1/k^2)$ rate of convergence. His acceleration scheme has inspired many researchers to develop accelerated variants of other existing iterative methods (for instance see [73] which proposes an accelerated variant of the proximal splitting method).

In [71], the authors propose a Nesterov-type acceleration of ADMM for problems of the form (3.1) in the special case where both $\boldsymbol{A}$ and $\boldsymbol{B}$ are identity matrices, and one of f_1 or f_2 is differentiable. Their accelerated scheme has $\mathcal{O}(1/k^2)$ convergence rate and is based on the "symmetric" ADMM method, which differs from Algorithm 2 in that it involves two dual updates per iteration rather than one. The authors handle weakly convex problems by introducing a "step-skipping" process that applies the acceleration selectively on certain iterations. However, the step-skipping process turns the algorithm to one with a more complicated sequence of steps than conventional ADMM. The major drawback of their analysis is that it requires the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ to be identity. Such assumption restricts the application of their accelerated version of ADMM.

In a related work [70], it was shown that by applying Nesterov's acceleration scheme, ADMM can have a $\mathcal{O}(1/k^2)$ convergence rate provided that some assumptions hold true about the problem. The proposed accelerated method is simply ADMM with a predictor-corrector type acceleration step. The convergence rate of this algorithm is analyzed in [70] under the assumptions that both objective terms are strongly convex and one of them is quadratic. These assumptions enable the authors to use a similar proof technique as in [72] to show fast convergence of their algorithm. Instead of analyzing the convergence rate of the algorithm in terms of decrease in the primal objective sequence, [70] considers the dual problem of (3.1) which involves maximizing the dual function

$$\max_{\boldsymbol{\lambda}} D(\boldsymbol{\lambda}) := -f_1^*(\boldsymbol{A}^T\boldsymbol{\lambda}) - f_2^*(\boldsymbol{B}^T\boldsymbol{\lambda}) + \langle \boldsymbol{\lambda}, \boldsymbol{c}\rangle \tag{3.2}$$

where f_1^* and f_2^* are Fenchel conjugate functions [74] of f_1 and f_2, respectively, defined as

$$F^*(\boldsymbol{u}) = \max_{\boldsymbol{v}} \langle \boldsymbol{u}, \boldsymbol{v}\rangle - F(\boldsymbol{v}), \tag{3.3}$$

for any closed convex function F. In the case where f_1 and f_2 are strongly convex, the conjugate functions turn out to be smooth with Lipschitz continuous gradients and hence the dual problem (3.2) simply becomes an unconstrained smooth convex optimization. As a result, if an accelerated gradient ascent method, as the one in [72], is applied to the

dual problem (3.2), a $\mathcal{O}(1/k^2)$ convergence rate will be obtained. However, since the ADMM algorithm exploits inexact gradient ascent type of update for the dual variable $\boldsymbol{\lambda}$, a further technical condition needs to be satisfied in every iteration of the accelerated algorithm. Therefore, in order to prove the iteration complexity of accelerated ADMM, the authors in [70] require both functions f_1 and f_2 to be strongly convex as well as f_2 to be quadratic.

In this chapter, we introduce a novel algorithm called Accelerated Alternating Direction Method of Multipliers (A2DM2), and prove that the algorithm has a $O(1/k^2)$ convergence rate as long as f_1, f_2 are strongly convex. In particular, unlike [71], the functions need not be differentiable, and unlike [70], neither of them needs to be quadratic. Further, the analysis works out without any restricting assumptions on the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. The analysis technique is similar to the ones in [70, 72]. To illustrate the versatility of the proposed A2DM2, we consider the problem of learning to rank with emphasis on accuracy at the top of the list, and show how A2DM2 can be applied to the problem. Through extensive empirical evaluation on a wide variety of datasets, we illustrate that A2DM2 is competitive, often by an order of magnitude, to specialized algorithms designed for the ranking problem. We also show the generality and wide applicability of A2DM2 by highlighting other problems, including superposition models and elastic net problems, where it is readily applicable.

The rest of the chapter is organized as follows. In Section 2, we introduce the Accelerated ADMM (A2DM2) algorithm, and prove the $O(1/k^2)$ convergence rate of the algorithm. In Section 3, we present the problem of learning to rank and illustrate how A2DM2 can be applied to solve the problem. In Section 4, we present experimental results comparing A2DM2 with ADMM on synthetic data, superposition models, and elastic nets. In Section 5, we present extensive comparisons of A2DM2 on the learning to rank problem with the state-of-the-art and basic ADMM in terms of both optimization time and accuracy. We conclude in Section 6.

3.1 Accelerated ADMM Algorithm

In this section we introduce our accelerated ADMM algorithm, which we call A2DM2, for solving (3.1). First, we need to assume strong convexity of f_1, and f_2 with corresponding

constants σ_1 and σ_2, i.e. for $i = 1, 2$ and every $\boldsymbol{x}, \boldsymbol{x}^{'} \in \mathbb{R}^{n_i}$

$$f_i(\boldsymbol{x}) - f_i(\boldsymbol{x}') \geq \langle g, \boldsymbol{x} - \boldsymbol{x}' \rangle + \frac{\sigma_i^2}{2} \|\boldsymbol{x} - \boldsymbol{x}'\|^2, \ \forall \ g \in \partial f_i(\boldsymbol{x}'),$$

where $\partial f_i(\cdot)$ denotes the sub-differential set of f_i. As a result of strong convexity of f_i, $i = 1, 2$, the conjugate function, defined in (3.3), would have a Lipschitz continuous gradient with constant $1/\sigma_i$, $i = 1, 2$.

Now we are ready to define the A2DM2 algorithm. Our accelerated ADMM algorithm uses Nesterov's method to extrapolate the update of $\boldsymbol{\lambda}$ in each iteration of ADMM. In order to guarantee the convergence, it is required to update the variable $\boldsymbol{y}$ based on this extrapolated version of $\boldsymbol{\lambda}$. The overall algorithm is summarized in Algorithm 3. Compared to the conventional ADMM algorithm, our method requires extrapolating $\boldsymbol{\lambda}$ and updating $\boldsymbol{y}$ twice; therefore, each iteration for accelerated ADMM would be more costly than the usual ADMM. But as we will see in the numerical experiments, in many scenarios this extra cost is negligible compared to the speed-up that is gained using accelerated ADMM.

One appealing feature of ADMM, which makes it suitable for large-scale problems, is the capability of being executed on parallel machines. Similar to ADMM, A2DM2 is also parallelizable. This is because the primal and dual updates of A2DM2 include the ones for ADMM (steps 2-4 of Algorithm 3). The extra update in A2DM2 for the dual variable $\widehat{\boldsymbol{\lambda}}$ (step 6) is an element-wise operation, which is easy to parallelize. Moreover, the additional variable $\widehat{\boldsymbol{y}}$ is iteratively set to the minimizer of the associated objective term f_2 augmented by a linear function. This can also be done in parallel provided that f_2 is decomposable across variables.

Regarding the convergence rate of the algorithm, the following theorem states $O(1/k^2)$ convergence of A2DM2.

Theorem 1. *Suppose that f_1 and f_2 are strongly convex with σ_1, σ_2. Moreover, assume that $\tau^3 \leq \frac{\sigma_1 \sigma_2^2}{\rho(\boldsymbol{A}^T\boldsymbol{A})\rho^2(\boldsymbol{B}^T\boldsymbol{B})}$, where $\rho(\cdot)$ denotes the maximum singular value of the matrix. Then the iterates $\boldsymbol{\lambda}_k$ generated by Algorithm 2 would satisfy*

$$D(\boldsymbol{\lambda}^*) - D(\boldsymbol{\lambda}_k) \leq \frac{2\|\widehat{\boldsymbol{\lambda}}_1 - \boldsymbol{\lambda}^*\|^2}{\tau(k+2)^2}, \tag{3.4}$$

where $\boldsymbol{\lambda}^*$ is an optimal solution of the dual problem (3.2).

Proof. For an optimal Lagrange multiplier $\boldsymbol{\lambda}^*$ of (3.2), define $\boldsymbol{s}_k = a_k\boldsymbol{\lambda}_k - (a_k-1)\boldsymbol{\lambda}_{k-1} - \boldsymbol{\lambda}^*$. Then using the following lemma helps us establish Theorem 1.

Lemma 1. *For the sequence $\boldsymbol{s}_k$ defined as $\boldsymbol{s}_k = a_k\boldsymbol{\lambda}_k - (a_k-1)\boldsymbol{\lambda}_{k-1} - \boldsymbol{\lambda}^*$ the assumptions of Theorem 1 imply*

$$\|\boldsymbol{s}_{k+1}\|^2 - \|\boldsymbol{s}_k\|^2 \leq 2a_k^2\tau(D(\boldsymbol{\lambda}^*) - D(\boldsymbol{\lambda}_k)) - 2a_{k+1}^2\tau(D(\boldsymbol{\lambda}^*) - D(\boldsymbol{\lambda}_{k+1})). \tag{3.5}$$

Now using Lemma 1, it is easy to see that [2]

$$2a_{k+1}^2\tau(D(\boldsymbol{\lambda}^*) - D(\boldsymbol{\lambda}_{k+1})) \leq 2a_k^2\tau(D(\boldsymbol{\lambda}^*) - D(\boldsymbol{\lambda}_k)) + \|\boldsymbol{s}_k\|^2.$$

Rewriting (3.5) and using induction, it is easy to see that

$$\|\boldsymbol{s}_k\|^2 + 2a_k^2\tau(D(\boldsymbol{\lambda}^*) - D(\boldsymbol{\lambda}_k)) \leq \|\boldsymbol{s}_1\|^2 + 2a_1^2\tau(D(\boldsymbol{\lambda}^*) - D(\boldsymbol{\lambda}_1)), \quad \forall\ k. \tag{3.6}$$

Now in order to prove the result, we need the following lemma, for which the proof is relegated to the appendix.

Lemma 2. *When the conditions of Theorem 1 are satisfied, then for any $\boldsymbol{\gamma} \in \mathbb{R}^m$,*

$$D(\boldsymbol{\lambda}_{k+1}) - D(\boldsymbol{\gamma}) \geq \frac{1}{\tau}\langle \boldsymbol{\gamma} - \widehat{\boldsymbol{\lambda}}_{k+1}, \widehat{\boldsymbol{\lambda}}_{k+1} - \boldsymbol{\lambda}_{k+1}\rangle + \frac{1}{2\tau}\|\boldsymbol{\lambda}_{k+1} - \widehat{\boldsymbol{\lambda}}_{k+1}\|^2, \ \ \forall k. \tag{3.7}$$

Applying Lemma 2 with $k = 0$ and $\boldsymbol{\gamma} = \boldsymbol{\lambda}^*$, we get

$$\begin{aligned} D(\boldsymbol{\lambda}_1) - D(\boldsymbol{\lambda}^*) &\geq \frac{1}{\tau}\langle \boldsymbol{\gamma} - \widehat{\boldsymbol{\lambda}}_1, \widehat{\boldsymbol{\lambda}}_1 - \boldsymbol{\lambda}_1\rangle + \frac{1}{2\tau}\|\boldsymbol{\lambda}_1 - \widehat{\boldsymbol{\lambda}}_1\|^2 \\ &= \frac{1}{2\tau}\bigg(\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}^*\|^2 - \|\widehat{\boldsymbol{\lambda}}_1 - \boldsymbol{\lambda}^*\|^2\bigg). \end{aligned} \tag{3.8}$$

Combining (3.6) and (3.8) plus using definition of $\boldsymbol{s}_1 = \boldsymbol{\lambda}_1 - \boldsymbol{\lambda}^*$ yields

$$2a_k^2\tau(D(\boldsymbol{\lambda}^*) - D(\boldsymbol{\lambda}^k)) \leq \|\widehat{\boldsymbol{\lambda}}_1 - \boldsymbol{\lambda}^*\|^2.$$

Note that we have ignored the term $\|\boldsymbol{s}_k\|^2 \geq 0$ on the left hand side of (3.6). In order to get the final result, note that $a_k > a_{k-1} + \frac{1}{2} > 1 + \frac{k}{2}$. Thus,

$$D(\boldsymbol{\lambda}^*) - D(\boldsymbol{\lambda}_k) \leq \frac{2\|\widehat{\boldsymbol{\lambda}}_1 - \boldsymbol{\lambda}^*\|^2}{\tau(k+2)^2}.$$

□

[2] Lemma 1 can be proved in a similar way as Lemma 5 of [70] except that since f_2 is not restricted here to be a quadratic function, we need our Lemma 2 to complete the proof.

In the convergence analysis of ADMM, primal and dual residuals play an important role [69, 70]. For the accelerated ADMM algorithm, the primal and dual residuals can also be defined as $\boldsymbol{r}_k := \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k - \boldsymbol{B}\boldsymbol{y}_k = \boldsymbol{\lambda}_k - \widehat{\boldsymbol{\lambda}}_k$ and $\boldsymbol{d}_k := \boldsymbol{A}^T\boldsymbol{B}(\boldsymbol{y}_k - \widehat{\boldsymbol{y}}_k)$, respectively. It can be shown that for A2DM2 these residuals will decrease in $O(1/k^2)$ (The proof is similar to Lemma 6 of [70]. However, because of the different update rule for $\widehat{\boldsymbol{y}}_k$, it can be shown through Lemma (1) that f_2 is no longer needed to be quadratic.) As we will see in the next section we can use these residuals to propose another variant of the accelerated ADMM.

3.1.1 Accelerated ADMM with Restarting

Here we introduce a variant of A2DM2 to address the issue of possible spiral movements around the optimal solution, which is quite common among accelerated algorithms [75]. One common way to reduce such movements is to use a restarting rule. Similar to [70], in order to find out when the good time to restart is, we use the sum of two terms, which are proportional to the primal and dual residuals respectively,

$$m_k := \frac{1}{\tau}\|\boldsymbol{\lambda}_k - \widehat{\boldsymbol{\lambda}}_k\|^2 + \tau\|\boldsymbol{B}(\boldsymbol{y}_k - \widehat{\boldsymbol{y}}_k)\|^2. \tag{3.9}$$

At every iteration of the accelerated algorithm with restarting, which we often refer to as A2DM2+Restart, we compare m_k with m_{k-1} and if $m_k > \eta\, m_{k-1}$, where $0 < \eta < 1$ is a constant close to one, we restart the method by setting $a_{k+1} = 1$, $\widehat{\boldsymbol{y}}_{k+1} = \boldsymbol{y}_k$ and $\widehat{\boldsymbol{\lambda}}_{k+1} = \boldsymbol{\lambda}_k$.

Interestingly, our empirical studies show that while in some cases restarting really helps to improve the performance of A2DM2 (see section 3.3), in others its performance is inferior (see section 3.4).

3.2 Top Ranking Optimization

Now we focus on using the A2DM2 framework to solve the problem of ranking on the top of a list. The goal is to provide an alternative optimization solution for pushing down the highest ranked negative example in the ranked list. In this section, we appropriately reformulate the TopPush optimization problem [76], which is the most efficient pairwise ranking solution up to date, and derive its updates within the A2DM2 framework. This

results in an algorithm competitive with the TopPush algorithm of [76], in terms of both ranking performance and computational efficiency.

3.2.1 Related Work

Ranking problems are prevalent in a wide range of domains where a long list of objects, such as web links or products, needs to be ranked. Typically, in such scenarios, what matters the most is the quality of ranking near the top of the ranked list. Towards this direction, a number of learning to rank algorithms which put more emphasis towards the top of the ranked list have been developed (see [77–81] and the references therein.)

One main group of ranking algorithms aims to optimize a convex upper-bound of specific metrics which look at the top of the ranked list. Examples of such metrics are Average Precision and Normalized Discounted Cumulative Gain [78]. Another line of work was initiated by the so-called p-Norm Push ranking method [82] which optimizes a novel measure that concentrates harder on the high ranked negative examples and *pushes* them down. Extending [82], Infinite Push [83] seeks to push down the *top* irrelevant item, by minimizing the maximum number of positive examples ranked below any negative. Both [82] and [83] are pairwise ranking approaches, thus inheriting the downside of computational cost proportional to the number of positive-negative pairs, which is prohibitive for large datasets. Recently, [76] addressed this issue by reformulating the Infinite Push objective as the number of positive examples ranked below the highest ranked negative. This results in a pairwise ranking algorithm, named TopPush, with time complexity linear in the number of training instances.

3.2.2 Bipartite Ranking

Consider the bipartite setup for ranking in which the samples are either relevant (positive) or irrelevant (negative). Assume that $\mathcal{X} \subseteq \mathbb{R}^d$ is the instance space and that we are given the training sample $S = (S_+, S_-) \in \mathcal{X}^m \times \mathcal{X}^n$, where $S_+ = (\boldsymbol{x}_1^+, \cdots, \boldsymbol{x}_m^+)$ and $S_- = (\boldsymbol{x}_1^-, \cdots, \boldsymbol{x}_n^-)$ are positive and negative samples, respectively. As in [76], our goal is to learn a ranking function $f : \mathcal{X} \to \mathbb{R}$ that maximizes the number of positive instances that are ranked higher than the top-ranked negative sample. In other words,

the ranking task is translated into minimizing the following loss over the choice of f

$$\mathcal{L}(f;S) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}\left(f(\boldsymbol{x}_i^+) \leq \max_{1\leq j\leq n} f(\boldsymbol{x}_j^-)\right), \tag{3.10}$$

where $\mathbb{I}(\cdot)$ is the indicator function which equals one if the input argument is true and zero otherwise. Here, we restrict ourselves to the class of linear scoring functions, i.e. $f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x}$ for some weight vector $\boldsymbol{w} \in \mathbb{R}^d$. Since the indicator function $\mathbb{I}(\cdot)$ is not convex, [76] suggests replacing it with a convex loss function $\ell(\cdot)$ and then solves the following problem

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} \frac{\tau}{2}\|\boldsymbol{w}\|^2 + \frac{1}{m}\sum_{i=1}^{m} \ell\left(\max_{1\leq j\leq n} \boldsymbol{w}^T\boldsymbol{x}_j^- - \boldsymbol{w}^T\boldsymbol{x}_i^+\right) \tag{3.11}$$

where the regularization term $\frac{\tau}{2}\|\boldsymbol{w}\|^2$ is added to avoid over-fitting. The loss function $\ell(\cdot)$ is further assumed to be non-decreasing and differentiable. When the loss is the truncated quadratic loss, i.e. $\ell(z) = ([1+z]_+)^2$, the authors of [76] solve the dual of (3.11) by using an accelerated gradient projection algorithm. Instead, A2DM2 solves (3.11) in its primal form. Even though in the sequel we restrict ourselves to the truncated quadratic loss, our framework is general enough to incorporate any other appropriate loss function.

3.2.3 A2DM2 for Ranking

We first illustrate how ADMM can be applied to solving (3.11). Define $a := \max_j \boldsymbol{w}^T\boldsymbol{x}_j^-$, then (3.11) can be cast as

$$\begin{aligned} \min_{\boldsymbol{w}\in\mathbb{R}^d, a\in\mathbb{R}} \ & \frac{\tau}{2}\|\boldsymbol{w}\|^2 + \frac{1}{m}\sum_{i=1}^{m} \ell(a - \boldsymbol{w}^T\boldsymbol{x}_i^+) \\ \text{subject to} \ \ & a = \max_{1\leq j\leq n} \boldsymbol{w}^T\boldsymbol{x}_j^-. \end{aligned}$$

Since the above constraint is not linear in $\boldsymbol{w}$, we need to define further extra variables. Let $s_j := \boldsymbol{w}^T\boldsymbol{x}_j^- - a$, $j = 1, \cdots, n$. Note that s_j has to be non-positive since a is, by definition, the maximum of all linear combinations $\boldsymbol{w}^T\boldsymbol{x}_j^-$, for $j = 1, \cdots, n$. Moreover,

let $b_i := a - \boldsymbol{w}^T \boldsymbol{x}_i^+$, $i = 1, \cdots, m$ and then the above problem translates to

$$\begin{aligned} \min_{\boldsymbol{w}\in\mathbb{R}^d, a\in\mathbb{R}, \boldsymbol{b}\in\mathbb{R}^m} \quad & \frac{\tau}{2}\|\boldsymbol{w}\|^2 + \frac{1}{m}\sum_{i=1}^m \ell(b_i) \\ \text{subject to} \quad & s_j = \boldsymbol{w}^T \boldsymbol{x}_j^- - a,\ 1 \le j \le n \\ & b_i = a - \boldsymbol{w}^T \boldsymbol{x}_i^+,\ 1 \le i \le m \\ & s_j \le 0,\ 1 \le j \le n. \end{aligned}$$

To write the constraints compactly, we stack the negative and positive samples as rows of $\boldsymbol{X}^-$ and $\boldsymbol{X}^+$, respectively, and let $\boldsymbol{X}^- := \left[(\boldsymbol{x}_1^-)^T; (\boldsymbol{x}_2^-)^T; \ldots; (\boldsymbol{x}_n^-)^T\right] \in \mathbb{R}^{n\times d}$ and $\boldsymbol{X}^+ := \left[(\boldsymbol{x}_1^+)^T; (\boldsymbol{x}_2^+)^T; \ldots; (\boldsymbol{x}_m^+)^T\right] \in \mathbb{R}^{m\times d}$. Defining the vector $\boldsymbol{s} := [s_1, s_2, \cdots, s_n]^T$, the problem becomes

$$\begin{aligned} \min_{\substack{\boldsymbol{w}\in\mathbb{R}^d, a\in R \\ \boldsymbol{b}\in\mathbb{R}^m, \boldsymbol{s}\in\mathbb{R}^n_-}} \quad & \frac{\tau}{2}\|\boldsymbol{w}\|^2 + \frac{1}{m}\sum_{i=1}^m \ell(b_i) \\ \text{subject to} \quad & a\mathbf{1}_n + s = \boldsymbol{X}^-\boldsymbol{w} \\ & \boldsymbol{b} = a\mathbf{1}_m - \boldsymbol{X}^+\boldsymbol{w},\ \ \boldsymbol{s} \le 0 \end{aligned} \tag{3.12}$$

where $\mathbf{1}_m$ and $\mathbf{1}_n$ are all-one vectors of lengths m and n, respectively. Introducing the dual variables $\boldsymbol{\gamma}_1 \in \mathbb{R}^n$ and $\boldsymbol{\gamma}_2 \in \mathbb{R}^m$ corresponding to the first and second sets of linear constraints, we can formulate the augmented Lagrangian as

$$\begin{aligned} \mathrm{L}(\boldsymbol{w}, a, \boldsymbol{b}, \boldsymbol{s}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2) = {} & \frac{\tau}{2}\|\boldsymbol{w}\|^2 + \frac{1}{m}\sum_{i=1}^m \ell(b_i) \\ & + \boldsymbol{\gamma}_1^T(a\mathbf{1}_n + \boldsymbol{s} - \boldsymbol{X}^-\boldsymbol{w}) + \boldsymbol{\gamma}_2^T(\boldsymbol{b} - a\mathbf{1}_m + \boldsymbol{X}^+\boldsymbol{w}) \\ & + \frac{\rho}{2}\|a\mathbf{1}_n + \boldsymbol{s} - \boldsymbol{X}^-\boldsymbol{w}\|^2 + \frac{\rho}{2}\|\boldsymbol{b} - a\mathbf{1}_m + \boldsymbol{X}^+\boldsymbol{w}\|^2 \end{aligned}$$

where $\rho > 0$ is a penalty parameter. Then the ADMM algorithm will consist of the following steps. At every iteration:

Step 1. Update $(\boldsymbol{w}, a)$ according to

$$\begin{aligned} (\boldsymbol{w}, a) \leftarrow \arg\min_{(\boldsymbol{w},a)} \quad & \frac{\tau}{2}\|\boldsymbol{w}\|^2 + \boldsymbol{\gamma}_1^T(a\mathbf{1}_n - \boldsymbol{X}^-\boldsymbol{w}) + \boldsymbol{\gamma}_2^T(-a\mathbf{1}_m + \boldsymbol{X}^+\boldsymbol{w}) \\ & + \frac{\rho}{2}\|a\mathbf{1}_n + \boldsymbol{s} - \boldsymbol{X}^-\boldsymbol{w}\|^2 + \frac{\rho}{2}\|\boldsymbol{b} - a\mathbf{1}_m + \boldsymbol{X}^+\boldsymbol{w}\|^2, \end{aligned}$$

which is a convex quadratic function of $(\boldsymbol{w}, a)$. Let $\boldsymbol{A} := \begin{bmatrix} \boldsymbol{X}^{+} & -\mathbf{1}_m \\ -\boldsymbol{X}^{-} & \mathbf{1}_n \end{bmatrix} \in \mathbb{R}^{(n+m)\times(d+1)}$, then the update of $(\boldsymbol{w}, a)$ will be

$$\begin{bmatrix} \boldsymbol{w} \\ a \end{bmatrix} \leftarrow -\left(\rho \boldsymbol{A}^T \boldsymbol{A} + \begin{bmatrix} \tau \boldsymbol{I}_d & 0 \\ 0 & 0 \end{bmatrix}\right)^{-1} \boldsymbol{A}^T \begin{bmatrix} \rho \boldsymbol{b} + \boldsymbol{\gamma}_2 \\ \rho \boldsymbol{s} + \boldsymbol{\gamma}_1 \end{bmatrix}.$$

Step 2. Update $(\boldsymbol{b}, \boldsymbol{s})$ according to

$$\begin{aligned}(\boldsymbol{b}, \boldsymbol{s}) \leftarrow \arg\min_{(\boldsymbol{b},\boldsymbol{s}),\boldsymbol{s}\le 0} \; & \frac{1}{m}\sum_{i=1}^{m} \ell(b_i) + \boldsymbol{\gamma}_1^T \boldsymbol{s} + \boldsymbol{\gamma}_2^T \boldsymbol{b} \\ & + \frac{\rho}{2}\|a\mathbf{1}_n + \boldsymbol{s} - \boldsymbol{X}^{-}\boldsymbol{w}\|^2 + \frac{\rho}{2}\|\boldsymbol{b} - a\mathbf{1}_m - \boldsymbol{X}^{+}\boldsymbol{w}\|^2.\end{aligned}$$

The variable $\boldsymbol{s}$ is simply updated as $\boldsymbol{s} \leftarrow \left[(\boldsymbol{X}^{-}\boldsymbol{w} - a\mathbf{1}_n) - \frac{1}{\rho}\boldsymbol{\gamma}_1\right]_{-}$, where $[\cdot]_{-}$ stands for projection onto the negative orthant. Depending on the choice of the loss function $\ell(\cdot)$, the update of $\boldsymbol{b}$ may vary. For the case of the truncated quadratic loss, the update has closed form. More specifically, for $i = 1, 2, \cdots, m$, let $c_i := a - (\boldsymbol{X}^{+}\boldsymbol{w})_i - \frac{1}{\rho}(\boldsymbol{\gamma}_2)_i$, then

$$b_i \leftarrow \begin{cases} c_i & \text{if } c_i \le -1 \\ \frac{1}{\rho + \frac{2}{m}}\left(-\frac{2}{m} + \rho c_i\right) & \text{if } c_i > -1. \end{cases}$$

Step 3. Update $(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)$ according to

$$\boldsymbol{\gamma}_1 \leftarrow \boldsymbol{\gamma}_1 + \rho(a\mathbf{1}_n + \boldsymbol{s} - \boldsymbol{X}^{-}\boldsymbol{w}), \;\text{ and }\; \boldsymbol{\gamma}_2 \leftarrow \boldsymbol{\gamma}_2 + \rho(\boldsymbol{b} - a\mathbf{1}_m + \boldsymbol{s} + \boldsymbol{X}^{+}\boldsymbol{w}).$$

By Theorem 1, A2DM2 requires strong convexity of the objective function with respect to both (w, a) and (b, s) pairs. In order to make this condition hold, we may add extra quadratic terms to the ranking objective function in equation (3.12) and change it to

$$\frac{\tau}{2}\|\boldsymbol{w}\|^2 + \frac{1}{m}\sum_{i=1}^{m} \ell(b_i) + \frac{\tau_1}{2}a^2 + \frac{\tau_2}{2}\|\boldsymbol{s}\|^2 + \frac{\tau_2}{2}\|\boldsymbol{b}\|^2 \tag{3.13}$$

where τ_1 and τ_2 are small positive constants.

Algorithm 4 A2DM2 for Ranking

Initialize: $\boldsymbol{b} = \widehat{\boldsymbol{b}}$, $\boldsymbol{s} = \widehat{\boldsymbol{s}}$, $\boldsymbol{\gamma}_1 = \widehat{\boldsymbol{\gamma}}_1$, $\boldsymbol{\gamma}_2 = \widehat{\boldsymbol{\gamma}}_2$, τ, τ_1, τ_2, ρ, $t_0 = 1$

repeat

$$\begin{bmatrix} \boldsymbol{w} \\ a \end{bmatrix} \leftarrow -\left(\rho \boldsymbol{A}^T \boldsymbol{A} + \begin{bmatrix} \tau \boldsymbol{I}_d & 0 \\ 0 & \tau_1 \end{bmatrix}\right)^{-1} \boldsymbol{A}^T \begin{bmatrix} \rho \widehat{\boldsymbol{b}} + \widehat{\boldsymbol{\gamma}}_2 \\ \rho \widehat{\boldsymbol{s}} + \widehat{\boldsymbol{\gamma}}_1 \end{bmatrix}$$

$$\boldsymbol{s} \leftarrow \left[\tfrac{\rho}{\rho+\tau_2}(\boldsymbol{X}^{-}\boldsymbol{w} - a\mathbf{1}_n) - \tfrac{1}{\rho+\tau_2}\widehat{\boldsymbol{\gamma}}_1\right]_{-}$$

$$\boldsymbol{c} \leftarrow a\mathbf{1}_n - \boldsymbol{X}^{+}\boldsymbol{w} - \tfrac{1}{\rho}\widehat{\boldsymbol{\gamma}}_2$$

$$b_i \leftarrow \begin{cases} \frac{\rho}{\rho+\tau_2} c_i & \text{if } c_i \leq -1 \\ \frac{1}{\rho + \frac{2}{m} + \tau_2}\left(-\frac{2}{m} + \rho c_i\right) & \text{if } c_i > -1. \end{cases}$$

$$\begin{bmatrix} \boldsymbol{\gamma}_1 \\ \boldsymbol{\gamma}_2 \end{bmatrix} \leftarrow \begin{bmatrix} \widehat{\boldsymbol{\gamma}}_1 \\ \widehat{\boldsymbol{\gamma}}_2 \end{bmatrix} + \rho \begin{bmatrix} a\mathbf{1}_n + \boldsymbol{s} - \boldsymbol{X}^{-}\boldsymbol{w} \\ \boldsymbol{b} - a\mathbf{1}_m + \boldsymbol{s} + \boldsymbol{X}^{+}\boldsymbol{w} \end{bmatrix}$$

$$t_0 \leftarrow t,\ t \leftarrow \tfrac{1+\sqrt{1+4t_0^2}}{2}$$

$$\begin{bmatrix} \widehat{\boldsymbol{\gamma}}_1 \\ \widehat{\boldsymbol{\gamma}}_2 \end{bmatrix} \leftarrow \begin{bmatrix} \boldsymbol{\gamma}_1 \\ \boldsymbol{\gamma}_2 \end{bmatrix} + \tfrac{t_0 - 1}{t} \begin{bmatrix} \boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_1^0 \\ \boldsymbol{\gamma}_2 - \boldsymbol{\gamma}_2^0 \end{bmatrix}$$

$$\widehat{\boldsymbol{s}} \leftarrow \tfrac{1}{\tau_2}\left[-\widehat{\boldsymbol{\gamma}}_1\right]_{-}$$

$$b_i \leftarrow \begin{cases} \frac{-(\widehat{\boldsymbol{\gamma}}_2)_i}{\tau_2} & \text{if } \frac{-(\widehat{\boldsymbol{\gamma}}_2)_i}{\tau_2} \leq -1 \\ \frac{(\widehat{\boldsymbol{\gamma}}_2)_i + \frac{2}{m}}{\tau_2 + \frac{2}{m}} & \text{otherwise.} \end{cases}$$

until convergence

Although adding these terms will slightly change the problem, our experimental results show the ranking performance of the algorithm is not worse than the existing state-of-the-art approaches. We avoid deriving the iterative updates of A2DM2 here as they are quite similar to ADMM and instead summarize them in Algorithm 4.

3.2.4 Computational Complexity

Updating the pair $(\boldsymbol{w}, a)$ requires $\mathcal{O}(d(m+n))$ operations provided that the matrix $\left(\rho \boldsymbol{A}^T \boldsymbol{A} + \begin{bmatrix} \tau \boldsymbol{I}_d & 0 \\ 0 & 0 \end{bmatrix}\right)^{-1} \boldsymbol{A}^T$ is computed before executing the algorithm and saved

in memory. The computational cost of updating $(\boldsymbol{b}, \boldsymbol{s})$ and $(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)$ is of the same order $\mathcal{O}(d(m+n))$. So, to achieve an ϵ-accuracy solution, the ADMM and A2DM2 require $\mathcal{O}(d(m+n)/\epsilon)$ and $\mathcal{O}(d(m+n)/\sqrt{\epsilon})$ operations, respectively. Therefore, in terms of the computational complexity, A2DM2 is comparable with the state-of-the-art in [76].

3.3 ADMM for Superposition Models

In this section we demonstrate the effectiveness of our method for solving an extensively studied family of statistical problems that are often known as "superposition models" or "dirty models" in the machine learning literature [13, 84].

Dirty Models. In many high dimensional statistical problems, the number of observations is far less than the dimension of the model to be estimated. Without any prior knowledge, the true model is not identifiable. Fortunately, in many practical applications, the model is known to have a low dimensional structure that can be used to resolve the identifiability issue. This prior knowledge can be exploited by adding to the objective function some appropriate convex regularizers which capture the structure of the model. Formally speaking, assume that given the linear observations $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{y}$, where $\boldsymbol{A}$, $\boldsymbol{B}$ are known matrices, the goal is to estimate $\boldsymbol{x}$ and $\boldsymbol{y}$. In addition, assume that $R_1(\cdot)$ and $R_2(\cdot)$ are the convex penalty functions which encode the prior knowledge of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. Then, the estimation problem can be formulated as follows

$$\min_{\boldsymbol{x},\boldsymbol{y}} \; R_1(\boldsymbol{x}) + R_2(\boldsymbol{y}) \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{y} = \boldsymbol{b}. \tag{3.14}$$

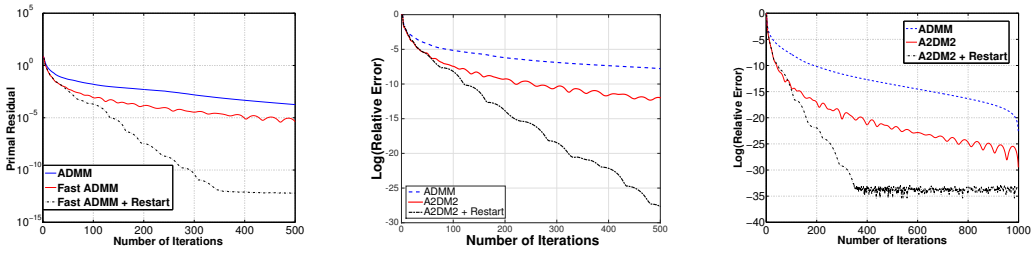
Many famous formulations can be interpreted by this model. For example, by specializing $R_2(\boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{y}\|^2$ (the ℓ_2-norm squared), $R_1(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$ (the ℓ_1-norm), $\boldsymbol{A}$ as the design matrix and $\boldsymbol{B}$ as the identity matrix, we obtain the Lasso formulation [11]. When working with real-world data, in order to increase the robustness of the estimation procedure, the elastic net regularizer, given by $R_1(\boldsymbol{x}) = \|\boldsymbol{x}\|_1 + \frac{1}{2}\|\boldsymbol{x}\|^2$, can be used instead of the ℓ_1-norm penalty [85].

Dirty Model with Elastic Net Regularizer. Another interesting special-case of problem (3.14) is where R_1 and R_2 are both elastic net regularization functions. Given the linear observation model $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{y}$, such a problem can be written as

$$\min_{\boldsymbol{x},\boldsymbol{y}} F(\boldsymbol{x},\boldsymbol{y}) := \|\boldsymbol{x}\|_1 + \frac{1}{2}\|\boldsymbol{x}\|^2 + \mu(\|\boldsymbol{y}\|_1 + \frac{1}{2}\|\boldsymbol{y}\|^2) \quad (3.15)$$
$$\text{subject to } \boldsymbol{A}\boldsymbol{x} + \boldsymbol{y} = \boldsymbol{b}.$$

where $\mu > 0$ is a constant. Minimizing the ℓ_1-norm of the variables here is to exploit the prior knowledge about the sparsity of such unknowns [86,87]. This problem has the same form as (3.1) with $f_1(\boldsymbol{x}) = \|\boldsymbol{x}\|_1 + \frac{1}{2}\|\boldsymbol{x}\|^2$ and $f_2(\boldsymbol{y}) = \mu(\|\boldsymbol{y}\|_1 + \frac{1}{2}\|\boldsymbol{y}\|^2)$. Note that the objective function components f_1 and f_2 are strongly convex in terms of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. Therefore, our analysis guarantees the fast convergence of A2DM2 for this problem. The application of A2DM2 to this problem is quite straight-forward. However, we omit the details of the variable updates here due to space limits. In the next section,we describe the results of our numerical tests with this problem.

Numerical Experiments. To test the performance of our accelerated method compared to ADMM, we carry out a set of simulations. We randomly generate the observation matrix $A \in \mathbb{R}^{m\times n}$ of size $m = 2^8$ and $n = 2^9$ from a standard Gaussian distribution. The true target vector $\boldsymbol{x}$ is sparse with only five percent of its entries being non-zero. The non-zero entries are standard Gaussian. The error vector $\boldsymbol{e} \in \mathbb{R}^m$ is generated from an exponential distribution with average 0.01. Figure 3.1 shows the convergence behavior of ADMM, A2DM2, and A2DM2 + Restart for solving problem (3.15). Figure 3.1(a) plots the primal residual sequence $r_k = \|\boldsymbol{A}\boldsymbol{x}_k - \boldsymbol{b} - \boldsymbol{y}_k\|$, Figure 3.1(b) shows the primal objective optimality gap $F(\boldsymbol{x}_k,\boldsymbol{y}_k) - F(\boldsymbol{x}^*,\boldsymbol{y}^*)$ and Figure 3.1(c) shows the dual objective optimality gap $D(\boldsymbol{\lambda}^*) - D(\boldsymbol{\lambda}_k)$. Both of the objective sequences are normalized with their initial values, i.e. with their values at iteration $k = 1$. As the three figures suggest, A2DM2 performs better than ADMM in terms of all three measures. The best performance is observed for A2DM2+Restart. Note that the improved performance is obtained at the cost of defining the auxiliary primal and dual variables $\widehat{\boldsymbol{y}}$ and $\widehat{\boldsymbol{\lambda}}$ whose updates can be done in $\mathcal{O}(m)$ (the details are omitted for the sake of brevity).



(a) (b) (c)

Figure 3.1: Convergence behavior of ADMM, A2DM2 and A2DM2 + Restart for the Elastic net with ℓ_1 regularization problem. A2DM2 performs better than the ADMM. ADM2+Restart has the best performance.

3.4 Experiments on Top Ranking

3.4.1 Settings

Datasets. To evaluate the performance of the proposed A2DM2 algorithm on the problem of learning to rank, we conduct a set of experiments on various datasets. The left column of Table 3.1 summarizes the datasets used in our experiments. All datasets used are publicly available binary classification datasets[3] having varying sizes and coming from different domains.

Some datasets come from the medical domain (`breast-cancer, diabetes`), ecology (`covtype`), biology (`cod-rna, splice`), others from email spam filtering (`spambase`), web data (`w8a`), census data (`a9a`), and credit card approval (`australian`). Also, competition data on generalization ability and text decoding (`ijcnn1`) were used. The `epsilon` dataset is an artificial data set from the Pascal large scale learning challenge 2008.

Setup & Parameters. On each dataset, we run experiments for ten trials and report the averaged results over those trials. In each run, the dataset is randomly divided into two subsets: 2/3 for training and 1/3 for test. For all algorithms, we set the precision parameter ϵ to 10^{-4}, choose other parameters by 3-fold cross validation (based on the average value of Pos@Top) on training set, and perform the evaluation on

[3] http://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/

the test set. In particular, the regularization parameter τ is chosen from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}\}$ based on cross-validation on the TopPush algorithm. The parameters τ_1, τ_2 were set to the value 0.01. The step size ρ of the proximal operator in the ADMM-based algorithms was cross validated as followed: For ADMM ρ was chosen from the set $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$, for A2DM2 from $\{10^{-4}, 10^{-3}, 10^{-2}\}$ and for A2DM2+Restart from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$.

Methods. We compared the ADMM, the proposed A2DM2 and A2DM2+Restart frameworks with TopPush [76], which is the state-of-the-art top rank algorithm. We implemented the ADMM-based algorithms in MATLAB and used the publicly available source code for TopPush.

Top-Ranking Metric. Since the objective of TopPush, and therefore of all compared approaches, is to push down the top ranked negative example in the ranked list, a natural performance measure is the number of positives ranked on top of the first negative example (Pos@Top) [83]. Larger values in this metric imply better top ranking performance.

Training Efficiency. In order to evaluate the computational efficiency of our proposed algorithms, we also report the average time (in seconds) it takes for the algorithms to be trained. For this experiment, we set the parameters of the different algorithms to be the best ones selected by cross validation and we run them on the training set. We do so for the ten different random runs and average out the training time. Regarding the stopping criterion, all three algorithms i.e., ADMM, A2DM2 and A2DM2+Restart, are stopped when the iteration number is greater than 10 and the sum of the primal and dual residuals is less than ϵ. The TopPush stopping criterion is kept in its original form as given in the source code[4] , i.e., when the iteration number is greater than 10 and the relative dual objective gap of the TopPush algorithm is less than ϵ.

[4] http://lamda.nju.edu.cn/code_TopPush.ashx

Data	Algorithm	Pos@top	Time (sec.)
brst-cncr	TopPush	$4.90\times10^{1} \pm 1.39\times10^{1}$	$5.58\times10^{-1} \pm 8.01\times10^{-1}$★
239 / 444	ADMM	$5.33\times10^{1} \pm 1.38\times10^{1}$	$1.74\times10^{0} \pm 1.46\times10^{0}$
d:10	A2DM2	$5.25\times10^{1} \pm 1.48\times10^{1}$	$9.24\times10^{-1} \pm 1.08\times10^{0}$★
	A2DM2+R	$5.23\times10^{1} \pm 1.50\times10^{1}$	$3.15\times10^{0} \pm 3.53\times10^{0}$
australian	TopPush	$1.13\times10^{1} \pm 5.14\times10^{0}$	$2.22\times10^{-1} \pm 2.00\times10^{-1}$★
307 / 383	ADMM	$1.62\times10^{1} \pm 7.18\times10^{0}$	$5.91\times10^{-1} \pm 4.79\times10^{-1}$★
d:14	A2DM2	$1.77\times10^{1} \pm 1.08\times10^{1}$	$5.12\times10^{-1} \pm 5.54\times10^{-1}$★
	A2DM2+R	$1.71\times10^{1} \pm 6.76\times10^{0}$	$1.34\times10^{0} \pm 2.05\times10^{0}$
diabetes	TopPush	$1.41\times10^{1} \pm 2.14\times10^{1}$	$4.66\times10^{-2} \pm 9.08\times10^{-2}$★
500 / 268	ADMM	$2.36\times10^{1} \pm 2.03\times10^{1}$	$4.58\times10^{-1} \pm 2.51\times10^{-1}$
d:34	A2DM2	$3.19\times10^{1} \pm 1.73\times10^{1}$	$5.03\times10^{-1} \pm 2.41\times10^{-1}$
	A2DM2+R	$1.87\times10^{1} \pm 1.89\times10^{1}$	$6.17\times10^{-1} \pm 4.60\times10^{-1}$
spambase	TopPush	$5.49\times10^{1} \pm 8.29\times10^{1}$	$1.63\times10^{1} \pm 9.74\times10^{0}$
1,813 / 2,788	ADMM	$4.48\times10^{1} \pm 3.86\times10^{1}$	$2.06\times10^{1} \pm 1.59\times10^{1}$
d:57	A2DM2	$5.02\times10^{1} \pm 3.64\times10^{1}$	$3.35\times10^{0} \pm 1.23\times10^{0}$★
	A2DM2+R	$5.48\times10^{1} \pm 3.55\times10^{1}$	$1.51\times10^{1} \pm 8.07\times10^{0}$
splice	TopPush	$8.78\times10^{1} \pm 4.85\times10^{1}$	$1.86\times10^{0} \pm 2.58\times10^{0}$★
1,648 / 1,527	ADMM	$9.99\times10^{1} \pm 2.22\times10^{1}$	$7.29\times10^{0} \pm 5.02\times10^{0}$★
d:60	A2DM2	$1.15\times10^{2} \pm 2.63\times10^{1}$•	$3.18\times10^{0} \pm 9.78\times10^{-1}$★
	A2DM2+R	$1.14\times10^{2} \pm 2.70\times10^{1}$•	$1.43\times10^{1} \pm 2.13\times10^{0}$
ijcnn1	TopPush	$6.08\times10^{1} \pm 2.06\times10^{1}$	$7.39\times10^{0} \pm 1.26\times10^{1}$★
4,853 / 45,137	ADMM	$1.24\times10^{2} \pm 3.76\times10^{1}$•	$1.82\times10^{2} \pm 6.77\times10^{1}$
d:22	A2DM2	$5.56\times10^{1} \pm 9.90\times10^{0}$	$1.50\times10^{0} \pm 2.32\times10^{0}$★
	A2DM2+R	$7.34\times10^{1} \pm 3.20\times10^{1}$	$1.13\times10^{2} \pm 1.39\times10^{2}$
a9a	TopPush	$1.78\times10^{1} \pm 1.30\times10^{1}$	$8.50\times10^{-1} \pm 1.85\times10^{-1}$★
11,687 / 37,155	ADMM	$4.57\times10^{1} \pm 2.19\times10^{1}$	$1.36\times10^{1} \pm 8.59\times10^{0}$
d:122	A2DM2	$5.47\times10^{1} \pm 2.89\times10^{1}$	$1.44\times10^{1} \pm 6.74\times10^{0}$
	A2DM2+R	$5.07\times10^{1} \pm 2.43\times10^{1}$	$5.29\times10^{1} \pm 3.43\times10^{1}$
w8a	TopPush	$1.39\times10^{2} \pm 3.42\times10^{1}$	$2.20\times10^{1} \pm 1.33\times10^{1}$★
1,933 / 62,767	ADMM	$1.37\times10^{2} \pm 4.19\times10^{1}$	$1.71\times10^{2} \pm 6.28\times10^{1}$
d:300	A2DM2	$1.38 \times 10^{2}\pm 4.98\times10^{1}$	$5.98\times10^{1} \pm 2.58\times10^{1}$★

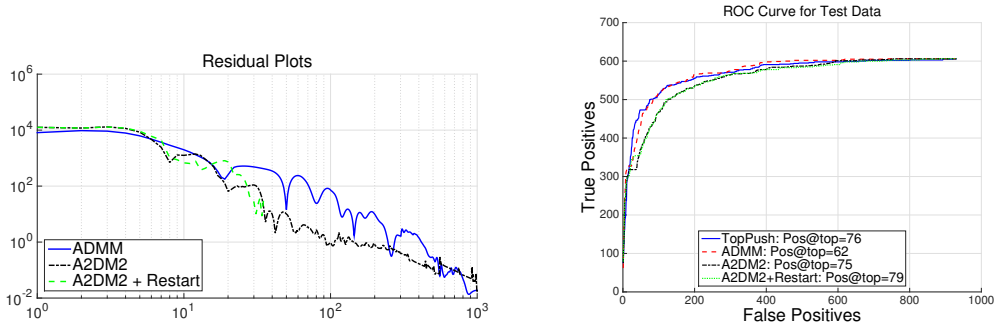
	A2DM2+R	$1.44\times10^2 \pm 3.86\times10^1$	$2.25\times10^2 \pm 7.54\times10^1$
`covtype`	TopPush	$7.97\times10^2 \pm 1.52\times10^2\bullet$	$1.78 \times10^1 \pm 4.26 \times 10^0$
283,301 / 297,711	A2DM2	$2.63 \times 10^1 \pm 2.48 \times 10^1$	$2.02 \times 10^1 \pm 2.81 \times 10^0$
d: 54			
`cod-rna`	TopPush	$1.97\times10^2 \pm 9.95\times10^1$	$8.34 \times10^2 \pm 4.59 \times 10^2$
162,855 / 325,710	ADMM	$1.03\times10^2 \pm 1.36\times10^2$	$6.50\times10^2 \pm 7.33 \times 10^2$
d:8	A2DM2	$2.24 \times 10^2 \pm 8.91 \times 10^1$	$2.01 \times 10^0 \pm 5.52 \times 10^{-1}\bigstar$
	A2DM2+R	$1.27 \times 10^2 \pm 8.38 \times 10^1$	$1.71 \times 10^2 \pm 4.75 \times 10^2$
`epsilon`	TopPush	$1.82 \times 10^3 \pm 3.07 \times 10^2$	$5.78 \times 10^2 \pm 1.85 \times 10^2$
249,778 / 250,222	A2DM2	$2.07 \times 10^3 \pm 4.16 \times 10^2$	$4.16 \times 10^2 \pm 1.79 \times 10^1$
d:2,000			

Table 3.1: Data statistics (left column) and experimental results. The mean and standard deviation of the training time (sec) and the Pos@Top over ten random splits of training-test sets are reported. For each dataset, the number of positive and negative instances is below the data name as m/n, together with dimensionality d. For training time comparison, one or more algorithms are marked as ★ if they are at least an order of magnitude faster compared to the remainings. For top-ranking performance (Pos@Top) comparison, the entries marked with • are those for which the number of positives at top is at least 10 times greater than the Pos@Top achieved by the rest of the algorithms. In most datasets, one can observe that A2DM2 is very competitive with TopPush in terms of the order of magnitude for both top-ranking accuracy and training time.

3.4.2 Results: Running Time Comparison

In the right-most column of Table 3.1 we report the training performance (in seconds) of the algorithms A2DM2, A2DM2 + Restart, compared to the TopPush algorithm and the standard ADMM. One can observe that in most datasets A2DM2 matches the training time of TopPush, and can be even one order of magnitude faster (`spambase`). In the cases where A2DM2 is slower than TopPush, it achieves better ranking performance (`diabetes, a9a`). In general, there is a tradeoff between accuracy at the top and time

for convergence. A2DM2 usually manages to balance the tradeoff and achieves good Pos@Top at time comparable (or better) with TopPush, while ADMM often does not.



(a) (b)

Figure 3.2: Study on `spambase` dataset. (a) Residuals decay faster for the accelerated variants of ADMM compared to ADMM. (b) ROC Curve for test data: One can observe similar top ranking performance for the four approaches.

3.4.3 Results: Top Ranking Accuracy

In addition, in Table 3.1 we report the performance of the compared approaches in terms of the average Pos@Top. The A2DM2 algorithm almost always matches the ranking performance of TopPush and in most datasets it results in slightly better results.

From the results of Table 3.1, we observe that as the size of the datasets increases at the bottom of the table, the acceleration that A2DM2 provides compared to ADMM becomes more considerable. For instance, for the cod-rna dataset, the value of Pos@Top for A2DM2 is around twice that for ADMM and yet A2DM2 is, in average, two orders of magnitude (100 times) faster than ADMM. Also, the results of ADMM for the two larger datasets, epsilon and covtype, are missing from Table 1 since the cross-validation study for ADMM was time consuming.

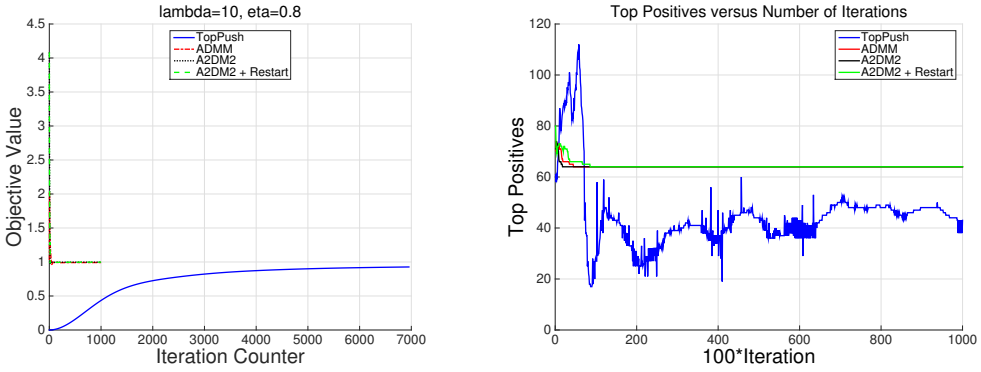
For the `spambase` dataset, for a single random training-test split, Figure 3.2 (b) shows the Receiver Operating Characteristic (ROC) curves of the four compared algorithms. Since we focus on the ranking model where accuracy at the top is critical, good performance in the left-most part of the ROC curve is necessary. In this regard, one can see similar ranking performance of the compared approaches. In fact, A2DM2 + Restart

achieves slightly higher Pos@Top performance followed by TopPush and A2DM2 and finally ADMM. Figure 3.2 (a) shows how the sum of the primal and dual residuals behaves vs. the number of training iterations. As the figure implies, the residuals decay faster for the accelerated variants of ADMM.

3.4.4 Effect of number of training iterations

In this subsection, we study the ranking performance of A2DM2 versus the number of iterations through some figures and compare it with TopPush, A2DM2 + Restart and ADMM. The shown plots are just for one random training-test split of the `spambase` dataset. However, we observed same trends as what follows with the other datasets and with different training-test splits.

Fixing the regularization parameter $\tau = 10$ and $\eta = 0.8$, Figure 3.3 (a) shows how the value of the objective evolves as the number of iterations grows. One can observe that ADMM, A2DM2 and A2DM2+Restart converge in only a few hundred iterations. In contrast, TopPush needs to run for a few thousands of iterations to reach the optimal objective value. In Figure 3.3 (b), we present how the number of Pos@Top in the test set evolves after every 100 iterations of the training phase. One interesting observation is that A2DM2 converges to its final test top-ranking performance after few hundreds of iteration, whereas TopPush does not seem to achieve a stable number of Pos@Top.



(a) (b)

Figure 3.3: Study on `spambase` dataset. (a) Objective value versus number of iterations (b) Top Ranking performance (Pos@Top) on the test set after every 100 iterations of the training phase. A2DM2 converges to its final ranking performance after few hundreds of iteration, whereas TopPush does not seem to achieve a stable number of Pos@Top.

3.5 Conclusion

In this chapter, we propose an Accelerated Alternating Direction Method of Multipliers, named A2DM2. We prove that it has $O(1/k^2)$ convergence rate when the objective terms are strongly convex. This guarantees a faster convergence rate compared to ADMM [68]. A large number of real world machine learning problems formulated under the ADMM framework can benefit from this improvement. We illustrate the applicability of A2DM2 on the problem of learning to rank, and show that it is competitive with the state-of-the-art TopPush algorithm [76] both in terms of ranking accuracy at the top and training efficiency.

Chapter 4

Estimation of Low-Rank Matrices via Regularized Factorization

The problem of estimating a high-dimensional (nearly) low-rank matrix from a small collection of acquired measurements has been extensively studied in the past two decades. Numerous instances of this general problem emerge in different fields across science and engineering. The matrix completion problem, appearing in applications related to recommendation systems [88–92], various forms of the principal component analysis [93–95], network monitoring [96], and phase retrieval problem [97–99] are just a few instances of this general framework.

Various optimization-based approaches have been proposed and analyzed recently for low-rank matrix estimation problems. Popular examples of such approaches involve minimizing convex regularized cost functions, where the minimization of cost function ensures the fit of the estimation model to acquired measurements and the presence of the regularization function is supposed to promote low-rank structure of the solution. Initiated after the pioneering works of [9, 100], incorporating nuclear norm regularization to yield low-rank matrix estimations has been recognized as a primary approach to obtain efficient convex optimization procedures that, under certain conditions, will verifiably obtain suitable estimates of the ground-truth low-rank matrix [101]. However, such convex methods can often demand high computation and memory resources, in particular when applied to contemporary high-dimensional problems.

An alternative line of efforts has been dedicated to studying more practical methods which do not utilize nuclear norm minimization to impose low-rank structure of the estimated matrix. Alternatively, and inspired by the works of Burer and Monteiro [102–104], various recent studies involve expressing the desired low-rank matrix as the product of two low-dimensional factor matrices whose common dimension is specified by the rank (or an approximation thereof) of the ground-truth matrix. In the case where the desired low-rank matrix is known to be symmetric (and real), this methodology represents the matrix as the product of a factor matrix and its transpose.

Having expressed the original low-rank matrix with its low-dimensional factor(s), a numerical procedure is then utilized for minimizing a properly-selected cost function with respect to the factor(s). Since the rank of the unknown matrix is typically smaller than its ambient dimensions, such a parametrization may significantly reduce the search space of the optimization problem, yielding potentially significant computational and memory utilization improvements. However, formulating low-rank matrix estimation as an optimization problem over its factor(s) will convert the estimation procedure into a non-convex problem, which will generally lack the theoretical recovery guarantees enjoyed for convex methodologies. Therefore, studying the behavior of potential spurious minima of such non-convex approaches has become a critical question in this domain and the motivation for several rigorous studies done recently [105–108].

A recent branch of works has been focused on designing and analyzing efficient numerical procedures for factorization-based formulations. Due to the non-convex nature of such problems, the main theme of these studies has been on analyzing the local convergence behavior of proposed algorithms, along with suggestions on procedures that could provide good initialization points to those methods. Alternating minimization [109–111], power methods [112], EM methods [113, 114], and variants of gradient descent algorithm [17,18,115–122] are among the simplest numerical procedures that are analyzed in this manner when applied to various instances of low-rank matrix estimation problem.

Among the studies that advocate the use of gradient descent method for the factorization of low-rank matrices, some have been dedicated to certain instances of this problem. Matrix sensing [116], matrix completion [117], phase retrieval [98] and robust

principal component analysis [121] are among the prominent instances of low-rank matrix estimation problem that have benefited from such specialized studies. The common theme in those works is studying the application of gradient descent method for minimizing a properly selected loss function with respect to the factor matrix. Another line of works, initiated by [17, 18, 118], has attempted to go beyond such problem instances and study the gradient descent method for more general factorization problems of low-rank matrices, allowing for general convex cost functions and representing the unknown matrix as a product of factors. Moreover, the comprehensive study in [17] includes both computational and statistical aspects of variants of gradient descent algorithm for the task of matrix factorization. In particular, that work is able to incorporate further known structure of the ground-truth matrix (beside being low-rank) into the problem by imposing constraints on the problem that capture such structure.

However, missing from the large and growing collection of existing studies is a study of formulations that promote special structure of the factor matrix via augmenting the cost function of matrix factorization with suitable regularization functions. In this chapter of the thesis, we introduce and rigorously study numerical algorithms that can be applied to solve instances of matrix factorization problem cast as minimizing a general cost function plus suitable regularization function, with respect to a factor matrix. Our regularized framework includes many non-regularized or constrained formulations as special cases, so that by proper specialization of our results, we are able to recover also many existing results. Under now-standard assumptions on the cost function and the regularization function, we demonstrate similar local convergence guarantees that are comparable to those provided for non-regularized factorization frameworks. Similar to [17, 122] we investigate both computational and statistical aspects, and provide experimental evidence for the efficiency of our numerical procedure.

Notation. Vectors and matrices are denoted by bold-face lowercase and uppercase letters, respectively. The i-th row and j-th column of a matrix $\boldsymbol{X}$ are represented by $\boldsymbol{X}_{i*}$ and $\boldsymbol{X}_{*j}$, respectively. The Euclidean norm of any vector $\boldsymbol{v}$ is denoted by $\|\boldsymbol{v}\|_2$. For any arbitrary matrix $\boldsymbol{X}$, we use $\|\boldsymbol{X}\|_2$ and $\|\boldsymbol{X}\|_F$ to denote its spectral and Frobenius norms, respectively. Moreover, given a positive number $p \geq 1$, $\|\boldsymbol{X}\|_{2,p}$ will denote the row-wise $\ell_{2,p}$ norm of $\boldsymbol{X}$, which is defined as $\|\boldsymbol{X}\|_{2,p} = (\sum_{i=1}^{p} \|\boldsymbol{X}_{i*}\|_2^p)^{1/p}$. Finally, assuming that $\boldsymbol{X}$ is a rank-r matrix, $\sigma_1(\boldsymbol{X})$ and $\sigma_r(\boldsymbol{X})$ will denote its maximum and

minimum singular values, respectively.

4.1 Problem Formulation

Inspired by the setup of [17, 123, 124], we assume that a collection of n samples $Z_1^n := \{Z_1, Z_2, \cdots, Z_n\}$ drawn from a marginal distribution $\mathbb{P}$ over a space $\mathcal{Z}$ is available. We consider an *empirical* loss function $\mathcal{L}_n : \mathbb{R}^{d_1 \times d_2} \times \mathcal{Z}^n \to \mathbb{R}$, where the value of $\mathcal{L}_n(\boldsymbol{X}, Z_1^n)$ quantifies a measure of the fit between an unknown parameter matrix $\boldsymbol{X} \in \mathbb{R}^{d_1 \times d_2}$ and the measurement data. This empirical loss function could be viewed as a surrogate to the *population* loss function $\mathcal{L} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}$ given as

$$\mathcal{L}(\boldsymbol{X}) := \mathbb{E}\left[\mathcal{L}_n(\boldsymbol{X}, Z_1^n)\right],$$

where the expectation is with respect to the distribution of all the n data points. Depending on the structure of the ground-truth matrix, which is defined as

$$\boldsymbol{X}^* := \underset{\boldsymbol{X} \in \mathbb{R}^{d_1 \times d_2}}{\operatorname{argmin}} \ \mathcal{L}(\boldsymbol{X}), \tag{4.1}$$

two cases are possible, which are studied in the following two subsections.

4.1.1 Case of PSD Matrices

In the first case, we assume that the ground-truth parameter matrix $\boldsymbol{X}^*$ is known to be symmetric positive semi-definite (PSD) and low-rank, with $\operatorname{rank}(\boldsymbol{X}^*) = r \ll d$, where $d := d_1 = d_2$ in this case. Under such assumptions, we follow the parameterization popularized by Burer and Monteiro [103, 104] to express the true parameter matrix as $\boldsymbol{X}^* = \boldsymbol{U}^* \boldsymbol{U}^{*T}$, where $\boldsymbol{U}^* \in \mathbb{R}^{d \times r}$ is the associated rank-r factor matrix. Then the following simple estimator can be utilized to estimate the desired factor matrix (and hence the true low-rank matrix)

$$\widehat{\boldsymbol{U}} \in \underset{\boldsymbol{U} \in \mathbb{R}^{d \times r}}{\operatorname{argmin}} \quad \mathcal{L}_n(\boldsymbol{U}\boldsymbol{U}^T, Z_1^n).$$

Moreover, if the ground-truth matrix $\boldsymbol{X}^*$ is known to meet further structural restrictions (beside being low-rank) and such restrictions can be captured via proper regularization

functions on the factor matrix $\boldsymbol{U} \in \mathbb{R}^{d\times r}$, then we propose to improve the above simple estimator by incorporating regularization terms into the estimation procedure as follows

$$\widehat{\boldsymbol{U}} \in \underset{\boldsymbol{U}\in\mathbb{R}^{d\times r}}{\operatorname{argmin}} \qquad \mathcal{L}_n(\boldsymbol{U}\boldsymbol{U}^T, Z_1^n) + \tau\,\omega(\boldsymbol{U}), \tag{4.2}$$

where $\omega(\cdot) : \mathbb{R}^{d\times r} \to \mathbb{R}$ is a regularization function that is supposed to promote "low-complexity" structure of $\boldsymbol{U}^*$ and $\tau > 0$ is a regularization constant. Notice that the above regularized estimation framework is somewhat more general than the analogous constrained ones studied in [17, 120, 122], since they can be obtained as special cases of our scenario upon setting the regularizer to be the indicator function of a corresponding, appropriately specified set.

To solve the regularized optimization problem in (4.5), we propose to use the proximal descent algorithm [125, 126], whose iterates are updated by the following rule:

$$\boldsymbol{U}_{t+1} = \operatorname{prox}_{\omega}(\widetilde{\boldsymbol{U}}_{t+1}; \tau\mu_t), \quad \text{where} \quad \widetilde{\boldsymbol{U}}_{t+1} = \boldsymbol{U}_t - \mu_t \nabla\mathcal{L}_n(\boldsymbol{U}_t\boldsymbol{U}_t^T)\boldsymbol{U}_t, \tag{4.3}$$

where $\operatorname{prox}_{\omega}(\boldsymbol{U};\alpha)$ denotes the proximal operator, associated with the function $\omega(\cdot)$, which is defined as

$$\operatorname{prox}_{\omega}(\boldsymbol{U};\alpha) := \underset{\check{\boldsymbol{U}}\in\mathbb{R}^{d\times r}}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{U} - \check{\boldsymbol{U}}\|_F^2 + \alpha\,\omega(\check{\boldsymbol{U}}), \tag{4.4}$$

and μ_t denotes the step-size parameter at iteration $t \geq 0$. Notice that the exact expression for the gradient of $\mathcal{L}_n(\boldsymbol{U}\boldsymbol{U}^T)$ with respect to $\boldsymbol{U}$ is $\big(\nabla\mathcal{L}_n(\boldsymbol{U}\boldsymbol{U}^T) + \nabla^T\mathcal{L}_n(\boldsymbol{U}\boldsymbol{U}^T)\big)\,\boldsymbol{U}$. However, assuming the symmetry of the loss function $\mathcal{L}_n(\cdot)$ with respect to its argument (as in [17, 116, 118] among others), we would then obtain a simpler expression for the gradient as $2\nabla\mathcal{L}_n(\boldsymbol{U}\boldsymbol{U}^T)\boldsymbol{U}$, which is utilized in the above update rule expression (after absorbing the constant 2 into the step-size parameter). We define $\boldsymbol{X}_{t+1} := \boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^T$, $\boldsymbol{X}_t := \boldsymbol{U}_t\boldsymbol{U}_t^T$, and $\widetilde{\boldsymbol{X}_t} := \widetilde{\boldsymbol{U}}_t\widetilde{\boldsymbol{U}}_t^T$ as the low-rank matrices associated with the above iteration . The overall scheme of the proposed algorithm is summarized in Algorithm 5. Here, we are proposing to employ a constant step-size value μ, which is directly obtained from applying Theorem 4.3.2 of section 4.3.

Algorithm 5 Proximal Descent Method for Factorizing PSD Matrices

Input: initial factor $\boldsymbol{U}_0$, target rank r, number of iterations T

Set $t = 0$, $\mu = 1/8M\sigma_1^2(\boldsymbol{U}_0)$

repeat

$\widetilde{\boldsymbol{U}}_{t+1} = \boldsymbol{U}_t - \mu\nabla\mathcal{L}_n(\boldsymbol{X}_t)\boldsymbol{U}_t$

$\boldsymbol{U}_{t+1} = \text{prox}_{\omega}(\widetilde{\boldsymbol{U}}_{t+1}; \mu\tau)$

$t = t + 1$

until $t = T$

Note that the proximal descent method can be viewed as a generalization of the standard gradient descent or projected gradient descent methods to estimation problems involving general convex regularizers. In particular, setting $\omega(\cdot)$ to be the zero function would reduce the above iterations to simple gradient descent iterations. Moreover, specializing $\omega(\cdot)$ to be the indicator function of a convex set would turn the iterations to those of the projected gradient descent algorithm.

4.1.2 General Case of Non-PSD Matrices

In the general case where $\boldsymbol{X}^* \in \mathbb{R}^{d_1 \times d_2}$ is non-symmetric PSD, we again pursue the Burer and Monteiro parameterization [103,104] to express the true parameter matrix as $\boldsymbol{X}^* = \boldsymbol{U}^*\boldsymbol{V}^{*T}$ this time, where $\boldsymbol{U}^* \in \mathbb{R}^{d_1 \times r}$ and $\boldsymbol{V}^* \in \mathbb{R}^{d_2 \times r}$ are the associated rank-r factor matrices. Then, a modified version of the empirical loss minimization approach could be utilized to estimate the desired factor matrices

$$(\widehat{\boldsymbol{U}}, \widehat{\boldsymbol{V}}) \in \underset{\substack{\boldsymbol{U}\in\mathbb{R}^{d_1\times r}\\ \boldsymbol{V}\in\mathbb{R}^{d_2\times r}}}{\text{argmin}} \quad \mathcal{L}_n(\boldsymbol{U}\boldsymbol{V}^T, Z_1^n) + \lambda\, g(\boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{V}^T\boldsymbol{V}),$$

where, as in [18,116], the addition of the second term $\lambda \cdot g(\boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{V}^T\boldsymbol{V})$, with $\lambda > 0$, to the objective function is to address the scaling mismatch that could be caused when adopting the factorized representation of the optimization variable. The assumptions on the convex function $g : \mathbb{R}^{r\times r} \to \mathbb{R}$ that underlie our analysis are discussed later.

Similar to the PSD case, if the ground-truth matrix $\boldsymbol{X}^*$ is known to meet further structural restrictions (beside being low-rank) and such restrictions can be captured via proper regularization functions on the factor matrices $\boldsymbol{U} \in \mathbb{R}^{d_1\times r}$ and $\boldsymbol{V} \in \mathbb{R}^{d_2\times r}$, then

we propose to improve the above estimator by incorporating extra regularization terms into the optimization procedure as follows

$$(\widehat{\boldsymbol{U}}, \widehat{\boldsymbol{V}}) \in \underset{\substack{\boldsymbol{U}\in\mathbb{R}^{d_1\times r}\\ \boldsymbol{V}\in\mathbb{R}^{d_2\times r}}}{\operatorname{argmin}} \quad \mathcal{L}_n(\boldsymbol{U}\boldsymbol{V}^T, Z_1^n) + \lambda\, g(\boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{V}^T\boldsymbol{V}) + \tau\, \omega_1(\boldsymbol{U}) + \tau\, \omega_2(\boldsymbol{V}), \quad (4.5)$$

where $\omega_i(\cdot) : \mathbb{R}^{d_i\times r} \to \mathbb{R}$, for $i = 1, 2$, is a regularization function that is supposed to promote "low-complexity" structure of the corresponding factor and $\tau > 0$ is a regularization constant.

To solve the regularized optimization problem in (4.5), we first compute the gradients of the empirical loss function $\mathcal{L}_n$ with respect to both $\boldsymbol{U}$ and $\boldsymbol{V}$, which through the rest of this chapter will be denoted as

$$\nabla_U \mathcal{L}_n(\boldsymbol{U}\boldsymbol{V}^T) := \nabla\mathcal{L}_n(\boldsymbol{U}\boldsymbol{V}^T)\boldsymbol{V} \text{ and } \nabla_V \mathcal{L}_n(\boldsymbol{U}\boldsymbol{V}^T) := \nabla\mathcal{L}_n(\boldsymbol{U}\boldsymbol{V}^T)^T\boldsymbol{U},$$

respectively [18]. Moreover, the partial gradients of $g(\boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{V}^T\boldsymbol{V})$ are defined as

$$\nabla_U g(\boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{V}^T\boldsymbol{V}) := \boldsymbol{U}\,\nabla g(\boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{V}^T\boldsymbol{V}), \text{ and}$$
$$\nabla_V g(\boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{V}^T\boldsymbol{V}) := -\boldsymbol{V}\,\nabla g(\boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{V}^T\boldsymbol{V}).$$

Having defined the partial gradients, we then suggest utilizing the proximal descent method as summarized in Algorithm 6.

Algorithm 6 Proximal Descent Algorithm for Bilinear Factorization

Input: initial factors $\boldsymbol{U}_0$ and $\boldsymbol{V}_0$, target rank r, number of iterations T

Set $t = 0$, $\mu = 1/8M\sigma_1^2(\boldsymbol{W}_0)$, where $\boldsymbol{W}_0 := \begin{bmatrix} \boldsymbol{U}_0 \\ \boldsymbol{V}_0 \end{bmatrix}$

repeat

$\widetilde{\boldsymbol{U}}_{t+1} = \boldsymbol{U}_t - \mu\nabla\mathcal{L}_n(\boldsymbol{U}_t\boldsymbol{V}_t^T)\boldsymbol{V}_t - \mu\lambda\cdot\boldsymbol{U}_t\,\nabla g(\boldsymbol{U}_t^T\boldsymbol{U}_t - \boldsymbol{V}_t^T\boldsymbol{V}_t)$

$\widetilde{\boldsymbol{V}}_{t+1} = \boldsymbol{V}_t - \mu\nabla\mathcal{L}_n(\boldsymbol{U}_t\boldsymbol{V}_t^T)^T\boldsymbol{U}_t + \mu\lambda\cdot\boldsymbol{V}_t\,\nabla g(\boldsymbol{U}_t^T\boldsymbol{U}_t - \boldsymbol{V}_t^T\boldsymbol{V}_t)$

$\boldsymbol{U}_{t+1} = \operatorname{prox}_{\omega_1}(\widetilde{\boldsymbol{U}}_{t+1}; \mu\tau)$

$\boldsymbol{V}_{t+1} = \operatorname{prox}_{\omega_2}(\widetilde{\boldsymbol{V}}_{t+1}; \mu\tau)$

$t = t + 1$

until $t = T$

As will be shown later, under the assumptions of our analysis, a constant step-size μ will be enough to attain attractive linear convergence rates for the proposed algorithm, once it is properly initiated.

4.2 Assumptions Underlying Our Analysis

In this section we describe the assumptions, on the empirical loss and regularization functions, upon which our theory will be developed. We also provide explanation for why these assumptions are reasonable for the purpose of our study and show how by imposing them we are building on the existing works of the literature. When stating the assumptions, we take the general non-PSD case as the default case and then comment on how they need to be adjusted to be applicable in the PSD case afterwards.

Assumption 1. The empirical loss function $\mathcal{L}_n$ satisfies the following two conditions:

(*i*) $\mathcal{L}_n$ meets the restricted strong convexity (RSC) condition over the set of rank-r matrices in $\mathbb{R}^{d_1 \times d_2}$, i.e.

$$\mathcal{L}_n(\boldsymbol{X}_2) \geq \mathcal{L}_n(\boldsymbol{X}_1) + \langle \nabla \mathcal{L}_n(\boldsymbol{X}_1), \boldsymbol{X}_2 - \boldsymbol{X}_1 \rangle + \frac{m}{2} \|\boldsymbol{X}_2 - \boldsymbol{X}_1\|_F^2, \tag{4.6}$$

holds for any $\boldsymbol{X}_1, \boldsymbol{X}_2 \in \mathbb{R}^{d_1 \times d_2}$ that are rank-r, where $m > 0$ is a constant.

(*ii*) $\mathcal{L}_n$ meets the restricted smoothness (RSM) condition, i.e.

$$\mathcal{L}_n(\boldsymbol{X}_2) \leq \mathcal{L}_n(\boldsymbol{X}_1) + \langle \nabla \mathcal{L}_n(\boldsymbol{X}_1), \boldsymbol{X}_2 - \boldsymbol{X}_1 \rangle + \frac{M}{2} \|\boldsymbol{X}_2 - \boldsymbol{X}_1\|_F^2, \tag{4.7}$$

where $M \geq m$ is a constant and $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ are arbitrary rank-r matrices in $\mathbb{R}^{d_1 \times d_2}$.

Various forms of the RSC and RSM assumptions widely appear in the computational and statistical analysis of numerical algorithms designed for solving regularized (mainly convex) optimization problems [28,123,124,127]. By imposing such conditions on the set of rank-r matrices, we are mimicking the recent studies in [18, 118, 120], which utilize such assumptions to guarantee local linear convergence rates of (projected) gradient descent type methods for minimizing $\mathcal{L}_n(\boldsymbol{U}\boldsymbol{V}^T)$ (or $\mathcal{L}_n(\boldsymbol{U}\boldsymbol{U}^T)$ in the symmetric PSD case). Finally we would like to note that related notions to RSC, such as *restricted isometry property*, are also widely used in some other existing analyses for low-rank matrix recovery [116,128,129]. In the PSD case of subsection 4.1.1, the above inequalities in (4.6) and (4.7) need to be satisfied for arbitrary rank-r matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ that belong to $\mathcal{S}_d$, the set of positive semi-definite matrices in $\mathbb{R}^{d \times d}$.

Assumption 2. As also required in the work of [18], the regularizer $g : \mathbb{R}^{r \times r} \to \mathbb{R}$ meets the following conditions:

(*i*) g is convex and minimized at zero; i.e. $\nabla g(\mathbf{0}) = \mathbf{0}$.

(*ii*) The gradient, $\nabla g(\boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{V}^T\boldsymbol{V}) \in \mathbb{R}^{r\times r}$ is symmetric for any $(\boldsymbol{U}, \boldsymbol{V})$ pair.

(*iii*) The function g is m_g-strongly convex and M_g-smooth.

To explain the above assumption, notice that an important issue in optimizing $\mathcal{L}_n$ over the factored space is the existence of non-unique factorizations for a given low-rank matrix $\boldsymbol{X}$. In fact, for any invertible matrix $\boldsymbol{G} \in \mathbb{R}^{r\times r}$, the pair $(\boldsymbol{U}\boldsymbol{G}, \boldsymbol{V}\boldsymbol{G}^{-T})$, with $\boldsymbol{G}^{-T}$ denoting the inverse transpose of $\boldsymbol{G}$, attains the same loss function as $(\boldsymbol{U}, \boldsymbol{V})$ since

$$\mathcal{L}_n\left((\boldsymbol{U}\boldsymbol{G})\left(\boldsymbol{V}\boldsymbol{G}^{-T}\right)^T\right) = \mathcal{L}_n\left(\boldsymbol{U}\boldsymbol{G}\boldsymbol{G}^{-1}\boldsymbol{V}^T\right) = \mathcal{L}_n\left(\boldsymbol{U}\boldsymbol{V}^T\right).$$

Therefore, setting $\boldsymbol{G} = t\boldsymbol{I}_r$, where $t > 0$ is a scalar and $\boldsymbol{I}_r$ denotes the $r \times r$ identity matrix, proves that the set of $(\boldsymbol{U}, \boldsymbol{V})$ pairs minimizing $\mathcal{L}_n$ is even unbounded.

Defining the *balanced* factors $\boldsymbol{U}^* := \boldsymbol{A}^*\boldsymbol{\Sigma}^{*1/2} \in \mathbb{R}^{d_1\times r}$ and $\boldsymbol{V}^* := \boldsymbol{B}^*\boldsymbol{\Sigma}^{*1/2} \in \mathbb{R}^{d_2\times r}$, where $\boldsymbol{A}^*\boldsymbol{\Sigma}^*\boldsymbol{B}^{*T}$ is the (truncated) singular value decomposition (SVD) of the low-rank matrix $\boldsymbol{X}^* \in \operatorname{argmin}_{\boldsymbol{X}\in\mathbb{R}^{d_1\times d_2}} \mathcal{L}(\boldsymbol{X})$, the set of "equally-footed" factorizations [18] for $\boldsymbol{X}^*$ may be expressed as below

$$\mathcal{X}^* := \{(\boldsymbol{U}, \boldsymbol{V}) | \boldsymbol{U} = \boldsymbol{U}^*\boldsymbol{R} \text{ and } \boldsymbol{V} = \boldsymbol{V}^*\boldsymbol{R}, \text{ for } \boldsymbol{R} \in \mathcal{O}_r\}, \tag{4.8}$$

where $\mathcal{O}_r := \{\boldsymbol{O} \in \mathbb{R}^{r\times r} : \boldsymbol{O}^T\boldsymbol{O} = \boldsymbol{I}_r\}$ is the set of orthonormal matrices in $\mathbb{R}^{r\times r}$. Interestingly, $\mathcal{X}^*$ forms a bounded subset of optimal factorization pairs, since for any $(\boldsymbol{U}, \boldsymbol{V}) \in \mathcal{X}^*$ we have

$$\|\boldsymbol{U}\|_F^2 = \|\boldsymbol{U}^*\|_F^2 = \|\boldsymbol{\Sigma}^{*1/2}\|_F^2 = \sum_{i=1}^{r} \sigma_i(\boldsymbol{X}^*),$$
$$\|\boldsymbol{V}\|_F^2 = \|\boldsymbol{V}^*\|_F^2 = \|\boldsymbol{\Sigma}^{*1/2}\|_F^2 = \sum_{i=1}^{r} \sigma_i(\boldsymbol{X}^*).$$

The function $g(\boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{V}^T\boldsymbol{V})$ attains its minimum value at any $(\boldsymbol{U}, \boldsymbol{V}) \in \mathcal{X}^*$, because for any such pair we have

$$g(\boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{V}^T\boldsymbol{V}) = g(\boldsymbol{R}^T\boldsymbol{U}^{*T}\boldsymbol{U}^*\boldsymbol{R} - \boldsymbol{R}^T\boldsymbol{V}^{*T}\boldsymbol{V}^*\boldsymbol{R}) = g(\boldsymbol{R}^T\boldsymbol{\Sigma}^*\boldsymbol{R} - \boldsymbol{R}^T\boldsymbol{\Sigma}^*\boldsymbol{R}) = g(\mathbf{0}),$$

and we know that $\mathbf{0}$ minimizes g by the first condition of Assumption 2. However, for other possible factorizations of $\boldsymbol{X}^*$, that are outside $\mathcal{X}^*$ and could be expressed

as $(\boldsymbol{U}^*\boldsymbol{G}, \boldsymbol{V}^*\boldsymbol{G}^{-T})$ with $\boldsymbol{G} \in \mathbb{R}^{r\times r}$ being an arbitrary invertible matrix, the matrix $\boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{V}^T\boldsymbol{V}$ does not equal $\mathbf{0}$ and so the regularization term will not be minimum. Therefore, the addition of $g(\boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{V}^T\boldsymbol{V})$ to the objective function will encourage the solutions to be balanced (see [18] for more details). Since the ultimate goal of our study is to introduce and analyze a factorization-based approach to estimate a (structured) low-rank matrix $\boldsymbol{X}^*$, therefore restricting ourselves to the set of equally-footed factorizations would cause no loss of generality in terms of the theory.

Moreover, as the convergence proof reveals, adding $g(\boldsymbol{U}^T\boldsymbol{U}-\boldsymbol{V}^T\boldsymbol{V})$, with a strongly convex g, is crucial to ensure restricted strong convexity of $\mathcal{L}_n(\boldsymbol{U}\boldsymbol{V}^T)+\lambda\, g(\boldsymbol{U}^T\boldsymbol{U}-\boldsymbol{V}^T\boldsymbol{V})$ with respect to the *lifted* variable $\boldsymbol{W}\boldsymbol{W}^T$, where $\boldsymbol{W} = \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix}$. This point is further discussed in the proof section.

Finally, we note that, in the case of PSD matrices, if $\boldsymbol{X}^* = \boldsymbol{A}^*\boldsymbol{\Sigma}^*\boldsymbol{A}^{*T}$ represents the SVD of the PSD matrix $\boldsymbol{X}^*$, with $\boldsymbol{A}^* \in \mathbb{R}^{d\times d}$ and $\boldsymbol{\Sigma}^* \in \mathbb{R}^{d\times r}$ being unitary and diagonal matrices, respectively, then the set of optimal factors could be defined as follows

$$\mathcal{X}^*_{\text{sym}} := \left\{ \boldsymbol{U} \middle| \boldsymbol{U} = \boldsymbol{A}^*\boldsymbol{\Sigma}^{*1/2}\boldsymbol{R},\ \text{for } \boldsymbol{R} \in \mathcal{O}_r \right\},$$

implying the rotation ambiguity in estimating an optimal factor.

Assumption 3. The regularizers ω_1 and ω_2 meet the following conditions

(*i*) The regularizers $\omega_1 : \mathbb{R}^{d_1\times r} \to \mathbb{R}$ and $\omega_2 : \mathbb{R}^{d_2\times r} \to \mathbb{R}$ are norms.

(*ii*) The regularizers are invariant to right multiplication of their operands by any orthonormal matrix $\boldsymbol{R} \in \mathcal{O}_r$, i.e.

$$\omega_1(\boldsymbol{U}\boldsymbol{R}) = \omega_1(\boldsymbol{U}),\ \text{and } \omega_2(\boldsymbol{V}\boldsymbol{R}) = \omega_2(\boldsymbol{V})$$

for any $\boldsymbol{U} \in \mathbb{R}^{d_1\times r}$ and $\boldsymbol{V} \in \mathbb{R}^{d_2\times r}$.

(*iii*) Given the subspace $\mathcal{S}_1$ in $\mathbb{R}^{d_1\times r}$, we assume the regularizer ω_1 is decomposable with respect to $\mathcal{S}_1$, i.e. for any $\boldsymbol{U}_1 \in \mathcal{S}_1$ and $\boldsymbol{U}_2 \in \mathcal{S}_1^{\perp}$, it holds that

$$\omega_1(\boldsymbol{U}_1 + \boldsymbol{U}_2) = \omega_1(\boldsymbol{U}_1) + \omega_1(\boldsymbol{U}_2).$$

Similarly, we assume ω_2 is decomposable with respect to the subspace $\mathcal{S}_2$ in $\mathbb{R}^{d_2\times r}$.

In the case of PSD matrices, the regularizer $\omega(\cdot) : \mathbb{R}^{d\times r} \to \mathbb{R}$ must obey the conditions of the above Assumption. Imposing the invariance condition on ω_1 and ω_2 would imply that the value of the regularized loss function in (4.5) remains the same for any $(\boldsymbol{U}, \boldsymbol{V}) \in \mathcal{X}^*$. In general, all rotations of an arbitrary pair of factors $(\boldsymbol{U}, \boldsymbol{V}) \in \mathbb{R}^{d_1\times r} \times \mathbb{R}^{d_2\times r}$ will attain the same objective value in (4.5). Due to this issue, a rotation-invariant type of distance is required for our analysis. One example of such a metric is the so-called *Procrustes distance* defined as follows:

Definition 4.2.1. *For any pair of factors $(\boldsymbol{U}, \boldsymbol{V})$, with $\boldsymbol{U} \in \mathbb{R}^{d_1\times r}$ and $\boldsymbol{V} \in \mathbb{R}^{d_2\times r}$, the Procrustes distance to $(\boldsymbol{U}^*, \boldsymbol{V}^*) \in \mathcal{X}^*$ is defined as follows*

$$dist(\boldsymbol{U}, \boldsymbol{V}; \boldsymbol{U}^*, \boldsymbol{V}^*) := \min_{\boldsymbol{R}\in\mathcal{O}_r} \left\| \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} - \begin{bmatrix} \boldsymbol{U}^*\boldsymbol{R} \\ \boldsymbol{V}^*\boldsymbol{R} \end{bmatrix} \right\|_F . \tag{4.9}$$

In the case of PSD matrices, the definition is slightly simpler as only one factor matrix $\boldsymbol{U} \in \mathbb{R}^{d\times r}$ is involved.

Definition 4.2.2. *For any factor $\boldsymbol{U} \in \mathbb{R}^{d\times r}$ the Procrustes distance to any $\boldsymbol{U}^* \in \mathcal{X}^*_{sym}$ is defined as follows*

$$dist(\boldsymbol{U}; \boldsymbol{U}^*) := \min_{\boldsymbol{R}\in\mathcal{O}_r} \|\boldsymbol{U} - \boldsymbol{U}^*\boldsymbol{R}\|_F . \tag{4.10}$$

By making the invariance assumption, we are in fact building upon the analyses of [17,120,122] which study non-convex *constrained* factorization problems, and assume the elements of the constraint set are closed under right multiplication by orthonormal matrices. Our assumption naturally brings this closeness condition to the regularized problem. There are numerous additional studies that utilize regularizers meeting this assumption in their factorization applications [95, 111, 117, 130, 131].

Decomposable regularizers have been extensively employed for analyzing (convex or non-convex) regularized M-estimators [28, 123, 132, 133]. As a short note, we remark that assuming ω_1 is a norm, then by the virtue of the triangle inequality, for any arbitrary $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ in $\mathbb{R}^{d_1\times r}$, we have $\omega_1(\boldsymbol{U}_1 + \boldsymbol{U}_2) \leq \omega_1(\boldsymbol{U}_1) + \omega_1(\boldsymbol{U}_2)$. Imposing the decomposability assumption on $\omega_1(\cdot)$ will then imply that the triangle inequality should hold with equality for any $(\boldsymbol{U}_1, \boldsymbol{U}_2) \in (\mathcal{S}_1, \mathcal{S}_1^{\perp})$. Ideally, the structure of $\boldsymbol{U}^*$ can be characterized by specifying a subspace $\mathcal{S}_1$ in $\mathbb{R}^{d_1\times r}$ such that $\boldsymbol{U}^* \in \mathcal{S}_1$. Then exploiting

a regularizer ω_1 that is decomposable with respect to $\mathcal{S}_1$, in the problem formulation will penalize the perturbation of iteration factors $\boldsymbol{U}_t$ from the low-dimensional subspace $\mathcal{S}_1$ as much as possible. A similar argument can be given for a decomposable ω_2.

Finally, to translate the factor complexity measured in terms of the regularizers's values to our commonly-used Frobenius norm, we need to define the following notion of subspace compatibility constant:

Definition 4.2.3. *For any subspace $\mathcal{S} \subseteq \mathbb{R}^{d\times r}$ and regularizer ω, the subspace compatibility constant of $\mathcal{S}$ with respect to ω and Frobenius norm is defined as*

$$\Psi_\omega(\mathcal{S}) := \sup_{\boldsymbol{U}\in\mathcal{S}/\{\boldsymbol{0}\}} \frac{\omega(\boldsymbol{U})}{\|\boldsymbol{U}\|_F}.$$

Notice that if $\mathcal{S}$ is set to be the entire space $\mathbb{R}^{d\times r}$, then the subspace compatibility constant equals the Lipschitz constant of ω, defined as the constant $L > 0$ for which

$$\omega(\boldsymbol{U}) \leq L\|\boldsymbol{U}\|_F$$

holds true for any $\boldsymbol{U} \in \mathbb{R}^{d\times r}$.

4.3 Convergence Result: PSD Case

We begin this section by introducing two piece of notation. At any iteration t, we define $\boldsymbol{R}_t := \operatorname{argmin}_{\boldsymbol{R}\in\mathcal{O}_r} \|\boldsymbol{U}_t - \boldsymbol{U}^*\boldsymbol{R}\|_F$ as the corresponding optimal rotation. To compress the notation, the Procrustes distance to optimality at t-th iteration, i.e. $\operatorname{dist}(\boldsymbol{U}_t, \boldsymbol{U}^*)$, will be denoted by d_t. Before jumping to the main theorem we state a useful lemma, which is proved in section C.2 of the appendix:

Lemma 4.3.1. *For any $\boldsymbol{U}_t$ satisfying $d_t \leq \rho\,\sigma_r(\boldsymbol{U}^*)$, with $\rho \leq \sqrt{m/32M}$, and step-size*

$$\mu_t \leq \frac{1}{4(m+M)\sigma_1^2(\boldsymbol{U}_t)}, \tag{4.11}$$

it holds that, if Assumptions 1 and 3 of the previous section are valid, then at the t-th iteration of the proximal descent method we have

$$d_{t+1}^2 + 2\tau\mu_t\left(\omega(\boldsymbol{U}_{t+1}) - \omega(\boldsymbol{U}^*)\right) \leq (1-\mu_t\alpha_1)\,d_t^2 + \eta^2,$$

where $\alpha_1 := mM\sigma_r^2(\boldsymbol{U}^*)/8(m+M)$ and $\eta^2 := r\,\|\nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_2^2/\big(m(m+M)\,\sigma_1^2(\boldsymbol{U}^*)\big)$.

Remark 4.3.1. *The above result can be viewed as a generalization of the recent local convergence guarantees established for (projected) gradient descent in the context of factorization-based approaches to low-rank matrix recovery [17, 116, 118, 120–122]. To see this, notice that standard gradient decent algorithm is indeed a special instance of the current framework upon setting the regularizer to be the zero function, i.e.* $\omega(\boldsymbol{U}) = 0$ *for any* $\boldsymbol{U} \in \mathbb{R}^{d\times r}$*. Under such a restriction, the above Lemma implies that, for a proper initialization* $d_t \leq \rho\, \sigma_r(\boldsymbol{U}^*)$*, with* $\rho \leq \sqrt{m/32M}$*, and a suitable choice of the step-size* μ_t*, the following claim holds about the* t*-th iteration of the gradient descent method:*

$$d_{t+1}^2 \leq (1 - \mu_t \alpha_1)\, d_t^2 + \eta^2, \tag{4.12}$$

where the only simplification here is by setting ω *to zero. This statement asserts that, under proper conditions, taking one gradient descent iteration reduce the square of the Procrustes distance to the optimal set by a constant factor* $(1 - \mu_t\alpha_1) \in (0, 1)$*. The scalar* α_1 *relies on the RSC and RSM constants of the loss function in a manner that agrees with classical results in convex optimization on the convergence rate of gradient methods for minimizing strongly convex functions [14].*

Remark 4.3.2. *The extra additive error term* η *in* (4.12) *does not vanish as the iterations progress. In fact, this term can be interpreted as the* statistical *imprecision of the estimatior. By recursively applying the inequality* (4.12) *for a sequence of consecutive iterations, we are able to demonstrate that the gradient descent iterations linearly converge to a neighborhood of* $\boldsymbol{U}^*$ *whose radius is quantified by the statistical error term* η*. Similar results, investigating the interplay between computational and statistical aspects of estimators are established in [17, 123, 124, 127].*

Remark 4.3.3. *In the statement of the Lemma,* $\boldsymbol{U}^*$ *is only used as an arbitrary representative of the optimal solution set* $\mathcal{X}^*_{sym}$*. In fact, by the definition of* $\mathcal{X}^*_{sym}$*, the spectral characteristics of all its members are the same, i.e.* $\sigma_i(\boldsymbol{U}^*)$ *is constant over* $\mathcal{X}^*_{sym}$*.*

In the next Lemma we take advantage of the decomposability condition of Assumption 3 to demonstrate a convergence result, in terms of the Procrustes distance.

Lemma 4.3.2. *For any* $\boldsymbol{U}_t$ *satisfying* $d_t \leq \rho\, \sigma_r(\boldsymbol{U}^*)$*, with* $\rho \leq \sqrt{m/32M}$*, and step-size* $\mu_t \leq 1/4(m + M)\sigma_1^2(\boldsymbol{U}_t)$*, one iteration of Algorithm 5 obeys*

$$d_{t+1} \leq \sqrt{1 - \mu_t\alpha_1} \cdot d_t + \eta + 2\tau\mu_t\, \Psi_\omega(\mathcal{S}), \tag{4.13}$$

where $\alpha_1 = mM\sigma_r^2(\boldsymbol{U}^*)/8(m+M)$ *and* $\eta = \sqrt{\frac{r}{m(m+M)}} \cdot \frac{\|\nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_2}{\sigma_1(\boldsymbol{U}^*)}$. *Moreover, if the following condition is true*

$$\eta + 2\tau\mu_t\Psi_\omega(\mathcal{S}) \leq (1-\gamma_t)\,\rho\,\sigma_r(\boldsymbol{U}^*),$$

where $\gamma_t := \sqrt{1-\mu_t\alpha_1}$, *then it is guaranteed that* $d_{t+1} \leq \rho\sigma_r(\boldsymbol{U}^*)$.

Lemma 4.3.2, which is proved in C.2, implies the reduction of Procrustes distance to optimality by a constant γ_t, when the initial distance is small enough.

Remark 4.3.4. *The term* $\eta + 2\tau\mu_t\,\Psi_\omega(\mathcal{S})$ *in* (4.13) *is related to the statistical error of the iterative procedure. As also pointed out in the discussion immediately following Theorem 1 of [17], the upper bound requirement on* $\eta + 2\tau\mu_t\Psi_\omega(\mathcal{S})$, *which is imposed to control the size of the statistical error term, entails no loss of generality. If this assumption fails, the current distance to optimality* d_t *is less than the statistical error and therefore the current iterate* $\boldsymbol{U}_t$ *already lies inside a ball of the statistical error radius, centered at the optimal solution* $\boldsymbol{U}^*$. *However, when the assumption holds, the first part of Theorem implies a geometric reduction of the distance to optimality.*

By recursive application of Lemma 4.3.2, while fixing the step-size parameter to be a (strictly) positive constant, the next Theorem guarantee a geometric reduction of distance to optimality that continues until the iterates enter a neighborhood of the ground-truth solution of the size equal to the statistical error.

Theorem 4.3.1. *Suppose Assumptions 1 and 3 are met by* $\mathcal{L}_n$ *and* ω *in* (4.2). *If the initial factor* $\boldsymbol{U}_0$ *satisfies* $dist(\boldsymbol{U}_0, \boldsymbol{U}^*) \leq \rho\sigma_r(\boldsymbol{U}^*)$, *with* $\rho \leq \sqrt{m/32M}$, *then setting*

$$\mu \leq \frac{1}{4(1+\rho)^2(m+M)\,\sigma_1^2(\boldsymbol{U}^*)},$$

and assuming the following statistical error condition

$$\epsilon_{stat} := \eta + 2\mu\tau\Psi_\omega(\mathcal{S}) \leq (1-\gamma)\rho\sigma_r(\boldsymbol{U}^*),$$

where $\eta = \sqrt{\frac{r}{m(m+M)}} \cdot \frac{\|\nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_2}{\sigma_1(\boldsymbol{U}^*)}$ *and* $\gamma = \sqrt{1 - \frac{mM}{8(m+M)}\mu\sigma_r^2(\boldsymbol{U}^*)} \in (0,1)$, *it holds that*

$$d_T \leq \gamma^T\, d_0 + \left(\frac{\gamma^{T-1}-1}{\gamma-1}\right)\epsilon_{stat}.$$

4.4 Convergence Result: Non-PSD Case

Before jumping to the results we introduce a few notations to ease the presentation. We let $M_{\max} := \max\{M, M_g\}$, $m_{\min} := \min\{m, m_g\}$, $d_t := \text{dist}(\boldsymbol{U}_t, \boldsymbol{V}_t; \boldsymbol{U}^*, \boldsymbol{V}^*)$, and finally $\boldsymbol{W}_t := \begin{bmatrix} \boldsymbol{U}_t \\ \boldsymbol{V}_t \end{bmatrix} \in \mathbb{R}^{(d_1+d_2)\times r}$. The following Lemma, proved in C.3, extends Lemma 4.3.1 to non-PSD matrices.

Lemma 4.4.1. *For any $(\boldsymbol{U}_t, \boldsymbol{V}_t) \in \mathbb{R}^{d_1\times r} \times \mathbb{R}^{d_2\times r}$, which satisfies $d_t \leq \rho \cdot \sqrt{\sigma_r(\boldsymbol{X}^*)}$, with $\rho^2 \leq m_{\min}/68\, M_{\max}$, if we choose the step-size $\mu_t \leq 1/16 M_{\max}\sigma_1^2(\boldsymbol{W}_t)$ and set the regularization parameter as $\lambda = 1/8$, then the t-th iteration of Algorithm 6 satisfies*

$$d_{t+1}^2 \leq (1 - \mu_t\, \alpha)\; d_t^2 - 2\mu_t\tau\,(\omega_1(\boldsymbol{U}_{t+1}) - \omega_1(\boldsymbol{U}^*) + \omega_2(\boldsymbol{V}_{t+1}) - \omega_2(\boldsymbol{V}^*)) + \eta^2,$$

where $\alpha := m_{\min} \cdot \sigma_r(\boldsymbol{X}^*)/16$ and $\eta := \sqrt{\frac{r}{4mM_{\max}\,\sigma_1(\boldsymbol{X}^*)}} \cdot \|\nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_2$.

By also incorporating Assumption 3 we will obtain a slightly simplified form of the above Lemma. The proof of this Lemma is given in section C.3 of the appendix.

Lemma 4.4.2. *For any $(\boldsymbol{U}_t, \boldsymbol{V}_t) \in \mathbb{R}^{d_1\times r} \times \mathbb{R}^{d_2\times r}$, which satisfies $d_t \leq \rho\sqrt{\sigma_r(\boldsymbol{X}^*)}$, with $\rho^2 \leq m_{\min}/68\, M_{\max}$, if we choose the step-size $\mu_t \leq 1/16 M_{\max}\sigma_1^2(\boldsymbol{W}_t)$ and set the regularization parameter to $\lambda = 1/8$, then the t-th iteration of Algorithm 6 satisfies*

$$d_{t+1} \leq \sqrt{1 - \mu_t\,\alpha} \cdot d_t + 2\mu_t\tau\,(\Psi_{\omega_1}(\mathcal{S}_1) + \Psi_{\omega_2}(\mathcal{S}_2)) + \eta, \tag{4.14}$$

where $\alpha = m_{\min} \cdot \sigma_r(\boldsymbol{X}^*)/16$ and $\eta = \sqrt{\frac{r}{4mM_{\max}\,\sigma_1(\boldsymbol{X}^*)}} \cdot \|\nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_2$. Moreover, assuming the following condition

$$2\mu_t\tau\,(\Psi_{\omega_1}(\mathcal{S}_1) + \Psi_{\omega_2}(\mathcal{S}_2)) + \eta \leq (1 - \gamma_t)\,\rho\sqrt{\sigma_r(\boldsymbol{X}^*)},$$

where $\gamma_t := \sqrt{1 - \mu_t\,\alpha}$ is the contraction factor, then guarantees that $d_{t+1} \leq \rho\sqrt{\sigma_r(\boldsymbol{X}^*)}$.

Remark 4.4.1. *As long as the additive error term $\eta + 2\mu_t\tau\,(\Psi_{\omega_1}(\mathcal{S}_1) + \Psi_{\omega_2}(\mathcal{S}_2))$ is small compared to the current distance dist$(\boldsymbol{U}_t, \boldsymbol{V}_t; \boldsymbol{U}^*, \boldsymbol{V}^*)$, the Lemma guarantees a geometric reduction of distance to optimality. As in the PSD case, the additive error term can be interpreted as the statistical imprecision of estimating $(\boldsymbol{U}^*, \boldsymbol{V}^*)$ via Algorithm 6. Similar results are provided in [123, 127] for analyzing the computational and statistical aspects of convex estimators as well as in [17, 124] for studying gradient-type methods that solve (constrained) non-convex formulations to estimate structured matrices.*

Remark 4.4.2. *As mentioned in Remark 4.3.4, the condition required by the second part of the Lemma is only to make sure that the radius of the statistical neighborhood is small enough so that the inequality in* (4.14) *implies an effective reduction in terms of the Procrustes distance to the ground-truth solution. If it fails to hold, then the current solution* $(\boldsymbol{U}_t, \boldsymbol{V}_t)$ *already satisfies an error bound better than what is guaranteed for subsequent iterates.*

Remark 4.4.3. *As we will illustrate later in the experiments section, the term* η*, which is a function of* $\|\nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_2$*, can be related to noise characteristics and the number of provided measurements.*

Theorem 4.4.1. *Suppose Assumptions 1, 2, and 3 are met by the functions* $\mathcal{L}_n, g, \omega_1,$ *and* ω_2 *appearing in the statement of problem* (4.5)*. Moreover, assume the initial pair* $(\boldsymbol{U}_0, \boldsymbol{V}_0)$ *obeys the condition* $d_0 \leq \rho\sqrt{\sigma_r(\boldsymbol{X}^*)}$*, with* $\rho^2 \leq m_{\min}/68M_{\max}$*. Then, setting*

$$\mu \leq \frac{1}{32M_{\max}(1+\rho)^2\,\sigma_1(\boldsymbol{X}^*)} \;\; \textit{and} \;\; \lambda = \frac{1}{8},$$

and assuming the following statistical error condition

$$\epsilon_{stat} := \eta + 2\mu\tau\left[\Psi_{\omega_1}(\mathcal{S}_1) + \Psi_{\omega_2}(\mathcal{S}_2)\right] \leq (1-\gamma)\,\rho\,\sqrt{\sigma_r(\boldsymbol{X}^*)}$$

where $\eta = \sqrt{\frac{r}{4mM_{\max}\,\sigma_1(\boldsymbol{X}^*)}} \cdot \|\nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_2$ *and* $\gamma = \sqrt{1 - \frac{\mu\, m_{\min}\,\sigma_r(\boldsymbol{X}^*)}{16}}$*, together imply*

$$d_T \leq \gamma^T\, d_0 + \left(\frac{\gamma^{T-1}-1}{\gamma-1}\right)\epsilon_{stat}.$$

Remark 4.4.4. *We note that the dependency of the step-size parameter* μ *on the RSM constant* $M_{\max} = \max\{M, M_g\}$ *resembles that of the step-size appearing in equation* (17) *of [18]. Due to the underlying assumption on the initial proximity of iterations to* $(\boldsymbol{U}^*, \boldsymbol{V}^*)$*, it can be shown (essentially by using Lemma C.1.6 in the appendix) that the singular value* $\sigma_1(\boldsymbol{X}^*)$ *is close to* $\sigma_1(\boldsymbol{X}_0)$*, which is practically computable.*

Remark 4.4.5. *Notice the dependency of the contraction factor* γ *on the RSC constant* $m_{\min}$*. As discussed in the experiments section, by acquiring more number of measurements, the value of the* m *can be improved (increased), which in turn implies a reduction in* γ*. Therefore, having more measurements yields computational speed up.*

4.5 Theorem Implications

In this section we study the implications of our theory in the context of certain well-known applications, namely matrix sensing and matrix completion problems. Before starting the discussion of each application, we would like to briefly mention some regularization functions that meet the conditions of Assumption 3 in our analysis.

4.5.1 Choice of Regularization

One class of convex regularization functions that meet the conditions of our analysis is the class of row-wise $\ell_{2,p}$ norms for $p \geq 1$, which are defined as

$$\|\boldsymbol{U}\|_{2,p} := \left(\sum_{i=1}^{d} \|\boldsymbol{U}_{i*}\|_2^p \right)^{\frac{1}{p}}$$

for any matrix $\boldsymbol{U} \in \mathbb{R}^{d \times r}$. Examples of these norms, that are used in the literature to impose certain low-dimensional structures for matrix factorization, are the row-wise $\ell_{2,1}$ and $\ell_{2,\infty}$. In fact, $\ell_{2,1}$ norm has been recommended by [95] in the context of sparse principal subspace estimation to enable the selection of a small subset of variables generating the principal subspace. Moreover, the row-wise $\ell_{2,\infty}$ norm is employed in [130] to form a heuristic for low-rank matrix completion, which outperforms Frobenius norm regularization when the low-rank factors are known to have bounded entries.

Notice that row-wise $\ell_{2,p}$ norms satisfy the rotation-invariance condition of Assumption 3, since for any orthonormal matrix $\boldsymbol{R} \in \mathcal{O}_r$ it holds that

$$\|\boldsymbol{U}\boldsymbol{R}\|_{2,p} = \left(\sum_{i=1}^{d} \|\boldsymbol{U}_{i*}\boldsymbol{R}\|_2^p \right)^{\frac{1}{p}} = \left(\sum_{i=1}^{d} \|\boldsymbol{U}_{i*}\|_2^p \right)^{\frac{1}{p}} = \|\boldsymbol{U}\|_{2,p}.$$

Moreover, depending on the choice of the subspace $\mathcal{S}$ in $\mathbb{R}^{d \times r}$, the norms meet the decomposability condition as well. Clearly, if we set $\mathcal{S} = \mathbb{R}^{d \times r}$, then $\omega(\boldsymbol{U}_1 + \boldsymbol{U}_2) = \omega(\boldsymbol{U}_1) + \omega(\boldsymbol{U}_2)$, for any $\boldsymbol{U}_1 \in \mathcal{S} = \mathbb{R}^{d \times r}$ and $\boldsymbol{U}_2 \in \mathcal{S}^{\perp} = \emptyset$. However, if $\mathcal{S}$ is set more carefully to capture the low-dimensional structure of $\boldsymbol{U}^*$, then the resulting expressions that emerge out of our convergence rate analysis become more revealing. For instance, suppose that $\boldsymbol{U}^*$ is known to be only supported on a subset of size $k < d$ of its rows which is indexed by $S \subset \{1, 2, \cdots, d\}$, i.e. assume

$$\|\boldsymbol{U}_{i*}\|_2 \neq 0 \iff i \in S.$$

Then, a clever choice for $\mathcal{S}$ will be the subspace of matrices that are only supported on the rows indexed by S, i.e. $\mathcal{S} = \big\{\boldsymbol{U} \in \mathbb{R}^{d\times r} : \boldsymbol{U}_{i*} = \boldsymbol{0} \text{ for any } i \notin S\big\}$. With this choice for $\mathcal{S}$, it can be easily shown that for any $\boldsymbol{U}_1 \in \mathcal{S}$ and $\boldsymbol{U}_2 \in \mathcal{S}^{\perp}$ we have

$$\omega(\boldsymbol{U}_1 + \boldsymbol{U}_2) = \omega(\boldsymbol{U}_1) + \omega(\boldsymbol{U}_2).$$

Moreover, the subspace compatibility constant $\Psi_\omega(\mathcal{S})$ then becomes $\Psi_\omega(\mathcal{S}) = \sqrt{s}$, which (depending on the size of s relative to the ambient dimension d) can be significantly smaller than $\sqrt{d}$, i.e. the Lipschitz constant of the $\ell_{2,1}$ norm.

4.5.2 Matrix Sensing

Assume the following noisy linear measurements are available:

$$y_i = \langle \boldsymbol{A}_i, \boldsymbol{X}^* \rangle + \epsilon_i,\ i = 1, 2, \cdots, n, \tag{4.15}$$

where $\boldsymbol{X}^* \in \mathbb{R}^{d_1\times d_2}$ is the ground-truth low-rank matrix, $\boldsymbol{A}_i$ is the sensing matrix associated with the i-th measurement, and ϵ_i is an additive observation noise corresponding to this measurement. To study the consequences of our developed convergence theory, we assume the measurement matrices $\{\boldsymbol{A}_i\}_{i=1}^n$ are independently-drawn from the $\boldsymbol{\Sigma}$-ensemble [123, 134]. This means that, defining $\boldsymbol{a}_i := \text{vec}(\boldsymbol{A}_i) \in \mathbb{R}^D$, with $D := d_1 d_2$, as the vectorization of $\boldsymbol{A}_i$, we assume each $\boldsymbol{a}_i$ is independently drawn from the Gaussian distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$. Furthermore, we assume that $\boldsymbol{X}^*$ is low-rank, i.e. $\text{rank}(\boldsymbol{X}^*) = r \ll \min\{d_1, d_2\}$, and therefore can be represented as $\boldsymbol{X}^* = \boldsymbol{U}^*\boldsymbol{V}^{*T}$, where $(\boldsymbol{U}^*, \boldsymbol{V}^*) \in \mathbb{R}^{d_1\times r} \times \mathbb{R}^{d_2\times r}$. Given the above assumptions, the natural empirical loss function that is suitable to be minimized, with respect to factor matrices $(\boldsymbol{U}, \boldsymbol{V})$, is the least squares [116, 122]

$$\mathcal{L}_n(\boldsymbol{X}) = \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \langle \boldsymbol{A}_i, \boldsymbol{X}\rangle\right)^2 = \frac{1}{2n}\|\boldsymbol{y} - \mathcal{A}(\boldsymbol{X})\|_2^2, \tag{4.16}$$

where $\mathcal{A} : \mathbb{R}^{d_1\times d_2} \to \mathbb{R}^n$ denotes the linear transformation which maps $\boldsymbol{X} \in \mathbb{R}^{d_1\times d_2}$ into a vector $\mathcal{A}(\boldsymbol{X}) \in \mathbb{R}^n$ whose i-th entry is given by $\langle \boldsymbol{A}_i, \boldsymbol{X}\rangle$. It can be easily shown that the population risk function is then given by $\mathcal{L}(\boldsymbol{X}) = \frac{1}{2}\|\boldsymbol{\Sigma}^{\frac{1}{2}}(\boldsymbol{x} - \boldsymbol{x}^*)\|_2^2$, where $\boldsymbol{x} := \text{vec}(\boldsymbol{X})$ and $\boldsymbol{x}^* := \text{vec}(\boldsymbol{X}^*)$.

Under the above presumptions on the generation of the measurement matrices $\boldsymbol{A}_i \in \mathbb{R}^{d_1 \times d_2}$, for $i = 1, 2, \cdots n$, the following Lemma (proved in section C.4 of the appendix) describes the sufficient condition, on the number of measurements, which guarantees that the RSC and RSM conditions of Assumption 1 hold true.

Lemma 4.5.1. *There exist universal positive constants (c_0, c_1) such that, with probability at least $1 - \exp(-c_0\, n)$, the loss function $\mathcal{L}_n(\boldsymbol{X}) = \frac{1}{2n}\|\mathcal{A}(\boldsymbol{X}) - \boldsymbol{y}\|_2^2$ satisfies Assumption 1, with $m = \lambda_{\min}(\boldsymbol{\Sigma})/4$ and $M = 4\,\lambda_{\max}(\boldsymbol{\Sigma})$, provided that*

$$n \geq \frac{c_1\, \xi(\boldsymbol{\Sigma})}{\lambda_{\min}(\boldsymbol{\Sigma})}\,(rd_1 + rd_2),$$

where $\xi(\boldsymbol{\Sigma}) := \sup_{\|\boldsymbol{u}\|_2=1, \|\boldsymbol{v}\|_2=1} var(\boldsymbol{u}^T \boldsymbol{A} \boldsymbol{v})$ for $vec(\boldsymbol{A}) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$.

In the particular case where the measurement matrices are drawn from the standard Gaussian distribution, i.e. where $\boldsymbol{\Sigma} = \boldsymbol{I}_D$, the D-dimensional identity matrix, then $\xi(\boldsymbol{\Sigma}) = 1$ and the above Lemma ensures the conditions of Assumption one, with $m = 1/4$ and $M = 4$, given that $n \geq c_1 r(d_1 + d_2)$.

To quantify the statistical error term η appearing in the statement of Theorem 4.4.1, we first notice that, in the context of the matrix sensing problem, we have $\|\nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_2 = \frac{1}{n}\,\|\sum_{i=1}^n \epsilon_i \boldsymbol{A}_i\|_2$. The following result, which is Lemma C.1 of [134], can be then utilized to bound the error term

Lemma 4.5.2. *Assume the additive noise vector $\epsilon \in \mathbb{R}^n$ is such that $\|\epsilon\|_2 \leq 2\nu\sqrt{n}$, then there exist universal positive constants c_2, c_3, c_4 such that*

$$\frac{1}{n}\left\|\sum_{i=1}^n \epsilon_i \boldsymbol{A}_i\right\|_2 \leq c_2 \nu\, \xi(\boldsymbol{\Sigma}) \sqrt{\frac{d_1 + d_2}{n}},$$

with probability at least $1 - c_3 \exp(-c_4(d_1 + d_2))$.

We choose the regularizer $g(\boldsymbol{Z}) = \frac{1}{2}\|\boldsymbol{Z}\|_F^2$ for $\boldsymbol{Z} \in \mathbb{R}^{r \times r}$, which meets conditions of Assumption 2 with $m_g = M_g = 1$. The overall optimization problem can be cast as

$$(\widehat{\boldsymbol{U}}, \widehat{\boldsymbol{V}}) \in \underset{\substack{\boldsymbol{U} \in \mathbb{R}^{d_1 \times r} \\ \boldsymbol{V} \in \mathbb{R}^{d_2 \times r}}}{\operatorname{argmin}} \quad \frac{1}{2n}\left\|\boldsymbol{y} - \mathcal{A}(\boldsymbol{U}\boldsymbol{V}^T)\right\|_2^2 + \frac{1}{16}\|\boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{V}^T\boldsymbol{V}\|_F^2. \tag{4.17}$$

The proposed proximal descent algorithm will take the following form

Algorithm 7 Proximal Descent Algorithm for Matrix Sensing Applications

Input: initial factors $\boldsymbol{U}_0$ and $\boldsymbol{V}_0$, target rank r, number of iterations T

Set $t = 0$, $\mu = 1/8M\sigma_1^2(\boldsymbol{W}_0)$, where $\boldsymbol{W}_0 := \begin{bmatrix} \boldsymbol{U}_0 \\ \boldsymbol{V}_0 \end{bmatrix}$

repeat

$\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \frac{\mu}{n}\sum_{i=1}^{n}\left(\langle \boldsymbol{A}_i, \boldsymbol{U}_t\boldsymbol{V}_t^T\rangle - y_i\right)\boldsymbol{A}_i\,\boldsymbol{V}_t - \frac{\mu}{4}\boldsymbol{U}_t\left(\boldsymbol{U}_t^T\boldsymbol{U}_t - \boldsymbol{V}_t^T\boldsymbol{V}_t\right)$

$\boldsymbol{V}_{t+1} = \boldsymbol{V}_t - \frac{\mu}{n}\sum_{i=1}^{n}\left(\langle \boldsymbol{A}_i, \boldsymbol{U}_t\boldsymbol{V}_t^T\rangle - y_i\right)\boldsymbol{A}_i^T\,\boldsymbol{U}_t + \frac{\mu}{4}\boldsymbol{V}_t\left(\boldsymbol{U}_t^T\boldsymbol{U}_t - \boldsymbol{V}_t^T\boldsymbol{V}_t\right)$

$t = t + 1$

until $t = T$.

The assumptions of this section imply the following Corollary of Theorem 4.4.1.

Corollary 4.5.1. *Suppose $n \geq c_1 r(d_1 + d_2)$ noisy linear measurements of the form (4.15) are gathered, where $\{\boldsymbol{A}_i\}_{i=1}^n$ are independently drawn from $\boldsymbol{\Sigma}$-ensemble, with $\boldsymbol{\Sigma} = \boldsymbol{I}$, and the noise vector meets the condition $\|\boldsymbol{\epsilon}\|_2 \leq 2\nu\sqrt{n}$. Furthermore, assume that dist$(\boldsymbol{W}_0, \boldsymbol{W}^*) = d_0 \leq \rho\sqrt{\sigma_r(\boldsymbol{X}^*)}$, with $\rho \leq \frac{1}{50}$. Then, setting $\mu \leq 0.0075/\sigma_1(\boldsymbol{X}^*)$ and assuming the following statistical error condition*

$$\epsilon_{stat} = \frac{c_2}{2}\,\nu\sqrt{\frac{rd_1 + rd_2}{n\,\sigma_1(\boldsymbol{X}^*)}} \leq (1-\gamma)\rho\sqrt{\sigma_r(\boldsymbol{X}^*)},$$

where $\gamma = \sqrt{1 - \frac{\mu\,\sigma_r(\boldsymbol{X}^*)}{64}}$, will ensure the following linear convergence result for T iterations of Algorithm 7:

$$d_T \leq \gamma^T d_0 + \epsilon_{stat},$$

with probability at least $1 - \exp(-c_0 n) - c_3\exp(-c_4(d_1 + d_2))$.

4.5.3 Sparse PCA

Principal component analysis (PCA) is one of the most important and classical techniques for dimensionality reduction. It reduces dimensionality by projecting the data points onto the subspace spanned by the leading eigenvectors of the population covariance matrix $\boldsymbol{\Sigma}$. Since in practice $\boldsymbol{\Sigma}$ is unknown, PCA estimates the leading principal eigenvectors of $\boldsymbol{\Sigma}$ by those of the sample covariance matrix $\boldsymbol{\Sigma}_n$. However, in high dimensional regimes, where the number of acquired data points n is significantly less than the ambient dimension d, classical PCA shows poor performance. Therefore, new variant

of this method are proposed, which incorporate extra available knowledge about the structure of the principal components into the estimation procedure.

To formally start the discussion on our approach to principal component estimation, we assume that the population covariance matrix $\boldsymbol{\Sigma} \in \mathcal{S}^{d\times d}$ can be expressed by the following *spiked* model [135]

$$\boldsymbol{\Sigma} = \boldsymbol{X}^* + \nu \boldsymbol{I}_d, \tag{4.18}$$

where $\boldsymbol{X}^*$ contains the $r \ll d$ principal data components, and $\nu > 0$ denotes the eigenvalue corresponding to weaker components. Then, given the eigen-decomposition $\boldsymbol{X}^* = \boldsymbol{Q}^*\boldsymbol{\Lambda}^*\boldsymbol{Q}^{*T}$, the columns of the orthonormal matrix $\boldsymbol{Q}^*$ specify the leading eigenvectors of $\boldsymbol{X}^*$ (and $\boldsymbol{\Sigma}$), which correspond to the eigenvalues $\lambda_1^*, \cdots, \lambda_r^*$ for $\boldsymbol{X}^*$ (the leading eigenvalues of $\boldsymbol{\Sigma}$ are $\lambda_1^* + \nu, \cdots, \lambda_r^* + \nu$).

In this section, we focus on the setting where the leading eigenvectors are *jointly k-sparse* [17, 95, 136]. Defining the row support of $\boldsymbol{Q}^*$ as follows

$$\mathrm{supp}(\boldsymbol{Q}^*) := \{i \in [d] : \boldsymbol{Q}^*_{i*} \neq \boldsymbol{0}\},$$

this assumption implies that $|\mathrm{supp}(\boldsymbol{Q}^*)| \leq k$, where $r \leq k \leq d$.

Since $\boldsymbol{X}^*$ is a symmetric PSD matrix of rank-r, adopting the Burer-Monteiro factorization, it can also be represented as $\boldsymbol{X}^* = \boldsymbol{U}^*\boldsymbol{U}^{*T}$, with $\boldsymbol{U}^* = \boldsymbol{Q}^*\boldsymbol{\Lambda}^{*1/2}\boldsymbol{R}$ for some orthonormal matrix $\boldsymbol{R} \in \mathcal{O}_r$. It is then easy to see that

$$\mathrm{supp}(\boldsymbol{U}^*) = \mathrm{supp}(\boldsymbol{Q}^*).$$

Given the above spiked covariance matrix, assume n i.i.d. measurement vectors $\boldsymbol{x}_i \in \mathbb{R}^d$, for $i = 1, 2, \cdots, n$, are randomly drawn from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$. Having defined the empirical covariance matrix $\boldsymbol{\Sigma}_n := \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T$, we want to investigate the quality of the estimate $\widehat{\boldsymbol{X}} := \widehat{\boldsymbol{U}}\widehat{\boldsymbol{U}}^T$, where $\widehat{\boldsymbol{U}}$ is a solution of the following problem

$$\min_{\boldsymbol{U}\in\mathbb{R}^{d\times r}} \frac{1}{2}\|\boldsymbol{\Sigma}_n - \boldsymbol{U}\boldsymbol{U}^T\|_F^2 + \tau\|\boldsymbol{U}\|_{2,1}. \tag{4.19}$$

Apparently, the above formulation lies in the general framework of subsection 4.1.1 for estimating PSD rank-r matrices $\boldsymbol{X}^*$, with $\mathcal{L}_n(\boldsymbol{X}) = \frac{1}{2}\|\boldsymbol{\Sigma}_n - \boldsymbol{X}\|_F^2$ and $\omega(\boldsymbol{U}) = \|\boldsymbol{U}\|_{2,1}$. Noticing that the proximal operator of the $\ell_{2,1}$ norm is row-wise defined as

$$\Big(\mathrm{prox}_{\ell_{2,1}}(\boldsymbol{U}; \mu\tau)\Big)_{i*} = \max\left\{1 - \frac{\mu\tau}{\|\boldsymbol{U}_{i*}\|_2}, 0\right\}\boldsymbol{U}_{i*}, \quad \text{for every } \ i = 1, 2, \cdots, d,$$

the following numeric scheme is an instantiation of the general Algorithm 5.

Algorithm 8 Proximal Descent Method for Sparse PCA

Input: initial factor $\boldsymbol{U}_0$, target rank r, number of iterations T

Set $t = 0$, $\mu = 1/8\sigma_1^2(\boldsymbol{U}_0)$

repeat

$\widetilde{\boldsymbol{U}}_{t+1} = (\boldsymbol{U}_t\boldsymbol{U}_t^T - \boldsymbol{\Sigma}_n)\boldsymbol{U}_t$

$\boldsymbol{U}_{t+1} = \text{prox}_{\ell_{2,1}}(\widetilde{\boldsymbol{U}}_{t+1}; \mu\tau)$

$t = t + 1$

until $t = T$

To study the implications of our convergence analysis here, it is required to first ensure that the assumptions which underlie our analysis, are valid. Indeed, as discussed in sub-section 4.5.1 of the current section, our choice of regularizer $\omega(\boldsymbol{U}) = \|\boldsymbol{U}\|_{2,1}$ meets the conditions of Assumption 3. In particular, this norm meets the invariance condition, i.e. $\omega(\boldsymbol{UR}) = \omega(\boldsymbol{U})$ for any $\boldsymbol{R} \in \mathcal{O}_r$, and is decomposable with respect to the subspace $\mathcal{S} = \left\{\boldsymbol{U} \in \mathbb{R}^{d\times r} : \boldsymbol{U}_{i*} = \boldsymbol{0} \text{ for any } i \notin \text{supp}(\boldsymbol{U}^*)\right\}$, for any $\boldsymbol{U}^*$ satisfying $\boldsymbol{U}^*\boldsymbol{U}^{*T} = \boldsymbol{X}^*$. Furthermore, the following Lemma, which is proved in the appendix, is helpful towards bounding the statistical error term:

Lemma 4.5.3. *Assume n i.i.d. samples are drown from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is structured as in* (4.18)*. Then, it holds that*

$$\|\boldsymbol{X}^* - \boldsymbol{\Sigma}_n\|_2 \leq c'\sigma_1(\boldsymbol{X}^*)\max\left\{\frac{k\log d}{n}, \sqrt{\frac{k\log d}{n}}\right\} + \nu^2\left(1 + 3\sqrt{\frac{d}{n}}\right)$$

with probability at least $1 - 2\exp(-d/2) - 2d^{-4}$.

The following Corollary is a direct consequence of Theorem 4.3.1:

Corollary 4.5.2. *Assume n i.i.d. samples are drawn from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is structured as in* (4.18)*, with $\boldsymbol{X}^* = \boldsymbol{U}^*\boldsymbol{U}^{*T}$ being rank-r and $\boldsymbol{U}^*$ having only $k \leq d$ non-zero rows. Furthermore, assume $\boldsymbol{U}_0$ in Algorithm 8 is such that dist$(\boldsymbol{U}_0, \boldsymbol{U}^*) \leq \rho\sqrt{\sigma_r(\boldsymbol{X}^*)}$, with $\rho \leq 1/6$, and the step-size μ satisfies $\mu \leq 1/12\sigma_1(\boldsymbol{X}^*)$. Then, if the statistical error condition*

$$\epsilon_{stat} = \sqrt{\frac{r}{2\,\sigma_1(\boldsymbol{X}^*)}} \cdot \|\boldsymbol{\Sigma}_n - \boldsymbol{X}^*\|_2 + \frac{\tau\sqrt{k}}{6\,\sigma_1(\boldsymbol{X}^*)} \leq \frac{1-\gamma}{6}\sqrt{\sigma_r(\boldsymbol{X}^*)}$$

holds with $\gamma = \sqrt{1 - \frac{1}{200} \cdot \frac{\sigma_r(\boldsymbol{X}^*)}{\sigma_1(\boldsymbol{X}^*)}}$*, then after* T *iterations of Algorithm 8, we have*

$$d_T \leq \gamma^T d_0 + \epsilon_{stat}.$$

4.5.4 Matrix Completion

Given exact (or noisy) observations of a subset of entries of a matrix $\boldsymbol{X}^* \in \mathbb{R}^{d_1 \times d_2}$, the task of matrix completion is to reconstruct $\boldsymbol{X}^*$ exactly (or approximately) [137]. This problem has applications in collaborative filtering and recommendation systems [88–92].

In general, the matrix completion problem is ill-posed and, unless if further structural assumptions are made on $\boldsymbol{X}^*$, its recoverability cannot be ensured. A popular and practically-sound structural assumption is the low-rankness of $\boldsymbol{X}^*$, i.e. that if $r := \text{rank}(\boldsymbol{X}^*)$, then it is significantly smaller than ambient dimensions: $r \ll \min\{d_1, d_2\}$.

With the low-rank assumption in place, various numerical procedures are proposed to estimate $\boldsymbol{X}^*$. In particular, an important branch of works have incorporated the nuclear norm of $\boldsymbol{X}^*$ (denoted by $\|\boldsymbol{X}^*\|_*$) as the convex approximation to its rank and yielded convex estimators with appealing numerical and theoretical properties [100].

As noticed in the theoretical works of [101, 138–140], which examined conditions to guarantee successful recovery of the low-rank $\boldsymbol{X}^*$ from partial entry-wise observations, the low-rank assumption by itself does not entirely solve the identifiability issue for matrix completion. For instance, there exist low-rank matrices which are also sparse (sometimes called "*spiky*") and cannot be recovered from any set of entry-wise observations, as long as the number of such observations is smaller than the total number of matrix entries $d_1 d_2$. Therefore, the so-called *incoherence* assumption was also assumed (on top of the low-rankness) in [101] to ensure the exact recovery of $\boldsymbol{X}^*$ via nuclear norm minimization. A related condition to incoherence assumption was later introduced by [140], which is expressed in terms of the *spikiness ratio* of $\boldsymbol{X}^*$ (definition to follow).

To start the mathematical discussion, assume that each (i, j) entry of $\boldsymbol{X}^*$ is observed with probability $p \in (0, 1)$ independent of other entries. The observations can then be collectively represented by the matrix $\boldsymbol{Y} \in \mathbb{R}^{d_1 \times d_2}$, which is element-wise defined as

follows

$$Y_{i,j} := \begin{cases} X^*_{i,j} + Z_{i,j}, & \text{with probability } p, \\ *, & \text{otherwise,} \end{cases}$$

where $Z_{i,j}$ denotes noise and inaccuracy in observing $X^*_{i,j}$. Following the existing literature, to enable the recovery of $\boldsymbol{X}^*$, we assume $\text{rank}(\boldsymbol{X}^*) = r \ll \min\{d_1, d_2\}$. Moreover, borrowing the definition of the spikiness ratio from [140]

$$\alpha(\boldsymbol{X}^*) := \frac{\|\boldsymbol{X}^*\|_\infty}{\|\boldsymbol{X}^*\|_F},$$

where $\|\boldsymbol{X}\|_\infty := \max_{(i,j)} |X_{i,j}|$, we also demand this ratio to be small. This requirement essentially guarantees that the energy of $\boldsymbol{X}^*$ is not captured by just a few of its entries and therefore it is not spiky. Notice that, essentially by the relationship between $\|\cdot\|_\infty$ and $\|\cdot\|_F$ norms, it can be shown that $\alpha(\boldsymbol{X}^*) \in \left[1/\sqrt{d_1 d_2}, 1\right]$.

With these assumptions in place, and before going to the discussion of our proposed factorization-based approach for matrix completion, we would like to mention a relevant convex estimator, which is reminiscent of the approach in [140]

$$\begin{aligned} \min_{\boldsymbol{X} \in \mathbb{R}^{d_1 \times d_2}} \quad & \frac{1}{2p} \|\Pi_\Omega(\boldsymbol{Y} - \boldsymbol{X})\|_F^2 + \tau \|\boldsymbol{X}\|_\star \\ \text{subject to} \quad & \|\boldsymbol{X}\|_\infty \leq \alpha^*, \end{aligned} \tag{4.20}$$

where $\alpha^* \geq 1$ is a scalar, $\Omega \subset \{1, 2, \cdots, d_1\} \times \{1, 2, \cdots, d_2\}$ denotes the index set for acquired partial measurements, and $\Pi_\Omega(\cdot)$ represents the projection operator associated with Ω as follows

$$\left(\Pi_\Omega(\boldsymbol{X})\right)_{i,j} = \begin{cases} \boldsymbol{X}_{i,j}, & \text{if } (i,j) \in \Omega, \\ 0, & \text{otherwise.} \end{cases}$$

In spite the convexity of the above estimator, which brings nice theoretical guarantees for its recovery characteristics, in high dimensional regimes, it demands high computational and memory resources. Therefore, an alternative line of work has recently analyzed factorization-based approaches for matrix completion [17, 117, 122]. For instance, upon representing $\boldsymbol{X}^*$ as $\boldsymbol{X}^* = \boldsymbol{U}^* \boldsymbol{V}^{*T}$ with $\boldsymbol{U}^* \in \mathbb{R}^{d_1 \times r}$ and $\boldsymbol{V}^* \in \mathbb{R}^{d_2 \times r}$, the work of [122] has adopted a constrained formulation like what follows:

$$\begin{aligned} \min_{\substack{\boldsymbol{U} \in \mathbb{R}^{d_1 \times r} \\ \boldsymbol{V} \in \mathbb{R}^{d_2 \times r}}} \quad & \frac{1}{2p} \left\|\Pi_\Omega(\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{V}^T)\right\|_F^2 + \frac{1}{8} \|\boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{V}^T\boldsymbol{V}\|_F^2, \\ \text{subject to} \quad & \|\boldsymbol{U}\|_{2,\infty} \leq \alpha \ \text{ and } \ \|\boldsymbol{V}\|_{2,\infty} \leq \alpha, \end{aligned}$$

where the constraints on $\|\boldsymbol{U}\|_{2,\infty}$ and $\|\boldsymbol{V}\|_{2,\infty}$ are used to avoid spiky solutions, since $\|\boldsymbol{X}^*\|_\infty \leq \|\boldsymbol{U}^*\|_{2,\infty} \times \|\boldsymbol{V}^*\|_{2,\infty}$. Clearly, the above constrained formulation can be viewed as an instance of (4.5), with the regularizers ω_1 and ω_2 set to the indicator functions of

$$\{\boldsymbol{U} \in \mathbb{R}^{d_1 \times r} : \|\boldsymbol{U}\|_{2,\infty} \leq \alpha\} \text{ and } \{\boldsymbol{V} \in \mathbb{R}^{d_2 \times r} : \|\boldsymbol{V}\|_{2,\infty} \leq \alpha\},$$

respectively. A similar approach is to reconstruct the low-rank matrix $\boldsymbol{X}^* = \boldsymbol{U}^* \boldsymbol{V}^{*T}$ via solving the following regularized formulation

$$\min_{\substack{\boldsymbol{U} \in \mathbb{R}^{d_1 \times r} \\ \boldsymbol{V} \in \mathbb{R}^{d_2 \times r}}} \frac{1}{2p} \left\| \Pi_\Omega(\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{V}^T) \right\|_F^2 + \frac{1}{16} \|\boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{V}^T\boldsymbol{V}\|_F^2 + \tau \|\boldsymbol{U}\|_{2,\infty} + \tau \|\boldsymbol{V}\|_{2,\infty}. \quad (4.21)$$

We will focus on this regularized recast of the problem through the rest of this subsection. To explain the specific proximal descent algorithm that can solve this problem, we first need to mention how the proximity operator associated with $\ell_{2,\infty}$ norm

$$\text{prox}_{\ell_{2,\infty}}(\boldsymbol{U}; \tau) = \underset{\check{\boldsymbol{U}} \in \mathbb{R}^{d \times r}}{\text{argmin}} \frac{1}{2} \|\boldsymbol{U} - \check{\boldsymbol{U}}\|_F^2 + \tau \|\check{\boldsymbol{U}}\|_{2,\infty}$$

can be computed. Fortunately, there is a simple procedure, based on bisection, that can be used for this purpose as detailed in the following pseudocode:

Algorithm 9 Compute the Proximity Operator $\widetilde{\boldsymbol{U}} = \text{prox}_{\ell_{2,\infty}}(\boldsymbol{U}; \tau)$

Input: Matrix $\boldsymbol{U} \in \mathbb{R}^{d \times r}$, positive scalar τ.

Find $\widetilde{t} \in [0, \|\boldsymbol{U}\|_{2,\infty}]$ which satisfies $\sum_{i=1}^{d} \left(\|\boldsymbol{U}_{i*}\|_2 - \widetilde{t} \right)_+ = \tau$ via bisection.

Set $\widetilde{\boldsymbol{U}}_{i*} = \left(1 - \frac{\widetilde{\rho}_i}{\|\boldsymbol{U}_{i*}\|_2}\right)_+ \boldsymbol{U}_{i*}$ where $\widetilde{\rho}_i = \left(\|\boldsymbol{U}_{i*}\|_2 - \widetilde{t} \right)_+$ for $i = 1, 2, \cdots, d$.

The following pseudocode describes iterations of the proximal descent method once instantiated to solve the matrix completion formulation in (4.21):

Algorithm 10 Proximal Descent Algorithm for Matrix Completion Application

Input: initial factors $\boldsymbol{U}_0$ and $\boldsymbol{V}_0$, target rank r, number of iterations T

Set $t = 0$, $\mu = 1/8M\sigma_1^2(\boldsymbol{W}_0)$, where $\boldsymbol{W}_0 := \begin{bmatrix} \boldsymbol{U}_0 \\ \boldsymbol{V}_0 \end{bmatrix}$

repeat

$\widetilde{\boldsymbol{U}}_{t+1} = \boldsymbol{U}_t - \frac{\mu}{p}\,\Pi_\Omega(\boldsymbol{U}_t\boldsymbol{V}_t^T - \boldsymbol{Y})\,\boldsymbol{V}_t - \frac{\mu}{4}\,\boldsymbol{U}_t\,(\boldsymbol{U}_t^T\boldsymbol{U}_t - \boldsymbol{V}_t^T\boldsymbol{V}_t)$

$\widetilde{\boldsymbol{V}}_{t+1} = \boldsymbol{V}_t - \frac{\mu}{p}\,\Pi_\Omega(\boldsymbol{V}_t\boldsymbol{U}_t^T - \boldsymbol{Y}^T)\,\boldsymbol{U}_t + \frac{\mu}{4}\,\boldsymbol{V}_t\,(\boldsymbol{U}_t^T\boldsymbol{U}_t - \boldsymbol{V}_t^T\boldsymbol{V}_t)$

$\boldsymbol{U}_{t+1} = \mathrm{prox}_{\ell_{2,\infty}}(\widetilde{\boldsymbol{U}}_{t+1}; \mu\tau)$

$\boldsymbol{V}_{t+1} = \mathrm{prox}_{\ell_{2,\infty}}(\widetilde{\boldsymbol{V}}_{t+1}; \mu\tau)$

$t = t + 1$

until $t = T$.

In order to apply the theoretical guarantees of the previous section in the context of matrix completion problem, we first need to ensure the Assumptions of section 4.2. Unfortunately, ensuring the RSC and RSM in the forms presented in Assumption 1 is not straightforward. However, other variants of these condition are verified in the literature, see e.g., Proposition 2 in Appendix E of [123]. Due to this issue, here we do not make statements about the theoretical implications of our study in the context of the matrix completion problem, and postpone such assertions until the required adjustments are made to ensure our framework encompasses other variants of Assumption 1.

4.6 Numerical Experiments

In this section, we experimentally evaluate the efficacy of the understudy factorization approach for low-rank estimation problems introduced in the past section.

4.6.1 Matrix Sensing

We investigate the performance of the proposed proximal descent algorithm for minimizing the low-rank matrix sensing problem (4.17). The entries of the factor matrices $\boldsymbol{U}^* \in \mathbb{R}^{d_1 \times r}$ and $\boldsymbol{V}^* \in \mathbb{R}^{d_2 \times r}$, with $d_1 = 128$, $d_2 = 256$, and $r = 10$, are set to i.i.d. samples drawn from $\mathcal{N}(0, 1)$. To ensure the equal-footedness of $\boldsymbol{U}^*$ and $\boldsymbol{V}^*$, they are normalized to have unit Frobenius norms. We have generated $n = 2(rd_1 + rd_2)$ noisy linear measurements of the form in (4.15). To generate the linear operator $\mathcal{A} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^n$,

we mimic the setup of [18, 118, 120] and use permuted and subsampled noiselets, due to their efficient implementation [141]. The noise vector $\epsilon \in \mathbb{R}^n$ is set to have i.i.d. entries drawn from $\mathcal{N}(0, \sigma^2)$. We chose a random initialization pair of factors $(\boldsymbol{U}_0, \boldsymbol{V}_0)$ drawn from the same distribution as for the ground-truth pair $(\boldsymbol{U}^*, \boldsymbol{V}^*)$. The step-size parameter is then set to $\mu = 1/5\|\boldsymbol{W}_0\|_2^2$, where $\boldsymbol{W}_0 = \begin{bmatrix} \boldsymbol{U}_0 \\ \boldsymbol{V}_0 \end{bmatrix}$.

Figure 4.1 illustrates the relative distance to optimality for the iterations of the algorithm. The three convergence curves, corresponding to three values of the noise variance , namely $\sigma = 10^{-5}, 10^{-4}$, and 10^{-3}, illustrate local linear convergence behavior. As the figure suggests, after a few thousand iterations, the curves floor at constant levels, which depend on the noise variance.

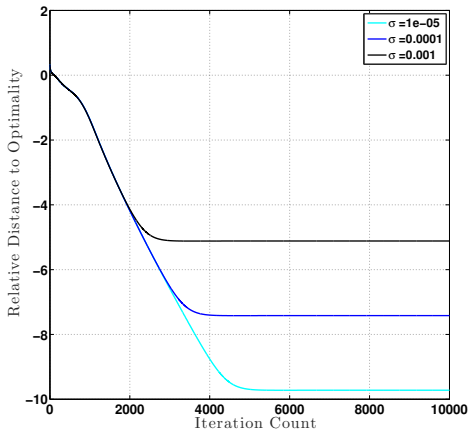


Figure 4.1: Convergence plots for three instances of the matrix sensing problem, with $\sigma = 10^{-5}, 10^{-4}, 10^{-3}$. The true matrix $\boldsymbol{X}^*$ is 128×256 dimensional and of rank $r = 10$.

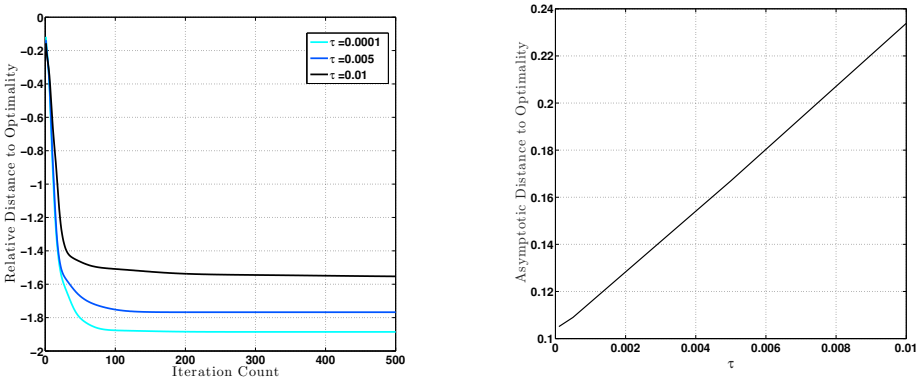
4.6.2 Sparse PCA

We also carried out experiments to validate the utility of Algorithm 8 for the problem of estimating a sparse covariance matrix $\boldsymbol{X}^*$, which is structured as described in subsection 4.5.3. The experiment's setting is as follows: the covariance matrix dimension, its rank r, and the row-sparsity level of the optimal factor $\boldsymbol{U}^* \in \mathbb{R}^{d \times r}$ were set to $d = 128$, $r = 10$, and $\alpha = 0.1$, respectively. The row-support of $\boldsymbol{U}^*$ was chosen uniformly at

random from all possible supports of size $\lfloor \alpha d \rfloor = 12$. The entries of all the 12 non-zero rows of $\boldsymbol{U}^*$ were set to i.i.d. samples of the standard Gaussian distribution $\mathcal{N}(0, 1)$ and the resulting factor matrix was normalized to have a unit Frobenius norm. Having generated $\boldsymbol{X}^*$ via $\boldsymbol{X}^* = \boldsymbol{U}^* \boldsymbol{U}^{*T}$, we set the noise variance σ to 10^{-5}, and drew $n = 2\alpha r d$ i.i.d. samples from the Gaussian distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \boldsymbol{X}^* + \sigma \boldsymbol{I}_d$.

To implement the steps of Algorithm 8, we first generated the initial factor matrix $\boldsymbol{U}_0$ with i.i.d. standard Gaussian entries and normalized it to have a unit Frobenius norm. The step size parameter was set to $\mu = 1/5\|\boldsymbol{U}_0\|_2$ and the regularization constant τ was swept from 10^{-4} to 10^{-2}.

Figure 4.2 shows the convergence behavior of Algorithm 8 for solving the problem. Panel (a) shows how the (relative) distance-to-optimality decays as the algorithm proceeds. For the sake of comparison, the convergence curves for four different values of the regularization constant τ are reported here. Consistent with the theory, the curves suggest a linear convergence decay followed by a constant asymptotic error, which changes by the value of the regularization constant τ. Panel (b) suggests a linear dependence between the asymptotic distance to optimality $\text{dist}(\widehat{\boldsymbol{U}}, \boldsymbol{U}^*)$, where $\widehat{\boldsymbol{U}}$ is the solution after 500 iterations of Algorithm 8 and the value of the regularization constant τ.



(a) (b)

Figure 4.2: Convergence Study of Algorithm 8 for solving the Sparse PCA problem. (a) Iterative decay of the relative distance to optimality. (b) The Asymptotic Procrustes distance to optimality as a function of the regularization constant τ.

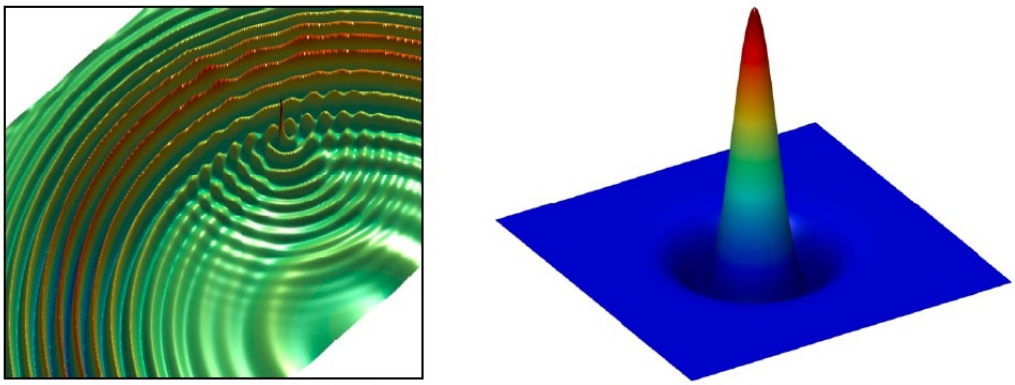
Chapter 5

Future Directions

Here, we describe a few directions to improve and extend the study of earlier chapters.

5.1 Estimation of Group Sparse Signals for Structural Diagnostics

The primary reason for adopting the identity matrix as the dictionary for the sparse anomalous component of wavefield measurements was the fact that, in the vicinity of defects, the wavefield is characterized by spatially localized features. The identity matrix will then provide the simplest dictionary to capture such morphological structure. Better choices for the sparse dictionary may better capture wavefield characteristics in the vicinity of defects. For instance, closely examining the wavefield measurements, in an extended neighborhood of a scatterer, reveals a pattern of concentric rings emanating from the scatterer with radially decaying amplitude; see panel (a) in Figure 5.1.



(a) (b)

Figure 5.1: Similarity of localized wavefield patterns with the Marr wavelet function. Panel (a) shows propagating wavefield incident upon a defect. Panel (b) shows a Marr wavelet that could be used to form a sparse dictionary.

This observation inspires us to search for a "morphologically germane" dictionary that more accurately captures this radial structure. One suitable example of such dictionaries is the (two dimensional) Marr wavelet [142], which is defined by the following equation

$$\mathbf{M}(r) = \frac{1}{\pi\sigma^4}\left(1 - \frac{r^2}{2\sigma^2}\right) e^{-r^2/2\sigma^2},$$

where the parameter σ controls how concentrated the wavelet is around its peak value, and r denotes the radial distance from the center. As illustrated in panel (b) of Figure 5.1, the Marr wavelet effectively displays one ring surrounding the peak.

We then adopt the following demixing procedure

$$\min_{\mathbf{b}_1\in\mathbb{R}^M, \mathbf{b}_2\in\mathbb{R}^N} \frac{1}{2}\|\mathbf{y} - \mathbf{X}_1\mathbf{b}_1 - \mathbf{X}_2\mathbf{b}_2\|_2^2 + \lambda_1\|\mathbf{b}_1\|_1 + \lambda_2\|\mathbf{b}_2\|_1$$

where λ_1 and λ_2 are regularization parameters, $\boldsymbol{X}_2$ is again a 2D-DCT dictionary and $\boldsymbol{X}_1$ is a dictionary whose columns are vectorized Marr wavelets, centered at different points of the structure's surface. To make the $\boldsymbol{X}_1$ dictionary expressive enough, (shifted) Marr wavelets for several values of the control parameter σ can be included, when creating the dictionary. Notice that to solve this problem numerically, we can use 2D-convolution for implementing the multiplication with the wavelet dictionary $\boldsymbol{X}_1$. Results of numerical

experiments with this approach are presented in [23], which indicate the supremacy of this method in identifying weak anomalies in challenging experimental configurations over the approach presented earlier in Chapter 2.

However, we need to still investigate the implications of the main theoretical results in the context of this new approach. Moreover, we recently came to know that, since the sparse dictionary $\boldsymbol{X}_1$ is essentially a concatenation of banded circulant matrices, the analytical results of [143] in the context of convolutional sparse coding, can be employed to analyze the new estimator.

5.2 Estimation of Low-Rank Models via Regularized Factorization

We can think of multiple directions to extend the presented study on regularized framework for low-rank matrix factorization problem. In particular, some required assumptions can be relaxed to strengthen the applicability of results. For instance, it is usful to extend the current analysis to the case where the matrix $\boldsymbol{X}^*$ is *approximately* rank-r. This will broaden the applicability of the result, since in many applications the exact rank r is unknown, or in some others, the ground-truth matrix is not even low-rank, due to noise and other inaccuracies. Doing so, we expect that an additional term, dependent on $\sum_{i=r+1}^{d} \sigma_i(\boldsymbol{X}^*)$, appears in the statistical error term expression of Theorem 4.3.1. The best approximation for the rank r is then the one that minimizes the following error expression

$$\eta + \sum_{i=r+1}^{d} \sigma_i(\boldsymbol{X}^*) = \sqrt{\frac{r}{m(m+M)}} \cdot \frac{\|\nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_2}{\sqrt{\sigma_1(\boldsymbol{X}^*)}} + \sum_{i=r+1}^{d} \sigma_i(\boldsymbol{X}^*).$$

The first term of the above expression can then be thought of as an estimation error term, while the second one is often referred to as the approximation error. Further discussions on this line are provided in [18, 118, 134, 140].

Another path for extending the framework of Chapter 4 from the computational perspective, is via analyzing *stochastic* variants of the proposed algorithms for solving similar factorization problems. As discussed in the introduction, the number of acquired measurements in many recent applications can be enormous. In those scenarios, solely

computing the gradient of the loss function $\mathcal{L}_n(\boldsymbol{X})$, with respect to $\boldsymbol{X}$, can be a huge burden. Algorithms that iteratively exploit (randomly selected) batches of the data will be then more efficient than those analyzed in this thesis. Efforts on this line are already made in, e.g., [144].

The local convergence study that was presented in Chapter 4 motivates the use of proximal descent-type methods for efficiently computing high-accuracy solutions to low-rank estimation problems. However, as the analysis reveals, the algorithms' performance highly depends on the quality of their initialization. Therefore, the current study will not be complete unless if it is accompanied by a discussion of some possible techniques for providing the presented algorithms with good initial points. Fortunately, the literature already contains many studies on such kind of techniques. For instance, [18, 122] propose easily-computable initializations that under certain (sometimes restrictive) assumptions will verifiably work. Beside those, there are other proposed methods for initializing factor matrices, which are based on one-time computing of the singular value decomposition of a rough estimate for $\boldsymbol{X}^*$. Examples of such studies can be found in [109, 131, 145, 146].

References

[1] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational mathematics*, 12(6):805–849, 2012.

[2] A. Agarwal. *Computational Trade-offs in Statistical Learning*. University of California, Berkeley, 2012.

[3] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.

[4] E. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, February 2006.

[5] S. A. Van De Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, pages 614–645, 2008.

[6] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[7] F. R. Bach. Consistency of the group lasso and multiple kernel learning. *The Journal of Machine Learning Research*, 9:1179–1225, 2008.

[8] L. Meier, S. Van De Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

[9] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[10] N. Srebro. Learning with matrix factorizations phd thesis, 2004.

[11] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[12] P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine learning research*, 7(Nov):2541–2563, 2006.

[13] E. Yang and P. K. Ravikumar. Dirty statistical models. In *Advances in Neural Information Processing Systems*, pages 611–619, 2013.

[14] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[15] M. Hong and Z.-Q. Luo. On the linear convergence of the alternating direction method of multipliers. *Mathematical Programming*, 162(1-2):165–199, 2017.

[16] R. Nishihara, L. Lessard, B. Recht, A. Packard, and M. I. Jordan. A general analysis of the convergence of admm. *arXiv preprint arXiv:1502.02009*, 2015.

[17] Y. Chen and M. J. Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

[18] D. Park, A. Kyrillidis, C. Caramanis, and S. Sanghavi. Finding low-rank solutions via non-convex matrix factorization, efficiently and provably. *arXiv preprint arXiv:1606.03168*, 2016.

[19] J. Druce, M. Kadkhodaie, J. D. Haupt, and S. Gonella. Structural diagnostics via anomaly-driven demixing of wavefield data. *In Proceedings of International Workshop on Structural Health Monitoring*, 2015.

[20] M. Kadkhodaie, S. Jain, J. D. Haupt, J. Druce, and S. Gonella. Locating rare and weak material anomalies by convex demixing of propagating wavefields. In

6th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), pages 373–376, 2015.

[21] J. Druce, S. Gonella, M. Kadkhodaie, S. Jain, and J. D. Haupt. Defect triangulation via demixing algorithms based on dictionaries with different morphological complexity. *Proceedings of 8th European Workshop on Structural Health Monitoring (IWSHM)*, 2016.

[22] M. K. Elyaderani, S. Jain, J. Druce, S. Gonella, and J. D. Haupt. Group-level support recovery guarantees for group lasso estimator. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 4366–4370. IEEE, 2017.

[23] J. Druce, S. Gonella, M. Kadkhodaie, S. Jain, and J. D. Haupt. Locating material defects via wavefield demixing with morphologically germane dictionaries. *Structural Health Monitoring*, 16(1):112–125, 2017.

[24] M. K. Elyaderani, S. Jain, J. Druce, S. Gonella, and J. D. Haupt. Improved support recovery guarantees for the group lasso with applications to structural health monitoring. *Submitted to IEEE Transactions on Signal Processing*, 2017.

[25] M. Kadkhodaie, K. Christakopoulou, M. Sanjabi, and A. Banerjee. Accelerated alternating direction method of multipliers. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 497–506. ACM, 2015.

[26] D. L. Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

[27] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

[28] S. Negahban, B. Yu, M. J. Wainwright, and P. K. Ravikumar. A unified framework for high-dimensional analysis of m -estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*, pages 1348–1356, 2009.

[29] J. A. Tropp. On the conditioning of random subdictionaries. *Applied and Computational Harmonic Analysis*, 25(1):1–24, 2008.

[30] E. J. Candès and Y. Plan. Near-ideal model selection by ℓ_1 minimization. *The Annals of Statistics*, 37(5A), 2009.

[31] M. F. Duarte and Y. C. Eldar. Structured compressed sensing: From theory to applications. *IEEE Transactions on Signal Processing*, 59(9):4053–4085, 2011.

[32] G. Obozinski, M. J. Wainwright, and M. I. Jordan. Support union recovery in high-dimensional multivariate regression. *The Annals of Statistics*, pages 1–47, 2011.

[33] N. S. Rao, B. Recht, and R. D. Nowak. Universal measurement bounds for structured sparse signal recovery. In *International Conference on Artificial Intelligence and Statistics*, pages 942–950, 2012.

[34] H. Liu and J. Zhang. Estimation consistency of the group lasso and its applications. In *International Conference on Artificial Intelligence and Statistics*, pages 376–383, 2009.

[35] P. Ravikumar, J. Lafferty, H. Liu, and L. Wasserman. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030, 2009.

[36] Y. Nardi and A. Rinaldo. On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal of Statistics*, 2:605–633, 2008.

[37] J. Huang and T. Zhang. The benefit of group sparsity. *The Annals of Statistics*, 38(4):1978–2004, 2010.

[38] K. Lounici, M. Pontil, S. Van De Geer, and A. B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, pages 2164–2204, 2011.

[39] A. S. Bandeira, E. Dobriban, D. G. Mixon, and W. F. Sawin. Certifying the restricted isometry property is hard. *arXiv preprint arXiv:1204.1580*, 2012.

[40] Y. C. Eldar, P. Kuppinger, and H. Bölcskei. Block-sparse signals: Uncertainty relations and efficient recovery. *Signal Processing, IEEE Transactions on*, 58(6):3042–3054, 2010.

[41] W. U. Bajwa, M. F. Duarte, and R. Calderbank. Conditioning of random block subdictionaries with applications to block-sparse recovery and regression. *IEEE Transactions on Information Theory*, 61(7):4060–4079, 2015.

[42] J-.B. Ihn and F-.K. Chang. Pitch-catch active sensing methods in structural health monitoring for aircraft structures. *Structural Health Monitoring*, 7(1):5–19, 2008.

[43] T. E. Michaels, J. E. Michaels, B. Mi, and M. Ruzzene. Damage detection in plate structures using sparse ultrasonic transducer arrays and acoustic wavefield imaging. In *AIP Conference Proceedings*, volume 760, page 938, 2005.

[44] Q. Wang and S. Yuan. Baseline-free imaging method based on new PZT sensor arrangements. *Journal of Intelligent Material Systems and Structures*, 20(14):1663–1673, 2009.

[45] C. Prada and M. Fink. Eigenmodes of the time reversal operator: A solution to selective focusing in multiple-target media. *Wave Motion*, 20(2):151–163, 1994.

[46] F. Foroozan and S. ShahbazPanahi. MUSIC-based array imaging in multi-modal ultrasonic non-destructive testing. In *Proceedings of IEEE Sensor Array and Multichannel Signal Processing Workshop*, 2012.

[47] V. Sharma, S. Hanagud, and M. Ruzzene. Damage index estimation in beams and plates using laser vibrometry. *AIAA Journal*, 44(4):919–923, 2006.

[48] T.E. Michaels, J.E. Michaels, and M. Ruzzene. Frequency-wavenumber domain analysis of guided wavefields. *Ultrasonics*, 51(4):452 – 466, 2011.

[49] M. Ruzzene. Frequency-wavenumber domain filtering for improved damage visualization. *Smart Materials and Structures*, 16(6):2116, 2007.

[50] Y. K. An, B. Park, and H. Sohn. Complete noncontact laser ultrasonic imaging for automated crack visualization in a plate. *Smart Materials and Structures*, 22(2):025022, 2013.

[51] C. Zhang, J. Qiu, and H. Ji. Laser ultrasonic imaging for impact damage visualization in composite structure. In *Proc. 7th European Workshop on Structural Health Monitoring*, Nantes, France, July 2014.

[52] J. A. Bucaro, A. J. Romano, P. Abraham, and S. Dey. Detection and localization of inclusions in plates using inversion of point actuated surface displacements. *Journal of the Acoustical Society of America*, 115(1):201 – 206, 2004.

[53] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), 2009.

[54] D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47(7):2845–2862, 2001.

[55] M. Elad, J. L. Starck, P. Querre, and D. L. Donoho. Simultaneous cartoon and texture image inpainting using morphological component analysis (mca). *Applied and Computational Harmonic Analysis*, 19(3):340–358, 2005.

[56] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference*, page iau005, 2014.

[57] R. Foygel and L. Mackey. Corrupted sensing: Novel guarantees for separating structured signals. *IEEE Transactions on Information Theory*, 60(2):1223–1247, 2014.

[58] R. Levine and J. E. Michaels. Block-sparse reconstruction and imaging for lamb wave structural health monitoring. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, 61(6):1006–1015, 2014.

[59] A. Golato, S. Santhanam, F. Ahmad, and M. G. Amin. Multimodal sparse reconstruction in guided wave imaging of defects in plates. *Journal of Electronic Imaging*, 25(4):043013–043013, 2016.

[60] D. L. Donoho and P. B. Stark. Uncertainty principles and signal recovery. *SIAM Journal on Applied Mathematics*, 49(3):906–931, 1989.

[61] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009.

[62] X. Lv, G. Bi, and C. Wan. The group lasso for stable recovery of block-sparse signal representations. *IEEE Transactions on Signal Processing*, 59(4):1371–1382, 2011.

[63] F. A. Potra and S. J. Wright. Interior-point methods. *Journal of Computational and Applied Mathematics*, 124(1):281–302, 2000.

[64] A. Juditsky and A. Nemirovski. First order methods for nonsmooth convex large-scale optimization, i: general purpose methods. *Optimization for Machine Learning*, pages 121–148, 2011.

[65] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[66] J. Duchi and Y. Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10(Dec):2899–2934, 2009.

[67] D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2(1):17–40, 1976.

[68] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[69] H. Wang and A. Banerjee. Online alternating direction method (longer version). *arXiv preprint arXiv:1306.3721*, 2013.

[70] T. Goldstein, B. O'Donoghue, S. Setzer, and R. Baraniuk. Fast alternating direction optimization methods. *SIAM Journal on Imaging Sciences*, 7(3):1588–1623, 2014.

[71] D. Goldfarb, S. Ma, and K. Scheinberg. Fast alternating linearization methods for minimizing the sum of two convex functions. *Mathematical Programming*, pages 1–34, 2013.

[72] Y. Nesterov. A method of solving a convex programming problem with convergence rate o (1/k2). In *Soviet Mathematics Doklady*, volume 27, pages 372–376, 1983.

[73] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[74] W. Fenchel. On conjugate convex functions. *Canad. J. Math*, 1(73-77), 1949.

[75] B. O'donoghue and E. Candes. Adaptive restart for accelerated gradient schemes. *Foundations of computational mathematics*, 15(3):715–732, 2015.

[76] N. Li, R. Jin, and Z.-H. Zhou. Top rank optimization in linear time. In *Advances in neural information processing systems*, pages 1502–1510, 2014.

[77] S. Boyd, C. Cortes, M. Mohri, and A. Radovanovic. Accuracy at the top. In *Advances in neural information processing systems*, pages 953–961, 2012.

[78] C. Burges. From ranknet to lambdarank to lambdamart: An overview. *Learning*, 11(23-581):81, 2010.

[79] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pages 129–136. ACM, 2007.

[80] T.-Y. Liu. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331, 2009.

[81] A. Rakotomamonjy. Sparse support vector infinite push. *arXiv preprint arXiv:1206.6432*, 2012.

[82] C. Rudin. The p-norm push: A simple convex ranking algorithm that concentrates at the top of the list. *Journal of Machine Learning Research*, 10(Oct):2233–2271, 2009.

[83] S. Agarwal. The infinite push: A new support vector ranking algorithm that directly optimizes accuracy at the absolute top of the list. In *Proceedings of the 2011 SIAM International Conference on Data Mining*, pages 839–850. SIAM, 2011.

[84] Ali Jalali. *Dirty statistical models*. PhD thesis, University of Texas, 2012.

[85] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[86] John Wright and Yi Ma. Dense error correction via l1-minimization. *CoRR*, abs/0809.0199, 2008.

[87] S. P. Kasiviswanathan, H. Wang, A. Banerjee, and P. Melville. Online l1-dictionary learning with application to novel document detection. In *Advances in Neural Information Processing Systems*, pages 2258–2266, 2012.

[88] N. Srebro, J. Rennie, and T. S. Jaakkola. Maximum-margin matrix factorization. In *Advances in neural information processing systems*, pages 1329–1336, 2005.

[89] J. D. Rennie and N. Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd international conference on Machine learning*, pages 713–719. ACM, 2005.

[90] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8), 2009.

[91] D. DeCoste. Collaborative prediction using ensembles of maximum margin matrix factorizations. In *Proceedings of the 23rd international conference on Machine learning*, pages 249–256. ACM, 2006.

[92] J. Bennett and S. Lanning. The netflix prize. In *Proceedings of KDD cup and workshop*, volume 2007, page 35. New York, NY, USA, 2007.

[93] A. A. Amini and M. J. Wainwright. High-dimensional analysis of semidefinite relaxations for sparse principal components. In *Information Theory, 2008. ISIT 2008. IEEE International Symposium on*, pages 2454–2458. IEEE, 2008.

[94] T. T. Cai, Z. Ma, and Y. Wu. Sparse pca: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(6):3074–3110, 2013.

[95] V.Q. Vu and J. Lei. Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics*, 41(6):2905–2947, 2013.

[96] M. Mardani, G. Mateos, and G. B. Giannakis. Dynamic of anomalography: tracking network anomalies via sparsity and low-rank. *IEEE Journal of Selected Topics in Signal Processing*, 7(1):50–66, 2013.

[97] E. J. Candes, Y. C. Eldar, T. Strohmer, and V. Voroninski. Phase retrieval via matrix completion. *SIAM review*, 57(2):225–251, 2015.

[98] E. J. Candes, X. Li, and M. Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[99] I. Waldspurger, A. d'Aspremont, and S. Mallat. Phase recovery, maxcut and complex semidefinite programming. *Mathematical Programming*, 149(1-2):47–81, 2015.

[100] M. Fazel. *Matrix rank minimization with applications*. PhD thesis, PhD thesis, Stanford University, 2002.

[101] E. Candes and B. Recht. Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119, 2012.

[102] S. Burer. Semidefinite programming in the space of partial positive semidefinite matrices. *SIAM Journal on Optimization*, 14(1):139–172, 2003.

[103] S. Burer and R. D. C. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

[104] S. Burer and R. D. C. Monteiro. Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming*, 103(3):427–444, 2005.

[105] R. Ge, J. D. Lee, and T. Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.

[106] S. Bhojanapalli, B. Neyshabur, and N. Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pages 3873–3881, 2016.

[107] R. Ge, C. Jin, and Y. Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. *arXiv preprint arXiv:1704.00708*, 2017.

[108] Z. Zhu, Q. Li, G. Tang, and M. B. Wakin. The global optimization geometry of nonsymmetric matrix factorization and sensing. *arXiv preprint arXiv:1703.01256*, 2017.

[109] P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674. ACM, 2013.

[110] M. Hardt. Understanding alternating minimization for matrix completion. In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*, pages 651–660. IEEE, 2014.

[111] M. Udell, C. Horn, R. Zadeh, and S. Boyd. Generalized low rank models. *Foundations and Trends® in Machine Learning*, 9(1):1–118, 2016.

[112] M. Hardt and E. Price. The noisy power method: A meta algorithm with applications. In *Advances in Neural Information Processing Systems*, pages 2861–2869, 2014.

[113] Z. Wang, Q. Gu, Y. Ning, and H. Liu. High dimensional expectation-maximization algorithm: Statistical optimization and asymptotic normality. *arXiv preprint arXiv:1412.8729*, 2014.

[114] S. Balakrishnan, M. J. Wainwright, and B. Yu. Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120, 2017.

[115] Q. Zheng and J. Lafferty. A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. In *Advances in Neural Information Processing Systems*, pages 109–117, 2015.

[116] S. Tu, R. Boczar, M. Simchowitz, M.Soltanolkotabi, and B. Recht. Low-rank solutions of linear matrix equations via procrustes flow. *arXiv preprint arXiv:1507.03566*, 2015.

[117] R. Sun and Z.-Q. Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016.

[118] S. Bhojanapalli, A. Kyrillidis, and S. Sanghavi. Dropping convexity for faster semi-definite optimization. *arXiv preprint*, 2015.

[119] T. Zhao, Z. Wang, and H. Liu. A nonconvex optimization framework for low rank matrix estimation. In *Advances in Neural Information Processing Systems*, pages 559–567, 2015.

[120] D. Park, A. Kyrillidis, S. Bhojanapalli, C. Caramanis, and S. Sanghavi. Provable non-convex projected gradient descent for a class of constrained matrix optimization problems. *stat*, 1050:4, 2016.

[121] X. Yi, D. Park, Y. Chen, and C. Caramanis. Fast algorithms for robust pca via gradient descent. In *Advances in neural information processing systems*, pages 4152–4160, 2016.

[122] L. Wang, X. Zhang, and Q. Gu. A unified computational and statistical framework for nonconvex low-rank matrix estimation. *arXiv preprint arXiv:1610.05275*, 2016.

[123] A. Agarwal, S. Negahban, and M. J. Wainwright. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. In *Advances in Neural Information Processing Systems*, pages 37–45, 2010.

[124] P.-L. Loh and M. J. Wainwright. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In *Advances in Neural Information Processing Systems*, pages 476–484, 2013.

[125] P. L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer, 2011.

[126] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

[127] S. Oymak, B. Recht, and M. Soltanolkotabi. Sharp time–data tradeoffs for linear inverse problems. *arXiv preprint arXiv:1507.04793*, 2015.

[128] E. J. Candes and Y. Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.

[129] Y.-K. Liu. Universal low-rank matrix recovery from pauli measurements. In *Advances in Neural Information Processing Systems*, pages 1638–1646, 2011.

[130] J. D. Lee, B. Recht, N. Srebro, J. Tropp, and R. R. Salakhutdinov. Practical large-scale optimization for max-norm regularization. In *Advances in Neural Information Processing Systems*, pages 1297–1305, 2010.

[131] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11(Jul):2057–2078, 2010.

[132] M. J. Wainwright. Structured regularizers for high-dimensional problems: Statistical and computational issues. *Annual Review of Statistics and Its Application*, 1:233–253, 2014.

[133] X. Yi and C. Caramanis. Regularized em algorithms: A unified framework and statistical guarantees. In *Advances in Neural Information Processing Systems*, pages 1567–1575, 2015.

[134] S. Negahban and M. J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, pages 1069–1097, 2011.

[135] I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of statistics*, pages 295–327, 2001.

[136] T. Cai, Z. Ma, and Y. Wu. Optimal estimation and rank detection for sparse spiked covariance matrices. *Probability theory and related fields*, 161(3-4):781–815, 2015.

[137] M. Laurent. Matrix completion problems. *Encyclopedia of Optimization*, 3:221–229, 2009.

[138] E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[139] B. Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(Dec):3413–3430, 2011.

[140] S. Negahban and M. J. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13(May):1665–1697, 2012.

[141] A. E. Waters, A. C. Sankaranarayanan, and R. Baraniuk. Sparcs: Recovering low-rank and sparse matrices from compressive measurements. In *Advances in neural information processing systems*, pages 1089–1097, 2011.

[142] Stéphane Mallat. *A wavelet tour of signal processing (2. ed.)*. Academic Press, 1999.

[143] V. Papyan, J. Sulam, and M. Elad. Working locally thinking globally: Theoretical guarantees for convolutional sparse coding. *IEEE Transactions on Signal Processing*, 2017.

[144] X. Zhang, L. Wang, and Q. Gu. Stochastic variance-reduced gradient descent for low-rank matrix recovery from linear measurements. *arXiv preprint arXiv:1701.00481*, 2017.

[145] P. Jain, R. Meka, and I. S. Dhillon. Guaranteed rank minimization via singular value projection. In *Advances in Neural Information Processing Systems*, pages 937–945, 2010.

[146] Z. Ma. Sparse principal component analysis and iterative thresholding. *The Annals of Statistics*, 41(2):772–801, 2013.

[147] M. Rudelson and R. Vershynin. Hanson-Wright inequality and sub-gaussian concentration. *Electron. Commun. Probab*, 18(0), 2013.

[148] A. S. Lewis. The convex analysis of unitarily invariant matrix functions. *Journal of Convex Analysis*, 2(1):173–183, 1995.

[149] P.-L. Loh and M. J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. In *Advances in Neural Information Processing Systems*, pages 2726–2734, 2011.

[150] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[151] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Appendix A

Proof of Results in Chapter 2

A.1 Overview of Proof Approach

In this section we describe an overview of our approach to prove the main Theorem 2.2.2. The details are provided in later sections .

Our analysis utilizes a basic result for characterizing the optimal solutions of the group Lasso problem (2.2). We state the result here as a lemma; its proof follows what are, by now, fairly standard methods in convex analysis so we omit it here[1] .

Lemma A.1.1. *A vector $\check{\boldsymbol{\beta}}$ solves problem* (2.2) *if and only if*

$$\boldsymbol{X}_{\mathcal{I}_g}^T \boldsymbol{X}(\check{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) - \boldsymbol{X}_{\mathcal{I}_g}^T \boldsymbol{w} + \lambda_g \check{\boldsymbol{z}}_{\mathcal{I}_g} = \boldsymbol{0}, \quad \forall\ g \in [G] \tag{A.1}$$

holds for some vector $\check{\boldsymbol{z}}$, whose elements satisfy

$$\begin{array}{rl} \check{\boldsymbol{z}}_{\mathcal{I}_g} = \frac{\check{\boldsymbol{\beta}}_{\mathcal{I}_g}}{\|\check{\boldsymbol{\beta}}_{\mathcal{I}_g}\|_2}, & \text{if } \check{\boldsymbol{\beta}}_{\mathcal{I}_g} \neq \boldsymbol{0} \\ \|\check{\boldsymbol{z}}_{\mathcal{I}_g}\|_2 \leq 1, & \text{otherwise} \end{array} . \tag{A.2}$$

If $\|\check{\boldsymbol{z}}_{\mathcal{I}_g}\|_2 < 1$ for all $g \notin \mathcal{G}(\check{\boldsymbol{\beta}})$ then any optimal solution $\check{\boldsymbol{\beta}}$ to (2.2) *satisfies $\check{\boldsymbol{\beta}}_{\mathcal{I}_g} = \boldsymbol{0}$ for all $g \notin \mathcal{G}(\check{\boldsymbol{\beta}})$; if, in addition, the matrix $\boldsymbol{X}_{\mathcal{S}(\check{\boldsymbol{\beta}})}^T \boldsymbol{X}_{\mathcal{S}(\check{\boldsymbol{\beta}})}$ is invertible, then $\check{\boldsymbol{\beta}}$ is the unique solution to* (2.2).

[1] A bit more specifically, we note that the proof of the lemma mirrors that of [61, Lemma 1], with appropriate changes arising from the group Lasso regularizer. We also note that an analogous result appears, for example, in [7, 35], among other works.

Note that the optimality condition (A.1) can be written in matrix form, as

$$\boldsymbol{X}^T\boldsymbol{X}(\check{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) - \boldsymbol{X}^T\boldsymbol{w} + \boldsymbol{\Lambda}\,\check{\boldsymbol{z}} = \boldsymbol{0}, \tag{A.3}$$

where $\boldsymbol{\Lambda}$ is the $p \times p$ diagonal matrix whose j-th diagonal entry $\Lambda_{j,j} = \lambda_{g(j)}$, where $g(j) = \{g \in [G] : j \in \mathcal{I}_g\}$. In other words, the diagonal elements of $\boldsymbol{\Lambda}$ are, for each index j, the regularization parameters associated with the group to which the corresponding element $\boldsymbol{\beta}_j$ of $\boldsymbol{\beta}$ belongs. We will find this alternative, equivalent formulation convenient in the analysis that follows.

The ultimate goal of this section is to find conditions under which the group-level support of $\check{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^*$ are identical, i.e. $\mathcal{G}(\check{\boldsymbol{\beta}}) = \mathcal{G}(\boldsymbol{\beta}^*)$. Our proof follows the so-called *Primal-Dual Witness* (PDW) technique utilized in [61] for the analysis of the Lasso problem and also in [32] for the analysis of the group Lasso problem arising in the context of multivariate regression. In our setting, a primal-dual certificate pair $(\check{\boldsymbol{\beta}}, \check{\boldsymbol{z}})$ is constructed according to the following steps:

1. We identify the solution of a *restricted* group Lasso problem over the true "group-level" support $\mathcal{S}_{\mathcal{G}}(\boldsymbol{\beta}^*)$. Specifically, we consider $\check{\boldsymbol{\beta}}_{\mathcal{S}^*_{\mathcal{G}}} \in \mathbb{R}^{d^*_{\mathcal{G}}}$ obtained according to

$$\check{\boldsymbol{\beta}}_{\mathcal{S}^*_{\mathcal{G}}} = \arg \min_{\boldsymbol{\beta}_{\mathcal{S}^*_{\mathcal{G}}} \in \mathbb{R}^{d^*_{\mathcal{G}}}} \frac{1}{2}||\boldsymbol{y} - \boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{\beta}_{\mathcal{S}^*_{\mathcal{G}}}||_2^2 + \sum_{g \in \mathcal{G}^*} \lambda_g \|\boldsymbol{\beta}_{\mathcal{I}_g}\|_2. \tag{A.4}$$

   Note that if $\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}}$ has full column-rank, there will be a unique vector $\check{\boldsymbol{\beta}}_{\mathcal{S}^*_{\mathcal{G}}}$ that solves (A.4).

2. We choose $\check{\boldsymbol{z}}_{\mathcal{S}^*_{\mathcal{G}}} \in \mathbb{R}^{d^*_{\mathcal{G}}}$ to be the optimal dual solution of the restricted group Lasso problem (A.4) such that the primal-dual pair $(\check{\boldsymbol{\beta}}_{\mathcal{S}^*_{\mathcal{G}}}, \check{\boldsymbol{z}}_{\mathcal{S}^*_{\mathcal{G}}})$ satisfies the optimality conditions of the restricted problem.

3. We set the "off group-level support" primal variable $\check{\boldsymbol{\beta}}_{(\mathcal{S}^*_{\mathcal{G}})^c}$ to be zero.

4. Finally, we solve for an "off group-level support" dual variable $\check{\boldsymbol{z}}_{(\mathcal{S}^*_{\mathcal{G}})^c} \in \mathbb{R}^{n - d^*_{\mathcal{G}}}$ which satisfies the optimality conditions for the full (unrestricted) group Lasso problem, as specified in (A.1) and (A.2), and identify conditions under which this vector satisfies $\|\check{\boldsymbol{z}}_{\mathcal{I}_g}\|_2 < 1$ for all $g \notin \mathcal{G}^*$.

Overall, the PDW approach can be viewed, essentially, as a method for evaluating the feasibility of one particular candidate solution $\check{\boldsymbol{\beta}}$ to the original group Lasso problem (2.2), constructed in a piece-wise manner. The first two steps identify conditions that the elements of the candidate solution must adhere to on the true "group-level" support. The strict dual feasibility condition ($\|\check{\boldsymbol{z}}_{\mathcal{I}_g}\|_2 < 1$ for all $g \notin \mathcal{G}^*$) in step 4 together with step 3 ensure that no "spurious" nonzero groups are present in $\check{\boldsymbol{\beta}}$. In other words, the success of the PDW approach outlined above ensures that the primal-dual pair $(\check{\boldsymbol{\beta}}, \check{\boldsymbol{z}})$ satisfies the optimality conditions of the general group Lasso problem (2.2) as given by Lemma A.1.1 and also meets the condition $\mathcal{G}(\check{\boldsymbol{\beta}}) \subseteq \mathcal{G}^*$. Moreover, it can be shown that if $\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}}$ has full column-rank, then $\check{\boldsymbol{\beta}}$ will be the unique optimal solution (see Lemma 2 in [32] for more details).

The last part of our analysis then relies on upper bounding the group-wise deviations between $\boldsymbol{\beta}^*_{\mathcal{S}^*_{\mathcal{G}}}$ and $\check{\boldsymbol{\beta}}_{\mathcal{S}^*_{\mathcal{G}}}$, from which we can identify conditions that the nonzero groups of the true parameter vector $\boldsymbol{\beta}^*$ must satisfy in order to ensure that no true signal groups are missed by the recovery procedure. Specifically, suppose that the condition

$$\|\boldsymbol{\beta}^*_{\mathcal{I}_g} - \check{\boldsymbol{\beta}}_{\mathcal{I}_g}\|_2 < \|\boldsymbol{\beta}^*_{\mathcal{I}_g}\|_2 \quad \text{for all } g \in \mathcal{G}^* \tag{A.5}$$

holds true. Then, it follows (essentially, by the triangle inequality) that $\check{\boldsymbol{\beta}}_{\mathcal{I}_g} \neq \boldsymbol{0}$, so that overall we have $\boldsymbol{\beta}^*_{\mathcal{I}_g} \neq \boldsymbol{0}$ implies $\check{\boldsymbol{\beta}}_{\mathcal{I}_g} \neq \boldsymbol{0}$. This is equivalent to $\mathcal{G}^* \subseteq \mathcal{G}(\check{\boldsymbol{\beta}})$; overall, the success of the PDW method *in addition to* a guarantee of the form (A.5) will ensure that $\mathcal{G}(\check{\boldsymbol{\beta}}) = \mathcal{G}^*$.

Throughout our analysis, we proceed under the assumption that the singular values of the block sub-dictionary $\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}}$ lie within the interval $[\sqrt{1/2}, \sqrt{3/2}\,]$. In other words, we assume the following probabilistic event

$$E_1 := \left\{ \|\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}} \boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}} - \boldsymbol{I}_{d^*_{\mathcal{G}} \times d^*_{\mathcal{G}}}\|_{2\to 2} \leq \frac{1}{2} \right\}, \tag{A.6}$$

happens to be true, in which $\mathcal{G}^*$ is randomly generated according to our previously-discussed statistical assumptions. This event, which is shown to happen with high probability under certain conditions, implies that the sub-dictionary $\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}}$ is well-conditioned and full column-rank (see Theorem 1 of [41] or its re-statement in Lemma A.2.6).

Strict Dual Feasibility Condition

According to Lemma A.1.1, the primal-dual pair $(\check{\boldsymbol{\beta}}, \check{\boldsymbol{z}})$, with $\check{\boldsymbol{\beta}}_{(\mathcal{S}^*_{\mathcal{G}})^c} = 0$, will be an optimal solution of the general group Lasso problem (2.2) if and only if

$$\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}} \boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}} (\check{\boldsymbol{\beta}}_{\mathcal{S}^*_{\mathcal{G}}} - \boldsymbol{\beta}^*_{\mathcal{S}^*_{\mathcal{G}}}) - \boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}} \boldsymbol{w} + \boldsymbol{\Lambda}_{\mathcal{S}^*_{\mathcal{G}}} \check{\boldsymbol{z}}_{\mathcal{S}^*_{\mathcal{G}}} = \boldsymbol{0}, \tag{A.7}$$

$$\boldsymbol{X}^T_{(\mathcal{S}^*_{\mathcal{G}})^c} \boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}} (\check{\boldsymbol{\beta}}_{\mathcal{S}^*_{\mathcal{G}}} - \boldsymbol{\beta}^*_{\mathcal{S}^*_{\mathcal{G}}}) - \boldsymbol{X}^T_{(\mathcal{S}^*_{\mathcal{G}})^c} \boldsymbol{w} + \boldsymbol{\Lambda}_{(\mathcal{S}^*_{\mathcal{G}})^c} \check{\boldsymbol{z}}_{(\mathcal{S}^*_{\mathcal{G}})^c} = \boldsymbol{0}, \tag{A.8}$$

where $\boldsymbol{\Lambda}_{\mathcal{S}^*_{\mathcal{G}}}$ and $\boldsymbol{\Lambda}_{(\mathcal{S}^*_{\mathcal{G}})^c}$ denote the sub-matrices of $\boldsymbol{\Lambda}$ obtained by sampling rows and columns at the locations in $\mathcal{S}^*_{\mathcal{G}}$ and $(\mathcal{S}^*_{\mathcal{G}})^c$, respectively, and $\check{\boldsymbol{z}}$ satisfies the subgradient condition (A.2), i.e. for every $g \in [G]$

$$\check{\boldsymbol{z}}_{\mathcal{I}_g} = \begin{array}{ll} \frac{\check{\boldsymbol{\beta}}_{\mathcal{I}_g}}{\|\check{\boldsymbol{\beta}}_{\mathcal{I}_g}\|_2}, & \text{if } \check{\boldsymbol{\beta}}_{\mathcal{I}_g} \neq \boldsymbol{0} \\ \|\check{\boldsymbol{z}}_{\mathcal{I}_g}\|_2 \leq 1, & \text{otherwise.} \end{array}, \tag{A.9}$$

The steps of the PDW construction method are designed such that the pair $(\check{\boldsymbol{\beta}}, \check{\boldsymbol{z}})$ constructed by this method meets the above conditions. To show this, note that (by Lemma A.1.1) the optimality condition for the restricted group Lasso problem (A.4) implies that (A.7) holds true for the support-restricted pair $(\check{\boldsymbol{\beta}}_{\mathcal{S}^*_{\mathcal{G}}}, \check{\boldsymbol{z}}_{\mathcal{S}^*_{\mathcal{G}}})$ constructed by steps 1 and 2 of the PDW method. Also, step 2 implies that the dual variable $\check{\boldsymbol{z}}_{\mathcal{S}^*_{\mathcal{G}}}$ will satisfy the group-wise condition (A.9). Since $\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}} \boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}}$ is invertible by the assumption that the event E_1 holds, we have that

$$\boldsymbol{\beta}^*_{\mathcal{S}^*_{\mathcal{G}}} - \check{\boldsymbol{\beta}}_{\mathcal{S}^*_{\mathcal{G}}} = (\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}} \boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}})^{-1} (\boldsymbol{\Lambda}_{\mathcal{S}^*_{\mathcal{G}}} \check{\boldsymbol{z}}_{\mathcal{S}^*_{\mathcal{G}}} - \boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}} \boldsymbol{w}). \tag{A.10}$$

Then, by step 3 of the PDW construction method, we take $\check{\boldsymbol{z}}_{(\mathcal{S}^*_{\mathcal{G}})^c}$ to be a vector that satisfies (A.8). This gives that

$$\check{\boldsymbol{z}}_{(\mathcal{S}^*_{\mathcal{G}})^c} = \boldsymbol{\Lambda}^{-1}_{(\mathcal{S}^*_{\mathcal{G}})^c} \boldsymbol{X}^T_{(\mathcal{S}^*_{\mathcal{G}})^c} \boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}} (\boldsymbol{\beta}^*_{\mathcal{S}^*_{\mathcal{G}}} - \check{\boldsymbol{\beta}}_{\mathcal{S}^*_{\mathcal{G}}}) + \boldsymbol{\Lambda}^{-1}_{(\mathcal{S}^*_{\mathcal{G}})^c} \boldsymbol{X}^T_{(\mathcal{S}^*_{\mathcal{G}})^c} \boldsymbol{w},$$

and we now aim to establish the strict dual feasibility condition, that $\|\check{\boldsymbol{z}}_{\mathcal{I}_g}\|_2 < 1$ for all $g \notin \mathcal{G}^*$. To that end, we note that for any fixed group index $g \notin \mathcal{G}^*$ we have

$$\begin{aligned} \check{\boldsymbol{z}}_{\mathcal{I}_g} &= \frac{1}{\lambda_g} \boldsymbol{X}^T_{\mathcal{I}_g} \left[\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}} (\boldsymbol{\beta}^*_{\mathcal{S}^*_{\mathcal{G}}} - \check{\boldsymbol{\beta}}_{\mathcal{S}^*_{\mathcal{G}}}) + \boldsymbol{w} \right] \\ &= \frac{1}{\lambda_g} \boldsymbol{X}^T_{\mathcal{I}_g} \left[\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}} (\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}} \boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}})^{-1} (\boldsymbol{\Lambda}_{\mathcal{S}^*_{\mathcal{G}}} \check{\boldsymbol{z}}_{\mathcal{S}^*_{\mathcal{G}}} - \boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}} \boldsymbol{w}) + \boldsymbol{w} \right] \\ &= \frac{1}{\lambda_g} \boldsymbol{X}^T_{\mathcal{I}_g} \left[\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}} (\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}} \boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}})^{-1} \boldsymbol{\Lambda}_{\mathcal{S}^*_{\mathcal{G}}} \check{\boldsymbol{z}}_{\mathcal{S}^*_{\mathcal{G}}} + \Pi_{(\mathcal{S}^*_{\mathcal{G}})^\perp} (\boldsymbol{w}) \right], \end{aligned}$$

where the second equality follows from the incorporation of (A.10), and the third one makes use of the definition $\Pi_{(\mathcal{S}^*_{\mathcal{G}})^\perp}(\boldsymbol{w}) := (\boldsymbol{I} - \boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}}(\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}})^{-1}\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}})\boldsymbol{w}$. Notice that by triangle inequality we will have that for any $g \notin \mathcal{G}^*$

$$\begin{aligned}\left\|\check{\boldsymbol{z}}_{\mathcal{I}_g}\right\|_2 \leq & \left\|\frac{1}{\lambda_g}\boldsymbol{X}^T_{\mathcal{I}_g}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}}(\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}})^{-1}\boldsymbol{\Lambda}_{\mathcal{S}^*_{\mathcal{G}}}\check{\boldsymbol{z}}_{\mathcal{S}^*_{\mathcal{G}}}\right\|_2 \\ & + \left\|\frac{1}{\lambda_g}\boldsymbol{X}^T_{\mathcal{I}_g}\Pi_{(\mathcal{S}^*_{\mathcal{G}})^\perp}(\boldsymbol{w})\right\|_2. \end{aligned} \tag{A.11}$$

Then sufficient conditions for $\|\check{\boldsymbol{z}}_{\mathcal{I}_g}\|_2 < 1, g \notin \mathcal{G}^*$, can be obtained by bounding the terms on the right hand-side of the above inequality. Note that bounding the first term on the right-hand side may proceed using any of a number of strategies. One (potentially loose) approach would entail applying the triangle inequality, utilizing standard matrix norm inequalities, and exploiting only magnitude information about the vectors $\check{\boldsymbol{z}}_{\mathcal{I}'_g}$ (e.g., that $\|\check{\boldsymbol{z}}_{\mathcal{I}'_g}\|_2 \leq 1$ for all $g' \in \mathcal{G}^*$). This strategy would lead us to the proof of Theorem 2.2.1, which relies on the following Lemma (proved in Appendix A.2) to bound the first term in the upper bound of (A.11).

Lemma A.1.2. *Suppose the group-level support $\mathcal{G}^*$, with $|\mathcal{G}^*| = s$, is fixed such that the event E_1 in* (A.6) *occurs. Then if*

$$s \leq \left(\frac{\lambda_{\min}}{\lambda_{\max}}\right) \cdot \frac{1}{4\,\mu_B\,(\boldsymbol{X})}, \tag{A.12}$$

it will follow that $\left\|\frac{1}{\lambda_g}\boldsymbol{X}^T_{\mathcal{I}_g}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}}(\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}})^{-1}\boldsymbol{\Lambda}_{\mathcal{S}^*_{\mathcal{G}}}\check{\boldsymbol{z}}_{\mathcal{S}^*_{\mathcal{G}}}\right\|_2 < \frac{1}{2}$ for any $g \notin \mathcal{G}^*$.

Alternatively, we can adopt a slightly different strategy that exploits our statistical assumptions, i.e. that the "direction" vectors $\boldsymbol{\beta}^*_{\mathcal{I}_g}/\|\boldsymbol{\beta}^*_{\mathcal{I}_g}\|_2$ associated with every nonzero block of $\boldsymbol{\beta}^*$ indexed by $g \in \mathcal{G}^*$ are random, and statistically independent. To this aim, we need to express the elements of the vector $\check{\boldsymbol{z}}_{\mathcal{S}^*_{\mathcal{G}}}$ (or more specifically, its individual blocks) in terms of the "direction' vectors associated with the corresponding nonzero blocks of the true vector $\boldsymbol{\beta}^*_{\mathcal{S}^*_{\mathcal{G}}}$. In particular, the following Lemma, which is proved in the Appendix A.2, states that every block of $\check{\boldsymbol{z}}_{\mathcal{S}^*_{\mathcal{G}}}$ can be expressed as the sum of the corresponding true direction vector and a bounded perturbation vector.

Lemma A.1.3. *Suppose that the group-level support $\mathcal{G}^*$ is fixed such that the event E_1 occurs. Defining $\boldsymbol{h}_{g'} := \check{\boldsymbol{\beta}}_{\mathcal{I}_{g'}} - \boldsymbol{\beta}^*_{\mathcal{I}_{g'}}$ for every $g' \in \mathcal{G}^*$, it follows that*

$$\|\boldsymbol{h}_{g'}\|_2 \leq \|\boldsymbol{X}^T_{\mathcal{I}_{g'}}\boldsymbol{w}\|_2 + \lambda_{g'} + \|\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{w}\|_2 + \|\boldsymbol{\lambda}_{\mathcal{G}^*}\|_2, \tag{A.13}$$

where $\boldsymbol{\lambda}_{\mathcal{G}^*} \in \mathbb{R}^{d^*_{\mathcal{G}}}$ *is a vector whose entries are the elements* $\{\lambda_{g'}\}_{g' \in \mathcal{G}^*}$. *Moreover, the blocks of the dual vector over the true support set* $\mathcal{G}^*$ *can be expressed as*

$$\check{\boldsymbol{z}}_{\mathcal{I}_{g'}} = \frac{\boldsymbol{\beta}^*_{\mathcal{I}_{g'}}}{\|\boldsymbol{\beta}^*_{\mathcal{I}_{g'}}\|_2} + \boldsymbol{u}_{g'}. \tag{A.14}$$

Further, if $\|\boldsymbol{h}_{g'}\|_2 \leq \frac{1}{2}\|\boldsymbol{\beta}^*_{g'}\|_2$ *for* $g' \in \mathcal{G}^*$, *then* $\|\boldsymbol{u}_{g'}\|_2 \leq 4\|\boldsymbol{h}_{g'}\|_2 / \|\boldsymbol{\beta}^*_{\mathcal{I}_{g'}}\|_2$, .

As the Lemma asserts for each $g' \in \mathcal{G}^*$, which satisfies $\|\check{\boldsymbol{\beta}}_{\mathcal{I}_{g'}} - \boldsymbol{\beta}^*_{\mathcal{I}_{g'}}\|_2 \leq \frac{1}{2}\|\boldsymbol{\beta}^*_{g'}\|_2$, we can write $\check{\boldsymbol{z}}_{\mathcal{I}_{g'}} = \left(\boldsymbol{\beta}^*_{\mathcal{I}_{g'}} / \|\boldsymbol{\beta}^*_{\mathcal{I}_{g'}}\|_2\right) + \boldsymbol{u}_{g'}$, where the norm of $\boldsymbol{u}_{g'}$ can be controlled in terms of the norm of the difference $\check{\boldsymbol{\beta}}_{\mathcal{I}_{g'}} - \boldsymbol{\beta}^*_{\mathcal{I}_{g'}}$. We can also express the condition (A.14) in the following compact form over the entire support $\mathcal{S}^*_{\mathcal{G}}$

$$\check{\boldsymbol{z}}_{\mathcal{S}^*_{\mathcal{G}}} = \overline{\boldsymbol{\beta}^*_{\mathcal{S}^*_{\mathcal{G}}}} + \boldsymbol{u}_{\mathcal{S}^*_{\mathcal{G}}},$$

where $\overline{\boldsymbol{\beta}^*_{\mathcal{S}^*_{\mathcal{G}}}}$ is obtained by concatenating the direction vectors $\boldsymbol{\beta}^*_{\mathcal{I}_{g'}} / \|\boldsymbol{\beta}^*_{\mathcal{I}_{g'}}\|_2$ for all $g' \in \mathcal{G}^*$ and similarly $\boldsymbol{u}_{\mathcal{S}^*_{\mathcal{G}}}$ is the result of stacking all $\{\boldsymbol{u}_{g'}\}_{g' \in \mathcal{G}^*}$. With this, we have overall that for each $g \notin \mathcal{G}^*$, we can write

$$\begin{aligned} \|\check{\boldsymbol{z}}_{\mathcal{I}_g}\|_2 &\leq \frac{1}{\lambda_g} \left\| \boldsymbol{X}^T_{\mathcal{I}_g} \boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}} (\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}} \boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}})^{-1} \boldsymbol{\Lambda}_{\mathcal{S}^*_{\mathcal{G}}} \overline{\boldsymbol{\beta}^*_{\mathcal{S}^*_{\mathcal{G}}}} \right\|_2 \\ &+ \frac{1}{\lambda_g} \left\| \boldsymbol{X}^T_{\mathcal{I}_g} \boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}} (\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}} \boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}})^{-1} \boldsymbol{\Lambda}_{\mathcal{S}^*_{\mathcal{G}}} \boldsymbol{u}_{\mathcal{S}^*_{\mathcal{G}}} \right\|_2 \\ &+ \frac{1}{\lambda_g} \left\| \boldsymbol{X}^T_{\mathcal{I}_g} \Pi_{(\mathcal{S}^*_{\mathcal{G}})^\perp}(\boldsymbol{w}) \right\|_2 . \end{aligned} \tag{A.15}$$

As noted before if we establish that the right-hand side is strictly less than 1 for each $g \notin \mathcal{G}^*$, then no "spurious" groups will be identified by the group Lasso procedure. This strategy will lead us to the proof of Theorem 2.2.2, which entails concentration theory arguments to control the terms in the above upper bound. Through the rest of the current section we will describe the proof of this theorem and relegate the (simpler) proof of Theorem 2.2.1 to Appendix A.3.

A.2 Proof of Theorem 2.2.2

Here we demonstrate that under the assumptions of Theorem 2.2.2, the dual variable $\check{\boldsymbol{z}}$ constructed by the PDW technique satisfies the strict dual feasibility condition,

i.e. $\|\check{\boldsymbol{z}}_{\mathcal{I}_g}\|_2 < 1$, $\forall g \in \mathcal{G}^*$, with high probability. As mentioned in the previous section, proceeding with the PDW technique will require the sub-dictionary $\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}}$ be well-conditioned. In other words, our analysis will be conditioned on that the true group-level support $\mathcal{G}^*$, which is randomly chosen as according to the statistical assumption $\mathbf{M_1}$ of our data model, is selected such that the event $E_1 = \left\{\|\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}} - \boldsymbol{I}_{d^*_{\mathcal{G}}\times d^*_{\mathcal{G}}}\|_{2\to 2} \leq \frac{1}{2}\right\}$ holds true (see Lemma A.2.6 for conditions under which this assumption is valid with high probability). In addition to E_1, the analysis is also conditioned on the following event being true

$$E_2 := \left\{\|\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{(\mathcal{S}^*_{\mathcal{G}})^c}\|_{B,1} = \max_{g\notin\mathcal{G}^*}\|\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{\mathcal{I}_g}\|_{2\to 2} \leq \gamma\right\},$$

for the specific choice of

$$\gamma = \frac{\lambda_{\min}}{\lambda_{\max}}\cdot\frac{c_4}{\sqrt{d_{\max}\cdot\log p}},$$

where c_4 is a positive constant (independent of problem parameters) that satisfies $c_4 \leq 1/8\sqrt{2(1+4\log 2)}$, as required later in the proof. Here also the randomness is over the choice of the true group-level support $\mathcal{G}^*$. Given this event, which is shown to hold with high probability by Lemma A.2.7 in Appendix A.2, we are ensured that blocks over the true group-level support $\mathcal{G}^*$ are distinct enough from the remaining blocks.

Conditioned on the events E_1 and E_2, to prove the strict dual feasibility condition we will show that for any $g \notin \mathcal{G}^*$, each of the terms appearing in the upper bound can be further bounded (e.g. by the constant $1/4$) under the assumptions $\mathbf{M_2}$ and $\mathbf{M_3}$ of our statistical model. To better organize the proof, we also define the three following probabilistic events, which correspond to the terms of the upper bound in (A.15):

$$E_3 := \left\{\left\|\boldsymbol{X}^T_{\mathcal{I}_g}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}}(\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}})^{-1}\boldsymbol{\Lambda}_{\mathcal{S}^*_{\mathcal{G}}}\overline{\boldsymbol{\beta}^*_{\mathcal{S}^*_{\mathcal{G}}}}\right\|_2 \leq \frac{\lambda_g}{4},\ \forall g\notin\mathcal{G}^*\right\}$$

$$E_4 := \left\{\left\|\boldsymbol{X}^T_{\mathcal{I}_g}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}}(\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}})^{-1}\boldsymbol{\Lambda}_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{u}_{\mathcal{S}^*_{\mathcal{G}}}\right\|_2 \leq \frac{\lambda_g}{4},\ \forall g\notin\mathcal{G}^*\right\}$$

$$E_5 := \left\{\|\boldsymbol{X}^T_{\mathcal{I}_g}\Pi_{(\mathcal{S}^*_{\mathcal{G}})^\perp}(\boldsymbol{w})\|_2 \leq \frac{\lambda_g}{4},\ \forall g\notin\mathcal{G}^*\right\}.$$

The Lemmata A.2.1, A.2.3, and A.2.4 that come through this sub-section will describe conditions under which these events hold with high probability. Having shown such bounds, the strict dual feasibility condition will naturally follow with high probability using a simple union bound argument.

Proof of E_3

The following Lemma provides a condition under which the event E_3, which corresponds to the first term of the upper bound (A.15), will be small. The proof of the Lemma is moved to Appendix A.2.

Lemma A.2.1. *Suppose the group-level support $\mathcal{G}^*$ is given such that the events E_1 and E_2 hold for the sub-dictionary $\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}}$ of the dictionary $\boldsymbol{X} \in \mathbb{R}^{n\times p}$. Then assuming $\boldsymbol{\beta}^*_{\mathcal{S}^*_{\mathcal{G}}}$ is a random vector generated according to the statistical model assumptions $\mathbf{M_2}$ and $\mathbf{M_3}$ described earlier we will have that*

$$\Pr\left(\bigcup_{g\notin\mathcal{G}^*}\left\|\boldsymbol{X}^T_{\mathcal{I}_g}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}}(\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}})^{-1}\boldsymbol{\Lambda}_{\mathcal{S}^*_{\mathcal{G}}}\overline{\boldsymbol{\beta}^*_{\mathcal{S}^*_{\mathcal{G}}}}\right\|_2 > \frac{\lambda_g}{4}\right) \tag{A.16}$$

is less than or equal to $\eta = 2p^{-4\,\log 2}$.

Proof of E_4

Next, we derive conditions under which the event E_4, which is associated with the second term of the upper bound in (A.15), holds as well. In order to show this, we will have to leverage Lemma A.1.3 to control the size of the $\{\boldsymbol{u}_{g'}\}_{g'\in\mathcal{G}^*}$ vectors and in turn the size of the $\{\boldsymbol{h}_{g'}\}_{g'\in\mathcal{G}^*}$ vectors. Since the upper bound in (A.13) for $\boldsymbol{h}_{g'}$, $g'\in\mathcal{G}^*$, is in terms of the noise-related terms $\|\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{w}\|_2$ and $\|\boldsymbol{X}^T_{\mathcal{I}_{g'}}\boldsymbol{w}\|_2$, we will start by providing probabilistic bounds on these quantities.

Lemma A.2.2. *Suppose the group-level support $\mathcal{G}^*$ is fixed. Moreover, assume $\boldsymbol{w}\sim\mathcal{N}(0,\sigma^2\boldsymbol{I}_{n\times n})$. Then there exists a universal constant $c_7\in(3,7)$ for which it holds: for any $t\geq 1$ and $\epsilon\geq\sqrt{(1+\mu_I(\boldsymbol{X}))\log\left(p^t\,|\mathcal{G}^*|\right)/c_7\,d_{\min}}$, the following events*

- $\|\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{w}\|_2\leq\sigma(1+\epsilon)\sqrt{d^*_{\mathcal{G}}}$
- $\bigcap_{g'\in\mathcal{G}^*}\left\{\|\boldsymbol{X}^T_{\mathcal{I}_{g'}}\boldsymbol{w}\|_2\leq\sigma(1+\epsilon)\sqrt{d_{g'}}\right\}$

hold simultaneously with probability at least $1-2\,p^{-t}-2\,\exp\left(-c_7\epsilon^2 d^*_{\mathcal{G}}/2\right)$.

The proof of this Lemma is brought in Appendix A.2. Now, by using this lemma together with Lemma A.1.3 we obtain the following result on the norm of the difference vectors $\boldsymbol{h}_{g'}=\check{\boldsymbol{\beta}}_{\mathcal{I}_{g'}}-\boldsymbol{\beta}^*_{\mathcal{I}_{g'}}$ for $g'\in\mathcal{G}^*$.

Corollary A.2.1. *Suppose the group-level support $\mathcal{G}^*$ is given such that the event E_1 holds. Furthermore, assume that $\boldsymbol{w} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_{n\times n})$. There exists a universal finite constant $c_7 > 0$ for which the following holds: for any $t \geq 1$ and*

$$\epsilon \geq \sqrt{(1+\mu_I(\boldsymbol{X})) \cdot \log\left(p^t\,|\mathcal{G}^*|\right)/c_7\, d_{\min}},$$

the following inequality

$$\|\boldsymbol{h}_{g'}\|_2 \leq \sigma(1+\epsilon)\left(\sqrt{d^*_{\mathcal{G}}} + \sqrt{d_{g'}}\right) + \lambda_{g'} + \|\boldsymbol{\lambda}_{\mathcal{G}^*}\|_2 \tag{A.17}$$

holds for every $g' \in \mathcal{G}^*$ with probability at least $1 - 2\,p^{-t} - 2\exp\left(-c_7\epsilon^2 d^*_{\mathcal{G}}/2\right)$.

Leveraging the above Corollary, we are able to bound the norm of the second term of the upper bound in (A.15).

Lemma A.2.3. *Suppose the group-level support $\mathcal{G}^*$ is given such that both events E_1 and E_2 hold for the sub-dictionary $\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}}$ of $\boldsymbol{X}$. Furthermore, assume $\boldsymbol{w} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_{n\times n})$ and that $\|\boldsymbol{\beta}^*_{\mathcal{I}_{g'}}\|_2 \geq t_2 \|\boldsymbol{h}_{g'}\|_2$ holds for all $g' \in \mathcal{G}^*$, for some value of t_2 satisfying*

$$t_2 \geq \max\left\{2, c_8\sqrt{\frac{|\mathcal{G}^*|}{d_{\max}\,\log p}}\right\}, \tag{A.18}$$

where c_8 is a universal constant which satisfies $c_8 \geq 4/\sqrt{2(1+4\log 2)}$, then we have that for all $g \notin \mathcal{G}^*$

$$\frac{1}{\lambda_g}\left\|\boldsymbol{X}^T_{\mathcal{I}_g}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}}(\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}})^{-1}\boldsymbol{\Lambda}_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{u}_{\mathcal{S}^*_{\mathcal{G}}}\right\|_2 \leq \frac{1}{4}. \tag{A.19}$$

Putting the result of Corollary A.2.1 together with the above Lemma and also setting $c_8 = 2 > 4/\sqrt{2(1+4\log 2)}$ will immediately obtain the next corollary.

Corollary A.2.2. *Suppose the group-level support $\mathcal{G}^*$ is given such that both events E_1 and E_2 hold for the sub-dictionary $\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}}$ of $\boldsymbol{X}$. Furthermore, assume $\boldsymbol{w} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_{n\times n})$, ϵ is set as in Theorem 2.2.1, and for all $g' \in \mathcal{G}^*$*

$$\begin{aligned}\|\boldsymbol{\beta}^*_{\mathcal{I}_{g'}}\|_2 \geq \max&\left\{2, 2\sqrt{\frac{|\mathcal{G}^*|}{d_{\max}\cdot\log p}}\right\} \\ &\cdot\left\{\sigma(1+\epsilon)\left(\sqrt{d^*_{\mathcal{G}}} + \sqrt{d_{g'}}\right) + \lambda_{g'} + \|\boldsymbol{\lambda}_{\mathcal{G}^*}\|_2\right\},\end{aligned}$$

then the following inequality

$$\left\|\boldsymbol{X}^T_{\mathcal{I}_g}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}}(\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}})^{-1}\boldsymbol{\Lambda}_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{u}_{\mathcal{S}^*_{\mathcal{G}}}\right\|_2 \leq \frac{\lambda_g}{4}, \tag{A.20}$$

holds with probability at least $1 - 2\,p^{-t} - 2\exp\left(-c_7\epsilon^2 d^*_{\mathcal{G}}/2\right)$.

Proof of E_5

Finally, we can show that, with high probability, the noise-dependent term of the upper bound in (A.15), i.e. $\frac{1}{\lambda_g}\|\boldsymbol{X}_{\mathcal{I}_g}^T \Pi_{(\mathcal{S}_{\mathcal{G}}^*)^\perp}(\boldsymbol{w})\|_2$, is also smaller than $1/4$ simultaneously for all $g \notin \mathcal{G}^*$. The proof of this Lemma is moved to Appendix A.2.

Lemma A.2.4. *Let $\boldsymbol{X}$ be as above with $\mathcal{S}_{\mathcal{G}}^*$ fixed, and let $\boldsymbol{w} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_{n\times n})$. There exists a universal finite constant $c_7 > 0$ for which the following holds: for any $t \geq 1$ and $\epsilon \geq \sqrt{(1+\mu_I(\boldsymbol{X}))\cdot \log\left(p^t\left(G - |\mathcal{G}^*|\right)\right)/c_7\, d_{\min}}$ if*

$$\lambda_g \geq 4\sigma(1+\epsilon)\sqrt{d_g} \quad \textit{for all } g \notin \mathcal{G}^*, \tag{A.21}$$

then

$$\Pr\left(\bigcup_{g\notin\mathcal{G}^*}\left\{\frac{1}{\lambda_g}\|\boldsymbol{X}_{\mathcal{I}_g}^T \Pi_{(\mathcal{S}_{\mathcal{G}}^*)^\perp}(\boldsymbol{w})\|_2 > \frac{1}{4}\right\}\right) \leq 2\,p^{-t}. \tag{A.22}$$

Completing the Proof of Theorem 2.2.2

Now we can put all the proof ingredients together to complete the overall argument. Let E denote the event that the group-level support $\mathcal{G}^*$ is exactly recovered via solving the group Lasso problem (2.2). As explained in sub-section A.2, to ensure E happens our approach is to first find conditions that guarantee E_1 and E_2 hold true; then conditioned on those two events, we impose extra assumptions to ensure E_3, E_4 and E_5 occur as well. Using a union bound then implies the following upper bound[2]

$$\begin{aligned}\Pr(E^c) \leq \Pr(E_1^c) + \Pr(E_2^c) + \Pr\left(E_3^c | E_1 \cap E_2\right) \\ + \Pr\left(E_4^c | E_1 \cap E_2\right) + \Pr\left(E_5^c | E_1 \cap E_2\right).\end{aligned}$$

The rest of the proof reviews conditions under which the probability terms on the right-hand side of the above inequality are bounded. First, by Lemma A.2.6, we know that if there exist positive constants c_0 and c_1 such that $\mu_I(\boldsymbol{X}) \leq c_0$, $\mu_B(\boldsymbol{X}) \leq \frac{c_1}{\log p}$, and

$$s \leq \min\left\{\frac{c_2}{\mu_B^2(\boldsymbol{X})\log p}, \frac{c_3\, G}{\|\boldsymbol{X}\|_{2\to 2}^2 \log p}\right\}, \tag{A.23}$$

[2] To show the inequality notice that for two probabilistic events A and B, we can write $A^c \cup B^c = A^c \cup (B^c \cap A)$. Setting $A = E_1 \cap E_2$ and $B = E_3 \cap E_4 \cap E_5$ and using the fact that $\Pr(A^c \cup B^c) \leq \Pr(A^c) + \Pr(B^c \cap A) \leq \Pr(A^c) + \Pr(B^c|A)$ would conclude the inequality.

where c_2 and c_3 are such that

$$(48c_1 + 6\sqrt{2(c_2+c_3)} + 2c_3 + 3c_0) \leq \frac{1}{4}, \tag{A.24}$$

then $\Pr(E_1^c) \leq 2p^{-4\log 2}$. Notice that the relationship in (A.24) requires c_0 and c_1 to be such that $48\,c_1 + 3\,c_0 < \frac{1}{4}$. Given this condition, a valid choice for c_2 and c_3 that satisfies (A.24) is

$$c_2 = c_3 \leq \frac{1}{14}\left(\frac{1}{4} - 3c_0 - 48c_1\right).$$

Second, utilizing Lemma A.2.7, with $\lambda_g = 4\sigma(1+\epsilon)\sqrt{d_g}$, $\gamma := \frac{\lambda_{\min}}{\lambda_{\max}} \cdot \frac{c_4}{\sqrt{d_{\max}\cdot \log p}}$, and $c_4 = 1/8\sqrt{2(1+4\log 2)}$, it will follow that as long as

$$\mu_B(\boldsymbol{X}) \leq \sqrt{\frac{d_{\min}}{d_{\max}^2}}\frac{c_5}{\log p} \;\text{ and }\; s \leq \frac{d_{\min}}{d_{\max}^2}\frac{c_6^2}{\mu_B^2(\boldsymbol{X})\,\log p}, \tag{A.25}$$

with constants c_5 and c_6 chosen such that $4\sqrt{2}c_5 + c_6 \leq \frac{c_4}{2}$, then $\Pr(E_2^c) \leq 2p^{-4\log 2}$. In particular,

$$c_4 = 0.04,\; c_5 = 0.004,\; c_6 = 0.02$$

are viable choices. To compactly express the upper bounds in (A.23) and (A.25) on the maximum recoverable group-sparsity level s, notice that since $d_{\min}/d_{\max}^2 \leq 1$, therefore

$$\begin{aligned} s &\leq \frac{d_{\min}}{d_{\max}^2}\cdot\frac{\min\{c_6^2, c_2\}}{\mu_B^2(\boldsymbol{X})\log p} \\ &\leq \min\left\{\frac{d_{\min}}{d_{\max}^2}\cdot\frac{c_6^2}{\mu_B^2(\boldsymbol{X})\log p}, \frac{c_2}{\mu_B^2(\boldsymbol{X})\log p}\right\}, \end{aligned}$$

together with $s \leq c_3\,G\,/\left(\|\boldsymbol{X}\|_{2\to 2}^2 \cdot \log p\right)$ guarantees the requirements on s are met. Similarly, $\sqrt{d_{\min}/d_{\max}^2} \leq 1$ implies that imposing

$$\mu_B(\boldsymbol{X}) \leq \sqrt{\frac{d_{\min}}{d_{\max}^2}}\cdot\frac{c_5}{\log p},$$

will ensure the block coherence parameter meets $\mu_B(\boldsymbol{X}) \leq c_1/\log p$ for $c_1 = c_5 = 0.004$.

Third, Lemma A.2.1 implies that $\Pr\left(E_3^c | E_1 \cap E_2\right) \leq 2p^{-4\log 2}$. Forth, Corollary A.2.2, with $\lambda_g = 4\sigma(1+\epsilon)\sqrt{d_g}$ and $t = 4\log 2$, implies that as long as for

$$\epsilon \geq \sqrt{\frac{(1+\mu_I(\boldsymbol{X}))\,\log\left(G\cdot p^{4\log 2}\right)}{c_7\,d_{\min}}} \tag{A.26}$$

we have

$$\|\boldsymbol{\beta}^*_{\mathcal{I}_{g'}}\|_2 \geq 10\,\sigma(1+\epsilon)\left(\sqrt{d^*_{\mathcal{G}}}+\sqrt{d_{g'}}\right)\cdot\max\left\{1,\sqrt{\frac{s}{d_{\max}\cdot\log p}}\right\}$$

for every $g' \in \mathcal{G}^*$, then $\Pr\left(E_4^c|E_1\cap E_2\right) \leq 2\,p^{-4\log 2} + 2\exp\left(-c_7\epsilon^2 d^*_{\mathcal{G}}/2\right)$. Finally, by Lemma A.2.4, we have that $\Pr\left(E_5^c|E_1\cap E_2\right) \leq 2p^{-4\log 2}$ whenever $\lambda_g = 4\sigma(1+\epsilon)\sqrt{d_g}$ for all $g \notin \mathcal{G}^*$. Therefore, under stated conditions of the theorem we have

$$\Pr(E^c) \leq 10p^{-4\log 2} + 2\exp\left(-c_7\epsilon^2 d^*_{\mathcal{G}}/2\right) \leq 12\,p^{-2\log 2},$$

where the last inequality follows from the lower bound on ϵ, namely that $\exp(-c_7\epsilon^2 d^*_{\mathcal{G}}/2) \leq p^{-2\log 2}$. Finally, since $c_7 > 3$, therefore the choice of ϵ in theorem statement is always above the threshold in (A.26).

Well-Conditioning of the Selected Sub-Dictionary

Here, we establish a general guarantee for the well-conditioning of the sub-dictionaries selected from an arbitrary dictionary $\boldsymbol{X}$ in terms of its intra- and inter-block coherence parameters, $\mu_I(\boldsymbol{X})$ and $\mu_B(\boldsymbol{X})$, which are defined as in Definition 2.2.1.

We begin by restating (and slightly expanding) the notation needed to develop our analysis. A predefined column-wise partition of the dictionary $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ into G column blocks is given by $\boldsymbol{X} = [\boldsymbol{X}_{\mathcal{I}_1}\boldsymbol{X}_{\mathcal{I}_2}\cdots\boldsymbol{X}_{\mathcal{I}_G}]$, where $\mathcal{I}_g$ denotes the indices for columns belonging to block g and $\boldsymbol{X}_{\mathcal{I}_g} \in \mathbb{R}^{n\times d_g}$ denotes the corresponding dictionary block, and $\sum_{g\in[G]} d_g = p$. Letting $\mathcal{G} \subset [G]$ be a set of group indices and $\mathcal{S}_{\mathcal{G}} = \cup_{g\in\mathcal{G}}\mathcal{I}_g$ be the corresponding column indices, then $\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}}$ will denote the sub-dictionary whose columns are in $\mathcal{S}_{\mathcal{G}}$. With this notation, here is the first Lemma which holds for any arbitrary dictionary $\boldsymbol{X}$.

Lemma A.2.5. *Suppose $\boldsymbol{X}$ denotes an arbitrary dictionary which is column-wise partitioned as above. For any sub-dictionary $\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}}$ of $\boldsymbol{X}$ which comprises $|\mathcal{G}|$ blocks we have*

$$\|\boldsymbol{X}^T_{\mathcal{S}_{\mathcal{G}}}\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}} - \boldsymbol{I}_{d_{\mathcal{G}}\times d_{\mathcal{G}}}\|_{2\to 2} \leq \mu_I(\boldsymbol{X}) + (|\mathcal{G}|-1)\mu_B(\boldsymbol{X}), \tag{A.27}$$

where $d_{\mathcal{G}} = \sum_{g\in\mathcal{G}} d_g$ is the total number of columns of $\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}}$.

Proof. Let λ denote an arbitrary eigenvalue of the hollow Gram matrix associated with $\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}}$, namely assume that for $\boldsymbol{H} := \boldsymbol{X}^T_{\mathcal{S}_{\mathcal{G}}}\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}} - \boldsymbol{I}_{d_{\mathcal{G}}\times d_{\mathcal{G}}}$ there exists some (non-zero)

eigenvector $\boldsymbol{x} \in \mathbb{R}^{d_{\mathcal{G}}}$ such that we have $\boldsymbol{H}\boldsymbol{x} = \lambda \boldsymbol{x}$. Furthermore, let $\boldsymbol{x} = [\boldsymbol{x}_1^T \boldsymbol{x}_2^T \cdots \boldsymbol{x}_{|\mathcal{G}|}^T]^T$ denote the partitioning of $\boldsymbol{x}$ with respect to the blocks incorporated in $\boldsymbol{H}$, with $\boldsymbol{x}_i \in \mathbb{R}^{d_{g_i}}$ for $i = 1, 2, \cdots, |\mathcal{G}|$. Since $\boldsymbol{x}$ is a non-zero vector, there must exist a block $\boldsymbol{x}_i$ with maximum Euclidean norm, i.e. we can find $1 \leq i \leq |\mathcal{G}|$ such that $\|\boldsymbol{x}_i\|_2 = \max_{1 \leq j \leq |\mathcal{G}|} \|\boldsymbol{x}_j\|_2$. If $\boldsymbol{H}_{i:}$ denotes the $i-$th row sub-dictionary of $\boldsymbol{H}$, then we have $\lambda \, \boldsymbol{x}_i = \boldsymbol{H}_{i:} \, \boldsymbol{x} = \sum_{1=j}^{|\mathcal{G}|} \boldsymbol{H}_{ij} \, \boldsymbol{x}_j$ that by taking Euclidean norm of its both sides and using the triangle inequality leads to

$$\lambda \|\boldsymbol{x}_i\|_2 \leq \sum_{j=1}^{|\mathcal{G}|} \|\boldsymbol{H}_{ij} \, \boldsymbol{x}_j\|_2 \leq \sum_{j=1}^{|\mathcal{G}|} \|\boldsymbol{H}_{ij}\|_{2\to 2} \|\boldsymbol{x}_j\|_2. \tag{A.28}$$

Therefore,

$$\lambda \leq \sum_{j=1}^{|\mathcal{G}|} \|\boldsymbol{H}_{ij}\|_{2\to 2} \frac{\|\boldsymbol{x}_j\|_2}{\|\boldsymbol{x}_i\|_2} \leq \sum_{j=1}^{|\mathcal{G}|} \|\boldsymbol{H}_{ij}\|_{2\to 2}, \tag{A.29}$$

where the last inequality uses $\|\boldsymbol{x}_i\|_2 \geq \|\boldsymbol{x}_j\|_2$ for any j. Since

$$\sum_{j=1}^{|\mathcal{G}|} \|\boldsymbol{H}_{ij}\|_{2\to 2} = \|\boldsymbol{H}_{ii}\|_{2\to 2} + \sum_{j \neq i} \|\boldsymbol{H}_{ij}\|_{2\to 2},$$

where $\|\boldsymbol{H}_{ii}\|_{2\to 2} = \|\boldsymbol{X}_{\mathcal{I}_{g_i}}^H \boldsymbol{X}_{\mathcal{I}_{g_i}} - \boldsymbol{I}_{d_{g_i} \times d_{g_i}}\|_{2\to 2} \leq \mu_I(\boldsymbol{X})$ by the definition of the intra-block coherence, and $\|\boldsymbol{H}_{ij}\|_{2\to 2} = \|\boldsymbol{X}_{\mathcal{I}_{g_i}}^H \boldsymbol{X}_{\mathcal{I}_{g_j}}\|_{2\to 2} \leq \mu_B(\boldsymbol{X})$ for $i \neq j$ by the definition of the inter-block coherence $\mu_B(\boldsymbol{X})$, it follows that $\lambda \leq \mu_I(\boldsymbol{X}) + (|\mathcal{G}| - 1)\mu_B(\boldsymbol{X})$.

□

Notice that since

$$\|\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}}^H \boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}} - \boldsymbol{I}_{d_{\mathcal{G}} \times d_{\mathcal{G}}}\|_{2\to 2} = \max\left\{\sigma_{\max}^2(\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}}) - 1, 1 - \sigma_{\min}^2(\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}})\right\},$$

the implication of having $\|\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}}^H \boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}} - \boldsymbol{I}_{d_{\mathcal{G}} \times d_{\mathcal{G}}}\|_{2\to 2} \leq \delta$ for some $\delta \in [0, 1)$ is that the singular values of $\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}}$ lie within the interval $[\sqrt{1-\delta}, \sqrt{1+\delta}]$. Now according to the above Lemma if it holds that

$$|\mathcal{G}| \leq \frac{\delta - \mu_I(\boldsymbol{X})}{\mu_B(\boldsymbol{X})} + 1 \tag{A.30}$$

with the upper bound being strictly positive, then the well-conditioning of the sub-dictionary $\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}}$ will be implied, i.e. $\sigma(\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}}) \in [\sqrt{1-\delta}, \sqrt{1+\delta}]$ for every $\sigma(\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}})$.

Using the statistical model assumption $\mathbf{M_1}$ a stronger probabilistic guarantee can be provided on the well-conditioning of $\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*}$, which is stated below.

Lemma A.2.6 (Theorem 1 of [41]). *Suppose the $n\times p$ dictionary $\boldsymbol{X} = [\boldsymbol{X}_{\mathcal{I}_1}\, \boldsymbol{X}_{\mathcal{I}_2} \cdots \boldsymbol{X}_{\mathcal{I}_G}]$, column-wise partitioned into G blocks, satisfies $\mu_I(\boldsymbol{X}) \le c_0$ and $\mu_B(\boldsymbol{X}) \le c_1/\log p$, with positive constants c_0 and c_1. Assume further that $\mathcal{G}^*$ is a subset of size $s := |\mathcal{G}^*|$ of the set $[G] = \{1, 2, \cdots, G\}$, which is drawn uniformly at random. Then, as long as*

$$s \le \min\left\{\frac{c_2}{\mu_B^2(\boldsymbol{X})\log p}, \frac{c_3\, G}{\|\boldsymbol{X}\|_{2\to 2}^2 \log p}\right\} \tag{A.31}$$

for some positive constants c_2 and c_3 that only depend on c_0 and c_1. Then the singular values of the sub-dictionary $\boldsymbol{X}_{\mathcal{S}_\mathcal{G}^*}$ lie within the $[\sqrt{1/2}, \sqrt{3/2}]$ interval, or equivalently $\|\boldsymbol{X}_{\mathcal{S}_\mathcal{G}^*}^T \boldsymbol{X}_{\mathcal{S}_\mathcal{G}^*} - \boldsymbol{I}_{d_\mathcal{G}^* \times d_\mathcal{G}^*}\|_{2\to 2} \le \frac{1}{2}$, with probability at least $1 - 2p^{-4\log 2}$ with respect to the random choice of the group index sets $\mathcal{G}^*$.

The above lemma is essentially identical to Theorem 1 of [41], with the difference that in (A.31) we have replaced $\mu_B(\boldsymbol{X})$ by $\overline{\mu}_B(\boldsymbol{X})$, where the latter is called quadratic-mean block coherence in [41]. This change yields a slightly more restrictive condition, since $\mu_B(\boldsymbol{X}) \ge \overline{\mu}_B(\boldsymbol{X})$; but it does not cause a significant difference in the context of our demixing problem. As a consequence of this lemma, it directly follows that under the described conditions and with high probability $\|(\boldsymbol{X}_{\mathcal{S}_\mathcal{G}^*}^T \boldsymbol{X}_{\mathcal{S}_\mathcal{G}^*})^{-1}\|_{2\to 2} \le 2$. We would like to note that in the above Lemma c_2 and c_3 are selected such that $(48c_1 + 6\sqrt{2(c_2+c_3)} + 2c_3 + 3c_0) \le \frac{1}{4}$ holds true [3] . The following lemma, which provides a probabilistic upper bound on the quantity $\|\boldsymbol{X}_{\mathcal{S}_\mathcal{G}^*}^T \boldsymbol{X}_{(\mathcal{S}_\mathcal{G}^*)^c}\|_{B,1} = \max_{g\notin\mathcal{G}^*} \|\boldsymbol{X}_{\mathcal{S}_\mathcal{G}^*}^T \boldsymbol{X}_{\mathcal{I}_g}\|_{2\to 2}$, also turns out to be useful in the proof of the strict dual feasibility condition.

Lemma A.2.7. *Suppose the $n\times p$ dictionary $\boldsymbol{X}$ is column-wise partitioned into G blocks as $\boldsymbol{X} = [\boldsymbol{X}_{\mathcal{I}_1}\, \boldsymbol{X}_{\mathcal{I}_2} \cdots \boldsymbol{X}_{\mathcal{I}_G}]$. Assume further that $\mathcal{G}^*$ is a subset of size $s = |\mathcal{G}^*|$ of the set $[G] = \{1, 2, \cdots, G\}$, which is drawn uniformly at random. Then for $\gamma > 0$ it follows*

$$\Pr\left(\left\|\boldsymbol{X}_{\mathcal{S}_\mathcal{G}^*}^T \boldsymbol{X}_{(\mathcal{S}_\mathcal{G}^*)^c}\right\|_{B,1} > \gamma\right) \le 2\left\{\frac{\mu_B(\boldsymbol{X})}{\gamma}\cdot\left(4\sqrt{2\log p} + \sqrt{s}\right)\right\}^{4\log p}. \tag{A.32}$$

In particular, for the choice of

$$\gamma := \frac{\lambda_{\min}}{\lambda_{\max}}\cdot\frac{c_4}{\sqrt{d_{\max}\cdot\log p}},$$

[3] This can be shown by using Eq. (5) in [41] and the discussion following that for bounding the expression appearing inside parentheses there.

where c_4 is an arbitrary positive constant, it holds that $\|\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{(\mathcal{S}^*_{\mathcal{G}})^c}\|_{B,1} \leq \gamma$, with probability at least $1-2\,p^{-4\log 2}$, as long as

$$\mu_B(\boldsymbol{X}) \leq \frac{\lambda_{\min}}{\lambda_{\max}} \cdot \frac{1}{\sqrt{d_{\max}\cdot \log p}} \cdot \min\left\{\frac{c_5}{\sqrt{\log p}}, \frac{c_6}{\sqrt{s}}\right\}, \tag{A.33}$$

where c_5 and c_6 are small enough universal constants that satisfy $4\sqrt{2}\,c_5 + c_6 \leq \frac{c_4}{2}$.

Proof. The proof first utilizes Lemma A.5 in [41] to show that

$$\begin{aligned}
\Pr\left(\left\|\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{(\mathcal{S}^*_{\mathcal{G}})^c}\right\|_{B,1} > \gamma\right) &\leq 2\gamma^{-q}\,\mathbb{E}\left\|\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{(\mathcal{S}^*_{\mathcal{G}})^c}\right\|^q_{B,1} \\
&\leq 2\gamma^{-q}\,\mathbb{E}\left\|\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}\right\|^q_{B,1} \\
&\leq 2\gamma^{-q}\,\left(2^{1.5}\sqrt{q}\,\mu_B(\boldsymbol{X}) + \sqrt{s}\,\mu_B(\boldsymbol{X})\right)^q
\end{aligned}$$

where $q := 4\log p$, the first inequality is due to the Markov inequality and a Poissonization argument (similar argument is used in the proof of Theorems 1 and 2 in [41]), the second inequality is due to the fact that $\boldsymbol{X}_{(\mathcal{S}^*_{\mathcal{G}})^c}$ is a sub-dictionary of $\boldsymbol{X}$, the third inequality is by Lemma A.5 in [41] along with that $\mu_B(\boldsymbol{X}) \geq \overline{\mu}_B(\boldsymbol{X})$. Rearranging the terms would then complete the proof of the first part. Setting $\gamma = c_4/(\frac{\lambda_{\max}}{\lambda_{\min}})\sqrt{d_{\max}\cdot\log p}$ will convert the upper bound of (A.32) into

$$\begin{aligned}
&\Pr\left(\left\|\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{(\mathcal{S}^*_{\mathcal{G}})^c}\right\|_{B,1} > \gamma\right) \\
&\leq 2\left(\frac{\mu_B(\boldsymbol{X})}{c_4}\frac{\lambda_{\max}}{\lambda_{\min}}\sqrt{d_{\max}\,\log p}\left(4\sqrt{2\log p} + \sqrt{s}\right)\right)^{4\log p} \\
&\leq 2\left(4\sqrt{2}\,\frac{c_5}{c_4} + \frac{c_6}{c_4}\right)^{4\log p} \leq 2\,p^{-4\log 2}
\end{aligned}$$

where the second inequality is by the condition (A.33) on $\mu_B(\boldsymbol{X})$ and the third one holds since $4\sqrt{2}\,\frac{c_5}{c_4} + \frac{c_6}{c_4} < 0.5$. □

Proof of Lemma A.1.2

Notice that by the sub-multiplicativity property of the spectral norm we have

$$\begin{aligned}
&\left\|\boldsymbol{X}^T_{\mathcal{I}_g}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}}(\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}})^{-1}\boldsymbol{\Lambda}_{\mathcal{S}^*_{\mathcal{G}}}\check{\boldsymbol{z}}_{\mathcal{S}^*_{\mathcal{G}}}\right\|_2 \\
&\leq \left\|\boldsymbol{X}^T_{\mathcal{I}_g}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}}\right\|_{2\to 2}\left\|(\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}})^{-1}\right\|_{2\to 2}\left\|\boldsymbol{\Lambda}_{\mathcal{S}^*_{\mathcal{G}}}\check{\boldsymbol{z}}_{\mathcal{S}^*_{\mathcal{G}}}\right\|_2.
\end{aligned} \tag{A.34}$$

In order to control the first and second spectral norms appearing on the right hand-side expression we invoke Lemma A.5.1 and the assumption on the well-conditioning of the selected sub-dictionary, respectively, which together imply that

$$\left\| \frac{1}{\lambda_g} \boldsymbol{X}_{\mathcal{I}_g}^T \boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*} (\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*}^T \boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*})^{-1} \boldsymbol{\Lambda}_{\mathcal{S}_{\mathcal{G}}^*} \check{\boldsymbol{z}}_{\mathcal{S}_{\mathcal{G}}^*} \right\|_2 \leq \frac{2\sqrt{s}\, \mu_B(\boldsymbol{X})}{\lambda_{\min}} \left\| \boldsymbol{\Lambda}_{\mathcal{S}_{\mathcal{G}}^*} \check{\boldsymbol{z}}_{\mathcal{S}_{\mathcal{G}}^*} \right\|_2$$

Since every diagonal element of $\boldsymbol{\Lambda}_{\mathcal{S}_{\mathcal{G}}^*}$ is no larger than $\lambda_{\max}$ and $\check{\boldsymbol{z}}_{\mathcal{S}_{\mathcal{G}}^*}$ is composed of s sub-vectors whose norms are less than or equal to one, it will follow that

$$\left\| \boldsymbol{\Lambda}_{\mathcal{S}_{\mathcal{G}}^*} \check{\boldsymbol{z}}_{\mathcal{S}_{\mathcal{G}}^*} \right\|_2 \leq \lambda_{\max} \sqrt{s}.$$

Substituting the upper bounds from the last two inequalities in the original inequality (A.34) and rearranging the terms will imply the Lemma.

Proof of Lemma A.1.3

Using the relationship in (A.10), and defining $\boldsymbol{S}_g \in \mathbb{R}^{d_g \times d_{\mathcal{G}}^*}$ as the selector matrix which selects indices corresponding to the block $g \in \mathcal{G}^*$, we have that for each $g \in \mathcal{G}^*$,

$$\check{\boldsymbol{\beta}}_{\mathcal{I}_g} = \boldsymbol{\beta}_{\mathcal{I}_g}^* + \boldsymbol{S}_g (\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*}^T \boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*})^{-1} (\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*}^T \boldsymbol{w} - \boldsymbol{\Lambda}_{\mathcal{S}_{\mathcal{G}}^*} \check{\boldsymbol{z}}_{\mathcal{S}_{\mathcal{G}}^*}). \tag{A.35}$$

We use the implication of (A.6), writing $(\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*}^T \boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*})^{-1} = \boldsymbol{I}_{d_{\mathcal{G}}^* \times d_{\mathcal{G}}^*} + \boldsymbol{\Delta}$, where $\|\boldsymbol{\Delta}\|_{2\to 2} \leq 1$, and note that

$$\begin{aligned}
\|\boldsymbol{h}_g\|_2 &= \|\boldsymbol{S}_g (\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*}^T \boldsymbol{w} - \boldsymbol{\Lambda}_{\mathcal{S}_{\mathcal{G}}^*} \check{\boldsymbol{z}}_{\mathcal{S}_{\mathcal{G}}^*}) + \boldsymbol{S}_g \boldsymbol{\Delta} (\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*}^T \boldsymbol{w} - \boldsymbol{\Lambda}_{\mathcal{S}_{\mathcal{G}}^*} \check{\boldsymbol{z}}_{\mathcal{S}_{\mathcal{G}}^*})\|_2 \\
&\leq \|\boldsymbol{X}_{\mathcal{I}_g}^T \boldsymbol{w}\|_2 + \|\lambda_g \check{\boldsymbol{z}}_{\mathcal{I}_g}\|_2 \\
&\quad + \|\boldsymbol{S}_g\|_{2\to 2} \|\boldsymbol{\Delta}\|_{2\to 2} \big(\|\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*}^T \boldsymbol{w}\|_2 + \|\boldsymbol{\Lambda}_{\mathcal{S}_{\mathcal{G}}^*} \check{\boldsymbol{z}}_{\mathcal{S}_{\mathcal{G}}^*}\|_2 \big).
\end{aligned}$$

The second result follows from the facts that $\|\boldsymbol{\Delta}\|_{2\to 2} \leq 2$, and $\|\boldsymbol{S}_g\|_{2\to 2} \leq 1$, and that

$$\|\boldsymbol{\Lambda}_{\mathcal{S}_{\mathcal{G}}^*} \check{\boldsymbol{z}}_{\mathcal{S}_{\mathcal{G}}^*}\|_2 = \Big(\sum_{g \in \mathcal{G}^*} \lambda_g^2 \|\check{\boldsymbol{z}}_{\mathcal{I}_g}\|_2^2 \Big)^{1/2} \leq \|\boldsymbol{\lambda}_{\mathcal{G}^*}\|_2, \tag{A.36}$$

where we have used the definition of $\boldsymbol{\lambda}_{\mathcal{G}^*}$, and the subgradient condition on each group of $\check{\boldsymbol{z}}$. The second result follows from a similar argument as that given for Lemma 3

in [32], which is brought here for completeness. First notice that $\|\boldsymbol{h}_{g'}\|_2 \leq \frac{1}{2}\|\boldsymbol{\beta}^*_{\mathcal{I}_{g'}}\|_2$ implies $\check{\boldsymbol{\beta}}_{\mathcal{I}_{g'}} \neq \mathbf{0}$ and so $\check{\boldsymbol{z}}_{\mathcal{I}_{g'}} = \frac{\check{\boldsymbol{\beta}}_{\mathcal{I}_{g'}}}{\|\check{\boldsymbol{\beta}}_{\mathcal{I}_{g'}}\|_2}$. Given this, we have that

$$\begin{aligned}
\boldsymbol{u}_{g'} &= \check{\boldsymbol{z}}_{\mathcal{I}_{g'}} - \frac{\boldsymbol{\beta}^*_{\mathcal{I}_{g'}}}{\|\boldsymbol{\beta}^*_{\mathcal{I}_{g'}}\|_2} = \frac{\check{\boldsymbol{\beta}}_{\mathcal{I}_{g'}}}{\|\check{\boldsymbol{\beta}}_{\mathcal{I}_{g'}}\|_2} - \frac{\boldsymbol{\beta}^*_{\mathcal{I}_{g'}}}{\|\boldsymbol{\beta}^*_{\mathcal{I}_{g'}}\|_2} \\
&= \frac{\boldsymbol{\beta}^*_{\mathcal{I}_{g'}} + \boldsymbol{h}_{g'}}{\|\boldsymbol{\beta}^*_{\mathcal{I}_{g'}} + \boldsymbol{h}_{g'}\|_2} - \frac{\boldsymbol{\beta}^*_{\mathcal{I}_{g'}}}{\|\boldsymbol{\beta}^*_{\mathcal{I}_{g'}}\|_2} \\
&= \boldsymbol{\beta}^*_{\mathcal{I}_{g'}} \left(\frac{1}{\|\boldsymbol{\beta}^*_{\mathcal{I}_{g'}} + \boldsymbol{h}_{g'}\|_2} - \frac{1}{\|\boldsymbol{\beta}^*_{\mathcal{I}_{g'}}\|_2} \right) + \frac{\boldsymbol{h}_{g'}}{\|\boldsymbol{\beta}^*_{\mathcal{I}_{g'}} + \boldsymbol{h}_{g'}\|_2}.
\end{aligned}$$

Now, since the function $f(\boldsymbol{\beta}, \boldsymbol{h}) := 1/\|\boldsymbol{\beta} + \boldsymbol{h}\|_2$, for $\boldsymbol{\beta} \neq \mathbf{0}$, is differentiable with respect to the vector $\boldsymbol{h}$, with gradient $\nabla_{\boldsymbol{h}} f(\boldsymbol{\beta}, \boldsymbol{h}) = -\frac{\boldsymbol{\beta}+\boldsymbol{h}}{2\|\boldsymbol{\beta}+\boldsymbol{h}\|_2^3}$, therefore, by the mean value theorem, there must exist a scalar $\alpha \in [0, 1]$ such that

$$\frac{1}{\|\boldsymbol{\beta} + \boldsymbol{h}\|_2} - \frac{1}{\|\boldsymbol{\beta}\|_2} = f(\boldsymbol{\beta}, \boldsymbol{h}) - f(\boldsymbol{\beta}, \mathbf{0}) = \nabla_{\boldsymbol{h}} f(\boldsymbol{\beta}, \alpha\boldsymbol{h})^T \boldsymbol{h} = -\frac{(\boldsymbol{\beta} + \alpha\boldsymbol{h})^T \boldsymbol{h}}{2\|\boldsymbol{\beta} + \alpha\boldsymbol{h}\|_2^3}.$$

This, together with the last expression in the above, implies

$$\boldsymbol{u}_{g'} = \boldsymbol{\beta}^*_{\mathcal{I}_{g'}} \left(-\frac{(\boldsymbol{\beta}^*_{\mathcal{I}_{g'}} + \alpha\boldsymbol{h}_{g'})^T \boldsymbol{h}_{g'}}{2\|\boldsymbol{\beta}^*_{\mathcal{I}_{g'}} + \alpha\boldsymbol{h}_{g'}\|_2^3} \right) + \frac{\boldsymbol{h}_{g'}}{\|\boldsymbol{\beta}^*_{\mathcal{I}_{g'}} + \boldsymbol{h}_{g'}\|_2}.$$

Using Cauchy-Schwartz inequality then obtains

$$\begin{aligned}
\|\boldsymbol{u}_{g'}\|_2 &\leq \frac{\|\boldsymbol{\beta}^*_{\mathcal{I}_{g'}}\|_2 \cdot \|\boldsymbol{h}_{g'}\|_2}{2\|\boldsymbol{\beta}^*_{\mathcal{I}_{g'}} + \alpha\boldsymbol{h}_{g'}\|_2^2} + \frac{\|\boldsymbol{h}_{g'}\|_2}{\|\boldsymbol{\beta}^*_{\mathcal{I}_{g'}} + \boldsymbol{h}_{g'}\|_2} \\
&\leq \frac{2\|\boldsymbol{h}_{g'}\|_2}{\|\boldsymbol{\beta}^*_{\mathcal{I}_{g'}}\|_2} + \frac{\|\boldsymbol{h}_{g'}\|_2}{\|\boldsymbol{\beta}^*_{\mathcal{I}_{g'}} + \boldsymbol{h}_{g'}\|_2} \leq \frac{4\|\boldsymbol{h}_{g'}\|_2}{\|\boldsymbol{\beta}^*_{\mathcal{I}_{g'}}\|_2}
\end{aligned}$$

where the last two inequalities follow from that since $\|\boldsymbol{h}_{g'}\|_2 \leq \frac{1}{2}\|\boldsymbol{\beta}^*_{\mathcal{I}_{g'}}\|_2$, we have

$$\|\boldsymbol{\beta}^*_{\mathcal{I}_{g'}} + \alpha\boldsymbol{h}_{g'}\|_2 \geq \|\boldsymbol{\beta}^*_{\mathcal{I}_{g'}}\|_2 - \alpha\|\boldsymbol{h}_{g'}\|_2 \geq \|\boldsymbol{\beta}^*_{\mathcal{I}_{g'}}\|_2 - \|\boldsymbol{h}_{g'}\|_2 \geq \frac{1}{2}\|\boldsymbol{\beta}^*_{\mathcal{I}_{g'}}\|_2,$$

for any $\alpha \in [0, 1]$.

Proof of Lemma A.2.1

The proof essentially follows the last step in the proof of Theorem 2 in [41]. First notice that (A.16) is equal to

$$\Pr\left(\left\|\boldsymbol{\Lambda}^{-1}_{(\mathcal{S}^*_{\mathcal{G}})^c}\boldsymbol{X}^T_{(\mathcal{S}^*_{\mathcal{G}})^c}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}}(\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}})^{-1}\boldsymbol{\Lambda}_{\mathcal{S}^*_{\mathcal{G}}}\overline{\boldsymbol{\beta}^*_{\mathcal{S}^*_{\mathcal{G}}}}\right\|_{2,\infty} > \frac{1}{4}\right),$$

where for a block-wise partitioned arbitrary vector $\boldsymbol{a} = \left[\boldsymbol{a}^T_{\mathcal{I}_1}\, \boldsymbol{a}^T_{\mathcal{I}_2}\, \cdots \boldsymbol{a}^T_{\mathcal{I}_G}\right]^T$, $\|\boldsymbol{a}\|_{2,\infty}$ denotes the maximum Euclidean norm of its constituent blocks, i.e.

$$\|\boldsymbol{a}\|_{2,\infty} := \max_{g\in[G]} \|\boldsymbol{a}_{\mathcal{I}_g}\|_2.$$

Furthermore, since $\|\boldsymbol{a}\|_{2,\infty} \leq \sqrt{d_{\max}}\|\boldsymbol{a}\|_\infty$ with $d_{\max}$ denoting the maximum block size, it is sufficient to show that the following inequality holds, with probability at least $1-\eta$, for the scalar random variable v defined as below

$$v := \left\|\boldsymbol{\Lambda}^{-1}_{(\mathcal{S}^*_{\mathcal{G}})^c}\boldsymbol{X}^T_{(\mathcal{S}^*_{\mathcal{G}})^c}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}}(\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}})^{-1}\boldsymbol{\Lambda}_{\mathcal{S}^*_{\mathcal{G}}}\overline{\boldsymbol{\beta}^*_{\mathcal{S}^*_{\mathcal{G}}}}\right\|_\infty \leq \frac{1}{4\sqrt{d_{\max}}}.$$

Letting $v_{g,j} := \frac{1}{\lambda_g}\boldsymbol{x}^T_{g,j}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}}(\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}})^{-1}\boldsymbol{\Lambda}_{\mathcal{S}^*_{\mathcal{G}}}\overline{\boldsymbol{\beta}^*_{\mathcal{S}^*_{\mathcal{G}}}}$, where $\boldsymbol{x}_{g,j}$ denotes the j-th column in the block sub-dictionary $\boldsymbol{X}_{\mathcal{I}_g} \in \mathbb{R}^{n\times d_g}$, with $j \in [d_g]$, we may write

$$v = \max_{g\notin\mathcal{G}^*,\, j\in[d_g]} |v_{g,j}|\,.$$

Moreover, by defining the vector $\boldsymbol{w}_{g,j} := \frac{1}{\lambda_g}(\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}})^{-1}\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{x}_{g,j}$ for $g \notin \mathcal{G}^*$ and $j \in [d_g]$, we can express each $v_{g,j}$ as the following inner product $v_{g,j} = \boldsymbol{w}^T_{g,j}\,\boldsymbol{\Lambda}_{\mathcal{S}^*_{\mathcal{G}}}\,\overline{\boldsymbol{\beta}^*_{\mathcal{S}^*_{\mathcal{G}}}}$. Notice that in the current lemma we are proceeding under the condition that the selected block support $\mathcal{G}^*$ is fixed, and therefore only random vector that appears on the right-hand side of the last expression is $\overline{\boldsymbol{\beta}^*_{\mathcal{S}^*_{\mathcal{G}}}}$. Now, by utilizing the definition of $\overline{\boldsymbol{\beta}^*_{\mathcal{S}^*_{\mathcal{G}}}}$ and that $\boldsymbol{w}_{g,j}$ is the concatenation of block vectors $\boldsymbol{w}_{g,j,g'} \in \mathbb{R}^{d_{g'}}$ (with $g' \in \mathcal{G}^*$) corresponding to row-wise blocks in the partition of $\left(\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}}\right)^{-1}$ we can express $v_{g,j}$ as follows

$$v_{g,j} = \sum_{g'\in\mathcal{G}^*} \lambda_{g'}\boldsymbol{w}^T_{g,j,g'}\left(\frac{\boldsymbol{\beta}^*_{\mathcal{I}_{g'}}}{\|\boldsymbol{\beta}^*_{\mathcal{I}_{g'}}\|_2}\right).$$

Since $v_{g,j}$ is now expressed in the form of the summation of random variables, its absolute value can be bounded by utilizing probabilistic concentration tools. To do so, first we

apply Cauchy-Schwartz inequality to every term in the summation to yield

$$\left|\lambda_{g'}\boldsymbol{w}_{g,j,g'}^T\left(\frac{\boldsymbol{\beta}^*_{\mathcal{I}_{g'}}}{\|\boldsymbol{\beta}^*_{\mathcal{I}_{g'}}\|_2}\right)\right| \leq \lambda_{g'}\left\|\boldsymbol{w}_{g,j,g'}\right\|_2,$$

where we also employed that $\boldsymbol{\beta}^*_{\mathcal{I}_{g'}}/\|\boldsymbol{\beta}^*_{\mathcal{I}_{g'}}\|_2$ is a unit-norm vector. Since $\mathbb{E}\left[\frac{\boldsymbol{\beta}^*_{\mathcal{I}_{g'}}}{\|\boldsymbol{\beta}^*_{\mathcal{I}_{g'}}\|_2}\right] = \mathbf{0}$ for every $g' \in \mathcal{G}^*$, using Hoeffding's inequality will then lead to the following statement

$$\Pr\left(|v_{g,j}| \geq t\right) \leq 2\exp\left(\frac{-t^2}{2\sum_{g'\in\mathcal{G}^*}\lambda_{g'}^2\|\boldsymbol{w}_{g,j,g'}\|_2^2}\right) = 2\exp\left(\frac{-t^2}{2\|\boldsymbol{\Lambda}_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{w}_{g,j}\|_2^2}\right).$$

By choosing $\kappa \geq \max_{g\notin\mathcal{G}^*,\, j\in[d_g]}\left\|\boldsymbol{\Lambda}_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{w}_{g,j}\right\|_2$ and applying a union bound we obtain $\Pr\left(v \geq t\right) \leq 2p\exp\left(\frac{-t^2}{2\kappa^2}\right)$. To find an appropriate choice for κ that is explicitly in terms of our defining parameters, we explore upper bounds on $\boldsymbol{w}_{g,j}$ as follows

$$\|\boldsymbol{w}_{g,j}\|_2 \leq \frac{1}{\lambda_g}\left\|\left(\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}}\right)^{-1}\right\|_{2\to 2}\left\|\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\,\boldsymbol{x}_{g,j}\right\|_2,$$

where since $\boldsymbol{x}_{g,j}$ is a column of the dictionary block $\boldsymbol{X}_{\mathcal{I}_g}$, it follows that

$$\begin{aligned}\left\|\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{x}_{g,j}\right\|_2 &\leq \left\|\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{\mathcal{I}_g}\right\|_{2\to 2} \leq \max_{g\notin\mathcal{G}^*}\left\|\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{\mathcal{I}_g}\right\|_{2\to 2}\\ &\leq \left\|\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{(\mathcal{S}^*_{\mathcal{G}})^c}\right\|_{B,1}.\end{aligned}$$

Since the selected sub-dictionary is well-conditioned, i.e. $\|(\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}})^{-1}\|_{2\to 2} \leq 2$, as guaranteed by E_1, and moreover that $\|\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{(\mathcal{S}^*_{\mathcal{G}})^c}\|_{B,1} \leq \gamma$, as guaranteed by E_2, then an upper bound on $\|\boldsymbol{w}_{g,j}\|_2$ would be $\|\boldsymbol{w}_{g,j}\|_2 \leq 2\gamma/\lambda_g \leq 2\gamma/\lambda_{\min}$ and therefore an appropriate choice for κ would be $\kappa = 2\gamma(\frac{\lambda_{\max}}{\lambda_{\min}})$ (also by that $\|\boldsymbol{\Lambda}_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{w}_{g,j}\|_2 \leq \lambda_{\max}\|\boldsymbol{w}_{g,j}\|_2$). Therefore, setting $t = \frac{1}{4\sqrt{d_{\max}}}$ and $\gamma = \frac{\lambda_{\min}}{\lambda_{\max}}\cdot\frac{c_4}{\sqrt{d_{\max}\cdot\log p}}$ implies

$$\begin{aligned}\Pr\left(v \geq \frac{1}{4\sqrt{d_{\max}}}\right) &\leq 2p\cdot\exp\left(\frac{-1}{32\,\kappa^2\,d_{\max}}\right)\\ &= 2p\cdot\exp\left(\frac{-1}{128\,d_{\max}\,(\frac{\lambda_{\max}}{\lambda_{\min}})^2\gamma^2}\right)\\ &= 2p^{\left(1-\frac{1}{128\,c_4^2}\right)}\end{aligned}$$

Therefore, assuming c_4 satisfies $1-\frac{1}{128\,c_4^2} \leq -4\log 2$, will imply the last expression on the right hand-side is less than $2p^{-4\log 2}$, which completes the proof.

Proof of Lemma A.2.2

We establish that the following events

$$\left\{\|\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*}^T\boldsymbol{w}\|_2 \leq \sigma(1+\epsilon)\sqrt{d_{\mathcal{G}}^*}\right\}, \text{ and } \left\{\|\boldsymbol{X}_{\mathcal{I}_{g'}}^T\boldsymbol{w}\|_2 \leq \sigma(1+\epsilon)\sqrt{d_{g'}}\,,\ \forall g' \in \mathcal{G}^*\right\}$$

hold with the specified probability using the *Hanson-Wright Inequality*; we state a useful (for our purposes) version of this inequality here as a lemma.

Lemma A.2.8 (Hanson Wright Inequality; From Thm. 2.1 of [147])**.** *Let $\boldsymbol{A}$ be a fixed matrix, and $\boldsymbol{x}$ be a vector whose elements are iid $\mathcal{N}(0,1)$ random variables (which are thus subgaussian). Then, there exists a finite constant $c_7 > 0$ such that for any $\tau > 0$,*

$$\Pr\left(|\|\boldsymbol{A}\boldsymbol{x}\|_2 - \|\boldsymbol{A}\|_F| > \tau\right) \leq 2\exp\left(-\frac{c_7\tau^2}{\|\boldsymbol{A}\|_{2\to 2}^2}\right). \tag{A.37}$$

We would like to note that for the case of Gaussian distribution, the constant c_7 lies in the interval $c_7 \in (3,7)$. Getting back to our proof, first fix any $g' \in \mathcal{G}^*$ and note that

$$\begin{aligned}\Pr\left(\|\boldsymbol{X}_{\mathcal{I}_{g'}}^T\boldsymbol{w}\|_2 > \sigma(1+\epsilon)\sqrt{d_{g'}}\right) &\\ \leq \Pr\left(\left|\|\boldsymbol{X}_{\mathcal{I}_{g'}}^T\boldsymbol{w}\|_2 - \sigma\sqrt{d_{g'}}\right| > \epsilon\sigma\sqrt{d_{g'}}\right) &\\ \leq 2\exp\left(-\frac{c_7\,\epsilon^2 d_{g'}}{1+\mu_I(\boldsymbol{X})}\right),&\end{aligned}$$

where the second inequality follows directly from Hanson-Wright inequality (setting $\boldsymbol{x} = \boldsymbol{w}/\sigma$, and $\boldsymbol{A} = \sigma\boldsymbol{X}_{\mathcal{I}_{g'}}^T$, and noting that $\|\boldsymbol{A}\|_F = \sigma\sqrt{d_{g'}}$ and $\|\boldsymbol{A}\|_{2\to 2} \leq \sigma\sqrt{1+\mu_I(\boldsymbol{X})}$, where the first statement follows from the fact that columns of $\boldsymbol{A}$ have unit Euclidean norms). Next, note that

$$\begin{aligned}\Pr\left(\|\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*}^T\boldsymbol{w}\|_2 > \sigma(1+\epsilon)\sqrt{d_{\mathcal{G}}^*}\right) &\leq \Pr\left(\left|\|\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*}^T\boldsymbol{w}\|_2 - \sigma\sqrt{d_{\mathcal{G}}^*}\right| > \epsilon\sigma\sqrt{d_{\mathcal{G}}^*}\right)\\ &\leq 2\exp\left(-\frac{2c_7\,\epsilon^2 d_{\mathcal{G}}^*}{3}\right).\end{aligned}$$

Here, the second inequality follows again from Hanson-Wright inequality, setting $\boldsymbol{x} = \boldsymbol{w}/\sigma$, and $\boldsymbol{A} = \sigma\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*}^T$, and noting that $\|\boldsymbol{A}\|_F = \sigma\sqrt{d_{\mathcal{G}}^*}$ (since each row of $\boldsymbol{A}$ is unit-norm) and $\|\boldsymbol{A}\|_{2\to 2} \leq \sigma\sqrt{3/2}$, which follows from the event E_1 (A.6).

Thus, both of the stated claims hold, except in an event of probability no larger than $2\exp\left(-2c_7\,\epsilon^2 d^*_{\mathcal{G}}/3\right) + 2\sum_{g'\in\mathcal{G}^*}\exp\left(-c_7\,\epsilon^2 d_{g'}/(1+\mu_I(\boldsymbol{X}))\right)$, which itself is upper-bounded by

$$2\exp\left(-c_7\,\epsilon^2 d^*_{\mathcal{G}}/2\right) + 2|\mathcal{G}^*|\exp\left(-c_7\epsilon^2 d_{\min}/(1+\mu_I(\boldsymbol{X}))\right),$$

where $d_{\min} := \min_{g\in[G]} d_g$. Finally, note that whenever $\epsilon \geq \sqrt{\frac{(1+\mu_I(\boldsymbol{X}))\cdot\log(p^t\,|\mathcal{G}^*|)}{c_7\,d_{\min}}}$ for any $t \geq 1$, we will have

$$2|\mathcal{G}^*|\exp\left(-c_7\epsilon^2 d_{\min}/(1+\mu_I(\boldsymbol{X}))\right) \leq 2\,p^{-t},$$

and the result follows. (Note that the constant c_7 in the stated result is the same as the constant c_7 arising in the Hanson-Wright Inequality).

Proof of Lemma A.2.3

We begin by using the sub-multiplicativity property of the spectral norm to obtain

$$\begin{aligned}
&\frac{1}{\lambda_g}\left\|\boldsymbol{X}^T_{\mathcal{I}_g}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}}(\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}})^{-1}\boldsymbol{\Lambda}_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{u}_{\mathcal{S}^*_{\mathcal{G}}}\right\|_2 \\
&\leq \frac{1}{\lambda_g}\left\|\boldsymbol{X}^T_{\mathcal{I}_g}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}}\right\|_{2\to 2}\left\|(\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}})^{-1}\right\|_{2\to 2}\left\|\boldsymbol{\Lambda}_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{u}_{\mathcal{S}^*_{\mathcal{G}}}\right\|_2 \\
&\leq \frac{2\gamma}{\lambda_g}\left\|\boldsymbol{\Lambda}_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{u}_{\mathcal{S}^*_{\mathcal{G}}}\right\|_2 \leq 2\gamma\,\frac{\lambda_{\max}}{\lambda_{\min}}\left\|\boldsymbol{u}_{\mathcal{S}^*_{\mathcal{G}}}\right\|_2
\end{aligned}$$

where the second inequality follows since we assume E_1 and E_2 hold true (therefore $\|(\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}})^{-1}\|_{2\to 2} \leq 2$ and $\|\boldsymbol{X}^T_{\mathcal{I}_g}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}}\|_{2\to 2} \leq \|\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{(\mathcal{S}^*_{\mathcal{G}})^c}\|_{B,1} \leq \gamma$) and the third inequality follows by the fact that $\|\boldsymbol{\Lambda}_{\mathcal{S}^*_{\mathcal{G}}}\|_{2\to 2} = \lambda_{\max}$ (and therefore $\|\boldsymbol{\Lambda}_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{u}_{\mathcal{S}^*_{\mathcal{G}}}\|_2 \leq \lambda_{\max}\|\boldsymbol{u}_{\mathcal{S}^*_{\mathcal{G}}}\|_2$). In addition, note that by assuming $\|\boldsymbol{\beta}^*_{\mathcal{I}_{g'}}\|_2 \geq t_2\|\boldsymbol{h}_{g'}\|_2 \geq 2\,\|\boldsymbol{h}_{g'}\|_2$ for all $g' \in \mathcal{G}^*$, Lemma A.1.3 implies

$$\|\boldsymbol{u}_{g'}\|_2 \leq 4\frac{\left\|\boldsymbol{h}_{g'}\right\|_2}{\|\boldsymbol{\beta}^*_{\mathcal{I}_{g'}}\|_2} \leq \frac{4}{t_2} \tag{A.38}$$

for all $g' \in \mathcal{G}^*$ and therefore $\|\boldsymbol{u}_{\mathcal{S}^*_{\mathcal{G}}}\|_2 \leq 4\sqrt{|\mathcal{G}^*|}/t_2$. Combining all of these we obtain

$$\frac{1}{\lambda_g}\left\|\boldsymbol{X}^T_{\mathcal{I}_g}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}}(\boldsymbol{X}^T_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}})^{-1}\boldsymbol{\Lambda}_{\mathcal{S}^*_{\mathcal{G}}}\boldsymbol{u}_{\mathcal{S}^*_{\mathcal{G}}}\right\|_2 \leq \frac{\lambda_{\max}}{\lambda_{\min}}\cdot\frac{8\gamma\sqrt{|\mathcal{G}^*|}}{t_2}$$

Therefore, assuming the event E_2 holds for the choice of $\gamma = c_4/(\frac{\lambda_{\max}}{\lambda_{\min}})\sqrt{d_{\max} \cdot \log p}$, where $c_4 \leq 1/8\sqrt{2(1+4\log 2)}$ is a finite positive constant as appeared in the proof of Lemma A.2.1, will ensure that

$$\frac{1}{\lambda_g}\left\|\boldsymbol{X}_{\mathcal{I}_g}^T\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*}(\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*}^T\boldsymbol{X}_{\mathcal{S}_{\mathcal{G}}^*})^{-1}\boldsymbol{\Lambda}_{\mathcal{S}_{\mathcal{G}}^*}\boldsymbol{u}_{\mathcal{S}_{\mathcal{G}}^*}\right\|_2 \leq \frac{8c_4}{t_2}\cdot\sqrt{\frac{|\mathcal{G}^*|}{d_{\max}\cdot\log p}}.$$

Then choosing $t_2 \geq c_8\sqrt{|\mathcal{G}^*|/d_{\max}\log p}$ as specified by the statement of the lemma (with $c_8 := 32\,c_4$) would complete the proof.

Proof of Lemma A.2.4

Fix any $g \notin \mathcal{G}^*$. Note that for any $\tau > 0$,

$$\begin{aligned}
&\Pr\big(\|\ \boldsymbol{X}_{\mathcal{I}_g}^T\Pi_{(\mathcal{S}_{\mathcal{G}}^*)^\perp}\boldsymbol{w}\|_2 > \sigma\sqrt{d_g} + \tau\Big) \\
&\quad\leq \Pr\left(\|\boldsymbol{X}_{\mathcal{I}_g}^T\Pi_{(\mathcal{S}_{\mathcal{G}}^*)^\perp}\boldsymbol{w}\|_2 > \sigma\|\boldsymbol{X}_{\mathcal{I}_g}^T\Pi_{(\mathcal{S}_{\mathcal{G}}^*)^\perp}\|_F + \tau\right) \\
&\quad\leq \Pr\left(\left|\|\boldsymbol{X}_{\mathcal{I}_g}^T\Pi_{(\mathcal{S}_{\mathcal{G}}^*)^\perp}\boldsymbol{w}\|_2 - \sigma\|\boldsymbol{X}_{\mathcal{I}_g}^T\Pi_{(\mathcal{S}_{\mathcal{G}}^*)^\perp}\|_F\right| > \tau\right),
\end{aligned}$$

where the first inequality follows from the fact that $\|\boldsymbol{X}_{\mathcal{I}_g}^T\Pi_{(\mathcal{S}_{\mathcal{G}}^*)^\perp}\|_F \leq \sqrt{d_g}$ (which is easy to verify by considering $\|\Pi_{(\mathcal{S}_{\mathcal{G}}^*)^\perp}^T\boldsymbol{X}_{\mathcal{I}_g}\|_F^2$, arranging the sum that arises in the definition of the squared Frobenius norm into a sum of sums over columns of $\boldsymbol{X}_{\mathcal{I}_g}$, and applying standard matrix inequalities along with the fact that $\|\Pi_{(\mathcal{S}_{\mathcal{G}}^*)^\perp}\|_{2\to 2} = 1$).

Now, the final upper bound above is of the form controllable by the Hanson-Wright Inequality (c.f., Lemma A.2.8). Specifically, setting $\boldsymbol{x} = \boldsymbol{w}/\sigma$, and $\boldsymbol{A} = \sigma\boldsymbol{X}_{\mathcal{I}_g}^T\Pi_{(\mathcal{S}_{\mathcal{G}}^*)^\perp}$, and using the fact that $\|\boldsymbol{X}_{\mathcal{I}_g}^T\Pi_{(\mathcal{S}_{\mathcal{G}}^*)^\perp}\|_{2\to 2} \leq \sigma\sqrt{1+\mu_I(\boldsymbol{X})}$ (which is easy to verify using the sub-multiplicativity of the spectral norm), we obtain overall that for the universal finite constant $c_7 > 0$, and the specific choice $\tau = \epsilon\sigma\sqrt{d_g}$,

$$\Pr\left(\|\boldsymbol{X}_{\mathcal{I}_g}^T\Pi_{(\mathcal{S}_{\mathcal{G}}^*)^\perp}\boldsymbol{w}\|_2 > \sigma(1+\epsilon)\sqrt{d_g}\right) \leq 2\exp\left(-c_7\epsilon^2 d_g\right).$$

Thus, it follows that

$$\begin{aligned}
\Pr\left(\bigcup_{g\notin\mathcal{G}^*}\left\{\|\boldsymbol{X}_{\mathcal{I}_g}^T\Pi_{(\mathcal{S}_{\mathcal{G}}^*)^\perp}\boldsymbol{w}\|_2 > \sigma(1+\epsilon)\sqrt{d_g}\right\}\right) &\leq 2\sum_{g\notin\mathcal{G}^*}\exp\left(-c_7\epsilon^2 d_g\right) \\
&\leq 2(G-|\mathcal{G}^*|)\exp\left(-c_7\epsilon^2 d_{\min}\right).
\end{aligned}$$

Next, note that whenever $\epsilon \geq \sqrt{\log\left(p^t\,(G-|\mathcal{G}^*|)\right)/c_7\,d_{\min}}$ the last term is no larger than $2\,p^{-t}$. Finally, note that the stated result holds if $\lambda_g \geq 4\sigma(1+\epsilon)\sqrt{d_g}$ for all $g \notin \mathcal{G}^*$.

A.3 Proof of Theorem 2.2.1

Since the number of non-zero groups of $\boldsymbol{\beta}^*$ satisfies the condition $|\mathcal{G}^*| \leq \frac{0.5-\mu_I(\boldsymbol{X})}{\mu_B(\boldsymbol{X})} + 1$, Lemma A.2.5 implies that the sub-dictionary $\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}}$, which incorporates the blocks of $\boldsymbol{X}$ corresponding to $\boldsymbol{\beta}^*_{\mathcal{S}^*_{\mathcal{G}}}$, will be well-conditioned and therefore the event E_1 holds true. Then, we can control the norm of the dual variable blocks $\check{\boldsymbol{z}}_{\mathcal{I}_g}$ for $g \notin \mathcal{G}^*$ by leveraging the inequality (A.11) and the bounds developed in Lemmata A.1.2 and A.2.4. To this end, first notice that if $|\mathcal{G}^*| \leq \frac{\lambda_{\min}}{\lambda_{\max}} \cdot \frac{1}{4\mu_B(\boldsymbol{X})}$ and $\lambda_g \geq 4\sigma(1+\epsilon)\sqrt{d_g}$ for all $g \in [G]$ and some $\epsilon \geq \sqrt{(1+\mu_I(\boldsymbol{X}))\,\log\left(p^t\,(G-|\mathcal{G}^*|)\right)/c_7\,d_{\min}}$, with $t \geq 1$, then $\|\check{\boldsymbol{z}}_{\mathcal{I}_g}\|_2 < 1$ holds for every $g \notin \mathcal{G}^*$, with probability at least $1-2p^{-t}$. Second, by Corollary A.2.1 we know that for $\epsilon \geq \sqrt{(1+\mu_I(\boldsymbol{X}))\,\log\left(p^t\,|\mathcal{G}^*|\right)/c_7\,d_{\min}}$, if for every $g' \in \mathcal{G}^*$

$$\|\boldsymbol{\beta}^*_{\mathcal{I}_{g'}}\|_2 \geq \sigma(1+\epsilon)\left(\sqrt{d'_g} + \sqrt{d^*_{\mathcal{G}}}\right) + \lambda_{g'} + \|\boldsymbol{\lambda}_{\mathcal{G}^*}\|_2,$$

then $\|\check{\boldsymbol{\beta}}_{\mathcal{I}_{g'}} - \boldsymbol{\beta}^*_{\mathcal{I}_{g'}}\|_2 \leq \|\boldsymbol{\beta}^*_{\mathcal{I}_{g'}}\|_2$ for every $g' \in \mathcal{G}^*$, with probability at least

$$1 - 4\,p^{-t/2} \leq 1 - 2\,p^{-t} - 2\exp\left(-c_7\epsilon^2 d^*_{\mathcal{G}}/2\right).$$

By using the union bound argument, it can be observed that if the assumptions of Theorem 2.2.1 are met, then with probability at least $1-6p^{-t/2}$, the sub-dictionary $\boldsymbol{X}_{\mathcal{S}^*_{\mathcal{G}}}$ will be well-conditioned, $\|\check{\boldsymbol{z}}_{\mathcal{I}_g}\| < 1$ for every $g \notin \mathcal{G}^*$, and $\|\check{\boldsymbol{\beta}}_{\mathcal{I}_{g'}} - \boldsymbol{\beta}^*_{\mathcal{I}_{g'}}\|_2 \leq \|\boldsymbol{\beta}^*_{\mathcal{I}_{g'}}\|_2$, which further imply that $\check{\boldsymbol{\beta}}$ will be the unique solution to the problem (2.2) and its group-level support will be exactly that of $\boldsymbol{\beta}^*$, i.e. $\mathcal{G}(\check{\boldsymbol{\beta}}) = \mathcal{G}(\boldsymbol{\beta}^*)$. To justify the choice for ϵ in the theorem statement, notice that specializing $t = 4\log 2$ along with the fact that $c_7 > 3$ implies that $\epsilon = \sqrt{(1+\mu_I(\boldsymbol{X}))\cdot\log(p\,G)/d_{\min}}$ meets both the above requirements on ϵ.

A.4 Proof of Corollary 2.3.1

This is a direct consequence of Theorem 2.2.1 for the anomaly detection framework studied in Section 2.3. There we assumed $\boldsymbol{X} = \left[\widetilde{\boldsymbol{X}}_{(1)}|\widetilde{\boldsymbol{X}}_{(2)}\right]$, where $\widetilde{\boldsymbol{X}}_{(1)} = \boldsymbol{I}_{T\times T} \otimes \boldsymbol{X}_{(1)}$ and $\widetilde{\boldsymbol{X}}_{(2)} = \boldsymbol{I}_{T\times T} \otimes \boldsymbol{X}_{(2)}$, with $\boldsymbol{X}_{(1)}$ and $\boldsymbol{X}_{(2)}$ specialized to two-dimensional DCT and identity matrices of size $N \times N$, respectively. Since in this setup d_g is either T (for the temporal groups defined over the support of the smooth component) or DT (for

the spatiotemporal groups defined over the support of the anomaly component), we set $\lambda_1 = 4\sigma(1+\epsilon)\sqrt{T}$ and $\lambda_2 = 4\sigma(1+\epsilon)\sqrt{DT}$ as in the statement of Theorem 2.2.1. Moreover, under the assumptions on the dictionary, we have $p = 2NT$, $\|\boldsymbol{X}\|_{2\to 2}^2 = 2$, the intra-block coherence parameter $\mu_I(\boldsymbol{X})$ will be zero and upper bounding $\mu_B(\boldsymbol{X})$ will amount to finding upper bounds on

$$\left\| \left(\widetilde{\boldsymbol{X}}_{(1)}\right)_{\mathcal{I}_i}^T \left(\widetilde{\boldsymbol{X}}_{(2)}\right)_{\mathcal{I}_j} \right\|_{2\to 2},$$

where $\left(\widetilde{\boldsymbol{X}}_{(1)}\right)_{\mathcal{I}_i}$ and $\left(\widetilde{\boldsymbol{X}}_{(2)}\right)_{\mathcal{I}_j}$ represent two column sub-matrices of $\widetilde{\boldsymbol{X}}_{(1)}$ and $\widetilde{\boldsymbol{X}}_{(2)}$ whose numbers of columns are given by the defined partition. More specifically, since the groups over the smooth component are temporal, we may write

$$\left(\widetilde{\boldsymbol{X}}_{(1)}\right)_{\mathcal{I}_i} = \boldsymbol{I}_{T\times T} \otimes \left(\boldsymbol{X}_{(1)}\right)_{\mathcal{I}_i} \in \mathbb{R}^{NT\times T}$$

for the $T\times T$ identity matrix $\boldsymbol{I}_{T\times T}$ and some column of $\boldsymbol{X}_{(1)}$ denoted by $\left(\boldsymbol{X}_{(1)}\right)_{\mathcal{I}_i}$. Also, since spatiotemporal groups are defined over the anomalous component, we may write $\left(\widetilde{\boldsymbol{X}}_{(2)}\right)_{\mathcal{I}_j} = \boldsymbol{I}_{T\times T} \otimes \left(\boldsymbol{X}_{(2)}\right)_{\mathcal{I}_j}$. Given these expressions for the sub-matrices of the two dictionaries, the associated inner products may be simplified as

$$\left(\widetilde{\boldsymbol{X}}_{(1)}\right)_{\mathcal{I}_i}^T \left(\widetilde{\boldsymbol{X}}_{(2)}\right)_{\mathcal{I}_j} = \boldsymbol{I}_{T\times T} \otimes \left(\left(\boldsymbol{X}_{(1)}\right)_{\mathcal{I}_i}^T \left(\boldsymbol{X}_{(2)}\right)_{\mathcal{I}_j} \right),$$

and it follows that

$$\left\| \boldsymbol{I}_{T\times T} \otimes \left((\boldsymbol{X}_{(1)})_{\mathcal{I}_i}^T (\boldsymbol{X}_{(2)})_{\mathcal{I}_j} \right) \right\|_{2\to 2} = \left\| (\boldsymbol{X}_{(1)})_{\mathcal{I}_i}^T (\boldsymbol{X}_{(2)})_{\mathcal{I}_j} \right\|_2.$$

Next, as $\boldsymbol{X}_{(1)} \in \mathbb{R}^{N\times N}$ is a two-dimensional DCT matrix, the absolute value of its largest entry is no larger than $\sqrt{4/N}$; see also [30]). Then since $\left(\boldsymbol{X}_{(2)}\right)_{\mathcal{I}_j}$ comprises D columns of the identity matrix, the Euclidean norm on the right hand-side of the above expression will not exceed $\sqrt{4D/N}$. Therefore, the block coherence parameter satisfies $\mu_B(\boldsymbol{X}) \leq \sqrt{4D/N}$.

The sufficient conditions stated in Corollary 2.3.1 are then simplifications of the conditions in Theorem 2.2.1 by specializing d_g and λ_g to their values mentioned in the above, and replacing $\mu_B(\boldsymbol{X})$ by its upper bound $\sqrt{4D/N}$. In particular, one can show

$$\min\left\{ \frac{0.5 - \mu_I(\boldsymbol{X})}{\mu_B(\boldsymbol{X})} + 1, \frac{\lambda_{\min}}{\lambda_{\max}} \cdot \frac{1}{4\mu_B(\boldsymbol{X})} \right\} = \frac{\lambda_{\min}}{\lambda_{\max}} \cdot \frac{1}{4\mu_B(\boldsymbol{X})} \geq \frac{\sqrt{N}}{8D} \tag{A.39}$$

and that the condition on the norms of the blocks of the coefficient matrices becomes

$$\min_{g_2 \in [G_2]} \left\| \left(\boldsymbol{B}^*_{(2)} \right)_{\mathcal{I}_{g_2}} \right\|_F \geq 5\sigma(1+\epsilon) \left(\sqrt{DT} + \sqrt{s_1 T + s_2 DT} \right),$$

for the anomalous component and

$$\min_{g_1 \in [G_1]} \left\| \left(\boldsymbol{B}^*_{(1)} \right)_{\mathcal{I}_{g_1}} \right\|_F \geq 5\sigma(1+\epsilon) \left(\sqrt{T} + \sqrt{s_1 T + s_2 DT} \right),$$

for the nominally smooth component.

A.5 Proof of Corollary 2.3.2

Similar to the proof of Corollary 2.3.1, in the context of the discussed anomaly detection problem we have $\|\boldsymbol{X}\|^2_{2\to 2} = 2$, $\mu_I(\boldsymbol{X}) = 0$, and $\mu_B(\boldsymbol{X}) \leq \sqrt{4D/N}$. The sufficient conditions stated in Corollary 2.3.2 are then simplifications of the conditions in Theorem 2.2.2. In particular, one can show that by imposing

$$\sqrt{N} \geq \frac{2\,\log(2NT)}{c_1} \sqrt{D^3\,T},$$

we are ensured

$$\mu_B(\boldsymbol{X}) \leq \sqrt{\frac{d_{\min}}{d^2_{\max}}} \cdot \frac{c_1}{\log(2NT)}.$$

Furthermore, the fact that $\mu_B(\boldsymbol{X}) \leq \sqrt{4D/N}$, along with that $d^2_{\max}/d_{\min} = D^2\,T$, can be used to demonstrate

$$\frac{d_{\min}}{d^2_{\max}} \cdot \frac{c_2'\,\mu_B^{-2}(\boldsymbol{X})}{\log(2NT)} \geq \frac{c_2'}{4\log(2NT)} \cdot \frac{N}{TD^3}.$$

Then the condition on the group-level sparsity in Theorem 2.2.2 will be ensured if

$$s = |\mathcal{G}^*| \leq \frac{c_2'\,N}{4\,TD^3\log(2NT)} = \min\left\{ \frac{c_2'\,N}{4\,TD^3\log(2NT)}, \frac{c_2\,G}{2\log(2NT)} \right\},$$

since $G = N(1+1/D) \geq N$, $c_0 = 0$, and

$$c_2 \leq 0.00028 \leq \left[\sqrt{9 + \frac{1}{2}\left(\frac{1}{4} - 3c_0 - 48c_1 \right)} - 3 \right]^2$$

so that $c_2' = 0.0001 = \min\{c_2, 0.0001\}$.

A Bound on the Spectral Norm of Concatenated Matrices

Lemma A.5.1. *Let* $\boldsymbol{X} = [\boldsymbol{X}_{\mathcal{I}_1}\boldsymbol{X}_{\mathcal{I}_2}\cdots\boldsymbol{X}_{\mathcal{I}_G}]$ *be a column-wise partitioned matrix with* $\boldsymbol{X}_{\mathcal{I}_g} \in \mathbb{R}^{n\times d_g}$ *for* $g \in [G]$*. Then it is always true that*

$$\|\boldsymbol{X}\|_{2\to 2}^2 \leq \sum_{g\in[G]} \|\boldsymbol{X}_{\mathcal{I}_g}\|_{2\to 2}^2 \tag{A.40}$$

Proof. Let $\boldsymbol{\beta} = [\boldsymbol{\beta}_{\mathcal{I}_1}^T\, \boldsymbol{\beta}_{\mathcal{I}_2}^T \cdots \boldsymbol{\beta}_{\mathcal{I}_G}^T]^T$ be an arbitrary unit norm vector partitioned according to the prescribed partition of $\boldsymbol{X}$. It follows by the triangle inequality that

$$\|\boldsymbol{X}\boldsymbol{\beta}\|_2^2 = \|\sum_{g\in[G]} \boldsymbol{X}_{\mathcal{I}_g}\boldsymbol{\beta}_{\mathcal{I}_g}\|_2^2 \leq \big(\sum_{g\in[G]} \|\boldsymbol{X}_{\mathcal{I}_g}\boldsymbol{\beta}_{\mathcal{I}_g}\|_2\big)^2 \leq \big(\sum_{g\in[G]} \|\boldsymbol{X}_{\mathcal{I}_g}\|_{2\to 2}\|\boldsymbol{\beta}_{\mathcal{I}_g}\|_2\big)^2$$

Then by using the Cauchy-Schwartz we will have that

$$\|\boldsymbol{X}\boldsymbol{\beta}\|_2^2 \leq \big(\sum_{g\in[G]} \|\boldsymbol{X}_{\mathcal{I}_g}\|_{2\to 2}^2\big)\big(\sum_{g\in[G]} \|\boldsymbol{\beta}_{\mathcal{I}_g}\|_2^2\big) \tag{A.41}$$

and since $\boldsymbol{\beta}$ is unit norm, the result will follow by the definition of the spectral norm. □

Appendix B

Proof of Results in Chapter 3

Here we prove Lemma 2. This lemma is the core in the analysis of accelerated ADMM algorithm. It is used in proof of Lemma 1 (for details see [70]). Our proof is similar to the proof presented for a similar lemma in [70].

Proof. To facilitate the proof, let us define $\boldsymbol{\lambda}_k^{1/2} := \widehat{\boldsymbol{\lambda}}_k + \tau(\boldsymbol{c} - \boldsymbol{A}\boldsymbol{x}_k - \boldsymbol{B}\boldsymbol{y}_k)$. The optimality condition corresponding to the update of $\boldsymbol{x}_k$ in Algorithm 3 gives

$$\begin{aligned} & \nabla f_1(\boldsymbol{x}_k) - \boldsymbol{A}^T\widehat{\boldsymbol{\lambda}}_k - \tau\boldsymbol{A}^T(\boldsymbol{c} - \boldsymbol{A}\boldsymbol{x}_k - \boldsymbol{B}\widehat{\boldsymbol{y}}_k) = \boldsymbol{0} \\ \Rightarrow \quad & \nabla f_1(\boldsymbol{x}_k) - \boldsymbol{A}^T\boldsymbol{\lambda}_k^{1/2} = \boldsymbol{0} \Rightarrow \ \boldsymbol{x}_k = \nabla f_1^*(\boldsymbol{A}^T\boldsymbol{\lambda}_k^{1/2}), \end{aligned} \tag{B.1}$$

where the last equality is due to the definition of dual functions and strong convexity of f_1. Using the same argument, it is easy to see that $\boldsymbol{y}_k = \nabla f_2^*(\boldsymbol{B}^T\boldsymbol{\lambda}_k)$.

As we mentioned earlier, when the function f_1 is strongly convex, its dual become smooth with Lipschitz gradient. Therefore, for any $\boldsymbol{\gamma}$

$$\begin{aligned} f_1^*(\boldsymbol{A}^T\boldsymbol{\gamma}) - f_1^*(\boldsymbol{A}^T\boldsymbol{\lambda}_{k+1}) &= \Big(f_1^*(\boldsymbol{A}^T\boldsymbol{\gamma}) - f_1^*(\boldsymbol{A}^T\boldsymbol{\lambda}_{k+1}^{1/2})\Big) + \Big(f_1^*(\boldsymbol{A}^T\boldsymbol{\lambda}_{k+1}^{1/2}) - f_1^*(\boldsymbol{A}^T\boldsymbol{\lambda}_{k+1})\Big) \\ &\geq \Big(f_1^*(\boldsymbol{A}^T\boldsymbol{\lambda}_{k+1}^{1/2}) + \langle \boldsymbol{A}\nabla f_1^*(\boldsymbol{A}^T\boldsymbol{\lambda}_{k+1}^{1/2}), \boldsymbol{\gamma} - \boldsymbol{\lambda}_{k+1}^{1/2}\rangle - f_1^*(\boldsymbol{A}^T\boldsymbol{\lambda}_{k+1}^{1/2})\Big) \\ &- \left(\langle \boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}_{k+1}^{1/2}, \boldsymbol{A}\nabla f_1^*(\boldsymbol{A}^T\boldsymbol{\lambda}_{k+1}^{1/2})\rangle + \frac{\rho^2(\boldsymbol{A})}{2\sigma_1}\|\boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}_{k+1}^{1/2}\|^2\right) \\ &= \langle \boldsymbol{\gamma} - \boldsymbol{\lambda}_{k+1}, \boldsymbol{A}\nabla f_1^*(\boldsymbol{A}^T\boldsymbol{\lambda}_{k+1}^{1/2})\rangle - \frac{\rho^2(\boldsymbol{A})}{2\sigma_1}\|\boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}_{k+1}^{1/2}\|^2. \end{aligned} \tag{B.2}$$

Our goal is to bound the term $\|\boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}_{k+1}^{1/2}\|$ in (B.2) using $\|\boldsymbol{\lambda}_{k+1} - \widehat{\boldsymbol{\lambda}}_{k+1}\|$.The updates

of A2DM2 then imply that $\boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}_{k+1}^{1/2} = -\tau \boldsymbol{B}(\boldsymbol{y}_{k+1} - \widehat{\boldsymbol{y}}_{k+1})$. Thus,

$$\boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}_{k+1}^{1/2} = -\tau \boldsymbol{B}(\nabla f_2^*(\boldsymbol{B}^T \boldsymbol{\lambda}_{k+1}) - \nabla f_2^*(\boldsymbol{B}^T \widehat{\boldsymbol{\lambda}}_{k+1})),$$

where the equality is due to the optimality conditions of the updates of $\boldsymbol{y}_{k+1}$ and $\widehat{\boldsymbol{y}}_{k+1}$. Now we use the Lipschitz continuity of the ∇f_2^* to get

$$\|\boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}_{k+1}^{1/2}\| \leq \tau \frac{\rho^2(\boldsymbol{B})}{\sigma_2} \|\boldsymbol{\lambda}_{k+1} - \widehat{\boldsymbol{\lambda}}_{k+1}\|. \tag{B.3}$$

Combining (B.2) and (B.3) and using the fact that $\tau \leq \frac{\sigma_1 \sigma_2^2}{\rho^2(\boldsymbol{A})\rho^4(\boldsymbol{B})}$

$$\begin{aligned} & f_1^*(\boldsymbol{A}^T \boldsymbol{\gamma}) - f_1^*(\boldsymbol{A}^T \boldsymbol{\lambda}_{k+1}) \\ & \geq \langle \boldsymbol{\gamma} - \boldsymbol{\lambda}_{k+1}, \boldsymbol{A} \nabla f_1^*(\boldsymbol{A}^T \boldsymbol{\lambda}_{k+1}^{1/2}) \rangle - \frac{1}{2\tau} \|\boldsymbol{\lambda}_{k+1} - \widehat{\boldsymbol{\lambda}}_{k+1}\|. \end{aligned} \tag{B.4}$$

Using the convexity of f_2^*, it is clear that

$$f_2^*(\boldsymbol{B}^T \boldsymbol{\gamma}) - f_2^*(\boldsymbol{B}^T \boldsymbol{\lambda}_{k+1}) \geq \langle \boldsymbol{\gamma} - \boldsymbol{\lambda}_{k+1}, \boldsymbol{B} \nabla f_2^*(\boldsymbol{B}^T \boldsymbol{\lambda}_{k+1}) \rangle. \tag{B.5}$$

Combining (B.4) and (B.5), we can easily get

$$\begin{aligned} D(\boldsymbol{\lambda}_{k+1}) - D(\boldsymbol{\gamma}) \geq \langle \boldsymbol{\gamma} - \boldsymbol{\lambda}_{k+1}, & \boldsymbol{A} \nabla f_1^*(\boldsymbol{A}^T \boldsymbol{\lambda}_k^{1/2}) \\ & + \boldsymbol{B} \nabla f_2^*(\boldsymbol{B}^T \boldsymbol{\lambda}_{k+1}) - \boldsymbol{c} \rangle - \frac{1}{2\tau} \|\boldsymbol{\lambda}_{k+1} - \widehat{\boldsymbol{\lambda}}_{k+1}\|. \end{aligned} \tag{B.6}$$

From the optimality conditions of the updates of accelerated ADMM it is clear that $\boldsymbol{x}_{k+1} = \nabla f_1^*(\boldsymbol{A}^T \boldsymbol{\lambda}_k^{1/2})$ and $\boldsymbol{y}_{k+1} = \nabla f_2^*(\boldsymbol{B}^T \boldsymbol{\lambda}_{k+1})$. Replacing these in (B.6) and noting $\tau(\widehat{\boldsymbol{\lambda}}_{k+1} - \boldsymbol{\lambda}_{k+1}) = \boldsymbol{A}\boldsymbol{x}_{k+1} + \boldsymbol{B}\boldsymbol{y}_{k+1} - \boldsymbol{c}$ yields

$$\begin{aligned} & D(\boldsymbol{\lambda}_{k+1}) - D(\boldsymbol{\gamma}) \\ & \geq \frac{1}{\tau} \langle \boldsymbol{\gamma} - \boldsymbol{\lambda}_{k+1}, \widehat{\boldsymbol{\lambda}}_{k+1} - \boldsymbol{\lambda}_{k+1} \rangle - \frac{1}{2\tau} \|\boldsymbol{\lambda}_{k+1} - \widehat{\boldsymbol{\lambda}}_{k+1}\|^2 \\ & = \frac{1}{\tau} \langle \boldsymbol{\gamma} - \widehat{\boldsymbol{\lambda}}_{k+1} + \widehat{\boldsymbol{\lambda}}_{k+1} - \boldsymbol{\lambda}_{k+1}, \widehat{\boldsymbol{\lambda}}_{k+1} - \boldsymbol{\lambda}_{k+1} \rangle - \frac{1}{2\tau} \|\boldsymbol{\lambda}_{k+1} - \widehat{\boldsymbol{\lambda}}_{k+1}\|^2 \\ & = \frac{1}{\tau} \langle \boldsymbol{\gamma} - \widehat{\boldsymbol{\lambda}}_{k+1}, \widehat{\boldsymbol{\lambda}}_{k+1} - \boldsymbol{\lambda}_{k+1} \rangle + \frac{1}{2\tau} \|\boldsymbol{\lambda}_{k+1} - \widehat{\boldsymbol{\lambda}}_{k+1}\|^2, \end{aligned}$$

which is the desired result. □

Appendix C

Proof of Results in Chapter 4

C.1 Auxiliary Lemmata

Despite the assumptions on the convexity of $\mathcal{L}_n$ and ω, the overall optimization problem cast in (4.2) is non-convex due to the factorized representation of the low-rank matrix. Therefore, the convergence of numerical procedures that are known to be well-behaved for convex problems cannot be directly applied to the current non-convex setup. In this section we develop the tools that enable us to generalize the familiar analysis of proximal descent methods for convex problems into our non-convex framework. In the current exposition of these Lemmata, we are assuming the symmetric PSD case. The extensions to the general non-PSD matrices are discussed in a section devoted to that case.

The first Lemma holds due to the fact that ω is invariant under right multiplication by any orthonormal matrix. Similar results for orthogonally-invariant functions (i.e. functions that are invariant to right and left multiplication of their operands by orthogonal matrices) are provided in [126, 148].

Lemma C.1.1. *For any $\boldsymbol{U} \in \mathbb{R}^{d\times r}$, $\alpha > 0$, and $\boldsymbol{R} \in \mathcal{O}_r$ it holds that*

$$prox_{\omega}(\boldsymbol{U}\boldsymbol{R}; \alpha) = prox_{\omega}(\boldsymbol{U}; \alpha)\boldsymbol{R}.$$

Proof. The proof follows from the definition of the proximal operator and the invariance

property of the regularization function ω. Note that

$$\begin{aligned}
\mathrm{prox}_{\omega}(\boldsymbol{U}\boldsymbol{R};\alpha) &= \operatorname*{argmin}_{\boldsymbol{V}\in\mathbb{R}^{d\times r}} \frac{1}{2}\|\boldsymbol{V}-\boldsymbol{U}\boldsymbol{R}\|_F^2 + \alpha\,\omega(\boldsymbol{V}) \\
&= \left(\operatorname*{argmin}_{\widetilde{\boldsymbol{V}}\in\mathbb{R}^{d\times r}} \frac{1}{2}\|\widetilde{\boldsymbol{V}}\boldsymbol{R}-\boldsymbol{U}\boldsymbol{R}\|_F^2 + \alpha\,\omega(\widetilde{\boldsymbol{V}}\boldsymbol{R})\right)\cdot\boldsymbol{R} \\
&= \left(\operatorname*{argmin}_{\widetilde{\boldsymbol{V}}\in\mathbb{R}^{d\times r}} \frac{1}{2}\|\widetilde{\boldsymbol{V}}-\boldsymbol{U}\|_F^2 + \alpha\,\omega(\widetilde{\boldsymbol{V}})\right)\cdot\boldsymbol{R} \\
&= \mathrm{prox}_{\omega}(\boldsymbol{U};\alpha)\cdot\boldsymbol{R},
\end{aligned}$$

where the second equality is by applying a change of variable $\boldsymbol{V} = \widetilde{\boldsymbol{V}}\boldsymbol{R}$ and the third one is by $\omega(\widetilde{\boldsymbol{V}}\boldsymbol{R}) = \omega(\widetilde{\boldsymbol{V}})$. □

Furthermore, we will utilize the following lemma in our analysis to relate the Procrustes distance between the algorithm iterations and the optimal solution to the values of the regularization function.

Lemma C.1.2. *For any convex regularization function $\omega : \mathbb{R}^{d\times r} \to \mathbb{R}$, any constant $\alpha > 0$, and any factor matrices $\boldsymbol{U}_1, \boldsymbol{U}_2 \in \mathbb{R}^{d\times r}$, it holds that*

$$\alpha\,\omega(\boldsymbol{U}_2) - \alpha\,\omega(\mathit{prox}_{\omega}(\boldsymbol{U}_1;\alpha)) \geq \langle \mathit{prox}_{\omega}(\boldsymbol{U}_1;\alpha) - \boldsymbol{U}_1, \mathit{prox}_{\omega}(\boldsymbol{U}_1;\alpha) - \boldsymbol{U}_2\rangle,$$

or equivalently, that

$$\begin{aligned}
&2\alpha\,\omega(\boldsymbol{U}_2) - 2\alpha\,\omega(\mathit{prox}_{\omega}(\boldsymbol{U}_1;\alpha)) \\
&\qquad \geq \|\mathit{prox}_{\omega}(\boldsymbol{U}_1;\alpha) - \boldsymbol{U}_1\|_F^2 + \|\mathit{prox}_{\omega}(\boldsymbol{U}_1;\alpha) - \boldsymbol{U}_2\|_F^2 - \|\boldsymbol{U}_2 - \boldsymbol{U}_1\|_F^2.
\end{aligned} \tag{C.1}$$

Proof. Writing the optimality condition for

$$\mathrm{prox}_{\omega}(\boldsymbol{U}_1;\alpha) = \operatorname*{argmin}_{\widetilde{\boldsymbol{U}}_1\in\mathbb{R}^{d\times r}} \frac{1}{2}\|\boldsymbol{U}_1 - \widetilde{\boldsymbol{U}}_1\|_F^2 + \alpha\,\omega(\widetilde{\boldsymbol{U}}_1)$$

implies

$$\boldsymbol{U}_1 - \mathrm{prox}_{\omega}(\boldsymbol{U}_1;\alpha) \in \alpha\,\partial\omega(\mathrm{prox}_{\omega}(\boldsymbol{U}_1;\alpha)),$$

where $\partial\omega(\mathrm{prox}_{\omega}(\boldsymbol{U}_1;\alpha))$ denotes the sub-differential set of the convex regularizer ω evaluated at $\mathrm{prox}_{\omega}(\boldsymbol{U}_1;\alpha)$. By the convexity of the regularizer ω we then have

$$\omega(\boldsymbol{U}_2) \geq \omega(\mathrm{prox}_{\omega}(\boldsymbol{U}_1;\alpha)) + \langle \boldsymbol{G}, \boldsymbol{U}_2 - \mathrm{prox}_{\omega}(\boldsymbol{U}_1;\alpha)\rangle,$$

for any $\boldsymbol{G} \in \partial\,\omega(\text{prox}_{\omega}(\boldsymbol{U}_1;\alpha))$. Setting $\boldsymbol{G} = \frac{1}{\alpha}(\boldsymbol{U}_1 - \text{prox}_{\omega}(\boldsymbol{U}_1;\alpha))$ would imply the first part of the claim. The second part of the Lemma simply follows by adding and subtracting the following terms

$$\|\boldsymbol{U}_2\|_F^2, \|\boldsymbol{U}_1\|_F^2, \text{ and } \|\text{prox}_{\omega}(\boldsymbol{U}_1;\alpha)\|_F^2$$

to the right-hand side expression of the first inequality and then completing the squares. □

We would like to note that, when ω is specialized to the indicator function of a convex set $\mathcal{C}$, and $\boldsymbol{U}_2$ is an arbitrary feasible point, i.e. $\boldsymbol{U}_2 \in \mathcal{C}$, then the above lemma would be reduced to the following form

$$\langle \text{proj}_{\mathcal{C}}(\boldsymbol{U}_1) - \boldsymbol{U}_1, \text{proj}_{\mathcal{C}}(\boldsymbol{U}_1) - \boldsymbol{U}_2 \rangle \leq 0.$$

This is because by the feasibility assumption of $\boldsymbol{U}_2$, along with the fact that $\text{proj}_{\mathcal{C}}(\boldsymbol{U}_1)$ is always a feasible point, we have $\omega(\boldsymbol{U}_2) - \omega(\text{proj}_{\mathcal{C}}(\boldsymbol{U}_1)) = 0$. This implication has been used in [120] for the convergence analysis of projected gradient descent algorithm (see Lemma 5.1 in there) in the context of similar low-rank factorization problem as well as in [127] (see Lemma 6.1).

The following Lemma will be useful to lower-bound the difference of regularization values as a scaling of the Procrustes distance.

Lemma C.1.3. *Suppose the regularization function $\omega(\cdot) : \mathbb{R}^{d\times r} \to \mathbb{R}$ meets the conditions of Assumption 3 in section 4.2 with respect to the subspace $\mathcal{S}$ in $\mathbb{R}^{d\times r}$, which contains $\mathcal{X}^*_{sym}$. Then, the following*

$$\omega(\boldsymbol{U}) - \omega(\boldsymbol{U}^*) \geq -\Psi_{\omega}(\mathcal{S})\, dist(\boldsymbol{U}, \boldsymbol{U}^*)$$

holds for any $\boldsymbol{U} \in \mathbb{R}^{d\times r}$ and $\boldsymbol{U}^* \in \mathcal{X}_{sym}$, where $\Psi_{\omega}(\mathcal{S})$ denotes the subspace compatibility constant of $\mathcal{S}$ with respect to ω, as defined by Definition 4.2.3.

Proof. Defining $\boldsymbol{R}^* = \text{argmin}_{\boldsymbol{R}\in\mathcal{O}_r} \|\boldsymbol{U} - \boldsymbol{U}^*\boldsymbol{R}\|_F$ and $\boldsymbol{\Delta} = \boldsymbol{U} - \boldsymbol{U}^*\boldsymbol{R}^*$, the proof proceeds

as follows

$$\begin{aligned}
\omega(\boldsymbol{U}) - \omega(\boldsymbol{U}^*) &\overset{(i)}{=} \omega(\boldsymbol{U}) - \omega(\boldsymbol{U}^*\boldsymbol{R}^*) \\
&\overset{(ii)}{=} \omega(\boldsymbol{U}^*\boldsymbol{R}^* + \boldsymbol{\Delta}) - \omega(\boldsymbol{U}^*\boldsymbol{R}^*) \\
&= \omega(\boldsymbol{U}^*\boldsymbol{R}^* + \boldsymbol{\Delta}_{\mathcal{S}} + \boldsymbol{\Delta}_{\mathcal{S}^\perp}) - \omega(\boldsymbol{U}^*\boldsymbol{R}^*) \\
&\overset{(iii)}{\geq} \omega(\boldsymbol{U}^*\boldsymbol{R}^* + \boldsymbol{\Delta}_{\mathcal{S}^\perp}) - \omega(\boldsymbol{\Delta}_{\mathcal{S}}) - \omega(\boldsymbol{U}^*\boldsymbol{R}^*) \\
&\overset{(iv)}{=} \omega(\boldsymbol{U}^*\boldsymbol{R}^*) + \omega(\boldsymbol{\Delta}_{\mathcal{S}^\perp}) - \omega(\boldsymbol{\Delta}_{\mathcal{S}}) - \omega(\boldsymbol{U}^*\boldsymbol{R}^*) \\
&= \omega(\boldsymbol{\Delta}_{\mathcal{S}^\perp}) - \omega(\boldsymbol{\Delta}_{\mathcal{S}}) \overset{(v)}{\geq} -\omega(\boldsymbol{\Delta}_{\mathcal{S}}) \\
&\overset{(vi)}{\geq} -\Psi_\omega(\mathcal{S})\|\boldsymbol{\Delta}_{\mathcal{S}}\|_F \\
&\overset{(vii)}{\geq} -\Psi_\omega(\mathcal{S})\|\boldsymbol{\Delta}\|_F
\end{aligned}$$

where (i) is by the rotation invariance condition of Assumption 3, (ii) is by the definition of $\boldsymbol{\Delta}$, (iii) uses the triangle inequality, (iv) employs the decomposability assumption along with that $\boldsymbol{U}^*\boldsymbol{R}^* \in \mathcal{X}^*_{\text{sym}} \subset \mathcal{S}$ and $\boldsymbol{\Delta}_{\mathcal{S}^\perp} \in \mathcal{S}^\perp$, (v) is by the non-negativity of ω, the definition of $\Psi_\omega(\mathcal{S})$ is utilized in (vi), and finally (vii) holds since $\|\boldsymbol{\Delta}_{\mathcal{S}}\|_F \leq \|\boldsymbol{\Delta}\|_F$. $\square$

The following lemma is a useful descent result that will be crucial in the proof of the linear convergence of our algorithm. Intuitively speaking, the lemma guarantees that moving along the negative gradient direction $\nabla\mathcal{L}_n(\boldsymbol{X})\boldsymbol{U}$ will reduce the remaining distance to the optimal point $\boldsymbol{U}^*$ since $\nabla\mathcal{L}_n(\boldsymbol{X})\boldsymbol{U}$ forms a positive inner product with the total distance to optimality $\boldsymbol{U} - \boldsymbol{U}^*\boldsymbol{R}^*$.

Lemma C.1.4. . *Let $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{U}^T$ be an arbitrary rank-r matrix, with $\boldsymbol{U} \in \mathbb{R}^{d\times r}$, which satisfies dist$(\boldsymbol{U}, \boldsymbol{U}^*) \leq \rho\,\sigma_r(\boldsymbol{U}^*)$, for $\rho \leq \sqrt{m/32M}$. Moreover, let the rotation matrix $\boldsymbol{R}^* \in \mathcal{O}_r$ be defined as $\boldsymbol{R}^* = \text{argmin}_{\boldsymbol{R}\in\mathcal{O}_r} \|\boldsymbol{U} - \boldsymbol{U}^*\boldsymbol{R}\|_F$. Then, under Assumption 1, we will have that*

$$\begin{aligned}
2\langle \nabla\mathcal{L}_n(\boldsymbol{X})\,\boldsymbol{U}, \boldsymbol{U} - \boldsymbol{U}^*\boldsymbol{R}^*\rangle \geq \frac{mM\cdot\sigma_r^2(\boldsymbol{U}^*)}{8(m+M)}\|\boldsymbol{\Delta}\|_F^2 - \frac{5r}{2m}\|\nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_2^2 \\
+ \frac{1}{2(m+M)}\|\nabla\mathcal{L}_n(\boldsymbol{X}) - \nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_F^2\,.
\end{aligned}$$

Proof. Before beginning the proof, we would like to state two Lemmata that we borrow from the existing literature. First, we will take advantage of the following Lemma, which

is Lemma 5.4 of [116] (after a slight simplification by noticing that $2 \geq 1/2(\sqrt{2}-1)$), to bound the Procrustes distance of factors in terms of the Euclidean distance between original low-rank matrices.

Lemma C.1.5. *For any $\boldsymbol{U} \in \mathbb{R}^{d\times r}$ we have*

$$dist^2(\boldsymbol{U}, \boldsymbol{U}^*) \leq \frac{2}{\sigma_r^2(\boldsymbol{U}^*)} \|\boldsymbol{U}\boldsymbol{U}^T - \boldsymbol{U}^*\boldsymbol{U}^{*T}\|_F^2.$$

Another Lemma, adopted from [118], will be useful to relate the singular values of the algorithm iterate $\boldsymbol{U}_t$ to those of $\boldsymbol{U}^*$, once $\boldsymbol{U}_t$ lies in a small neighborhood of $\boldsymbol{U}^*$.

Lemma C.1.6. *Let $\boldsymbol{U}$ and $\boldsymbol{U}^*$ be $d \times r$ matrices such that $dist(\boldsymbol{U}, \boldsymbol{U}^*) \leq \rho\, \sigma_r(\boldsymbol{U}^*)$, for $\rho \in (0, 1)$. Then, the following bounds hold true:*

$$(1-\rho)\sigma_1(\boldsymbol{U}^*) \leq \sigma_1(\boldsymbol{U}) \leq (1+\rho)\sigma_1(\boldsymbol{U}^*)$$
$$(1-\rho)\sigma_r(\boldsymbol{U}^*) \leq \sigma_r(\boldsymbol{U}) \leq (1+\rho)\sigma_r(\boldsymbol{U}^*).$$

We now have all the ingredients required to do the proof of Lemma C.1.4. To begin, inspired by the steps of similar proofs in [116, 118, 120, 122], we notice that

$$\begin{aligned}
2\langle \nabla\mathcal{L}_n(\boldsymbol{X})\,\boldsymbol{U}, \boldsymbol{U} - \boldsymbol{U}^*\boldsymbol{R}^*\rangle &= 2\,\langle (\nabla\mathcal{L}_n(\boldsymbol{X}) - \nabla\mathcal{L}_n(\boldsymbol{X}^*))\,\boldsymbol{U}, \boldsymbol{U} - \boldsymbol{U}^*\boldsymbol{R}^*\rangle \\
&\quad + 2\,\langle \nabla\mathcal{L}_n(\boldsymbol{X}^*)\,\boldsymbol{U}, \boldsymbol{U} - \boldsymbol{U}^*\boldsymbol{R}^*\rangle \\
&= \langle \nabla\mathcal{L}_n(\boldsymbol{X}) - \nabla\mathcal{L}_n(\boldsymbol{X}^*), \boldsymbol{X} - \boldsymbol{X}^*\rangle \\
&\quad + \langle \nabla\mathcal{L}_n(\boldsymbol{X}) - \nabla\mathcal{L}_n(\boldsymbol{X}^*), \boldsymbol{\Delta}\boldsymbol{\Delta}^T\rangle \\
&\quad + \langle \nabla\mathcal{L}_n(\boldsymbol{X}^*), \boldsymbol{X} - \boldsymbol{X}^*\rangle + \langle \nabla\mathcal{L}_n(\boldsymbol{X}^*), \boldsymbol{\Delta}\boldsymbol{\Delta}^T\rangle,
\end{aligned} \tag{C.2}$$

where $\boldsymbol{\Delta} := \boldsymbol{U} - \boldsymbol{U}^*\boldsymbol{R}^*$, and to derive the second equality we leveraged the equality $2(\boldsymbol{U} - \boldsymbol{U}^*\boldsymbol{R}^*)\boldsymbol{U}^T = \boldsymbol{X} - \boldsymbol{X}^* + \boldsymbol{\Delta}\boldsymbol{\Delta}^T$. For the first term of the above summation, we take advantage of that $\mathcal{L}_n$ meets RSC and RSM conditions to yield (via utilizing Theorem 2.1.11 in [14] and adapting it to the set of rank-r symmetric PSD matrices)

$$\begin{aligned}
\langle \nabla\mathcal{L}_n(\boldsymbol{X}) - \nabla\mathcal{L}_n(\boldsymbol{X}^*), \boldsymbol{X} - \boldsymbol{X}^*\rangle &\geq \frac{mM}{m+M}\|\boldsymbol{X} - \boldsymbol{X}^*\|_F^2 \\
&\quad + \frac{1}{m+M}\|\nabla\mathcal{L}_n(\boldsymbol{X}) - \nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_F^2.
\end{aligned}$$

For the second inner product of the summation beginning on (C.2), we simply use Cauchy-Schwartz inequality and the definition of matrix spectral norm to obtain

$$\begin{aligned}
\langle \nabla\mathcal{L}_n(\boldsymbol{X}) - \nabla\mathcal{L}_n(\boldsymbol{X}^*), \boldsymbol{\Delta}\boldsymbol{\Delta}^T\rangle &\geq -|\langle \nabla\mathcal{L}_n(\boldsymbol{X}) - \nabla\mathcal{L}_n(\boldsymbol{X}^*), \boldsymbol{\Delta}\boldsymbol{\Delta}^T\rangle| \\
&= -|\langle (\nabla\mathcal{L}_n(\boldsymbol{X}) - \nabla\mathcal{L}_n(\boldsymbol{X}^*))\boldsymbol{\Delta}, \boldsymbol{\Delta}\rangle| \\
&\overset{(i)}{\geq} -\|(\nabla\mathcal{L}_n(\boldsymbol{X}) - \nabla\mathcal{L}_n(\boldsymbol{X}^*))\boldsymbol{\Delta}\|_F \cdot \|\boldsymbol{\Delta}\|_F \\
&\overset{(ii)}{\geq} -\|\nabla\mathcal{L}_n(\boldsymbol{X}) - \nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_2 \cdot \|\boldsymbol{\Delta}\|_F^2 \\
&\overset{(iii)}{\geq} -\|\nabla\mathcal{L}_n(\boldsymbol{X}) - \nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_F \cdot \|\boldsymbol{\Delta}\|_F^2 \\
&\overset{(iv)}{\geq} -\frac{1}{2(m+M)}\|\nabla\mathcal{L}_n(\boldsymbol{X}) - \nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_F^2 - \frac{m+M}{2}\|\boldsymbol{\Delta}\|_F^4
\end{aligned}$$

where (i) is by Cauchy-Schwartz inequality, (ii) follows from that

$$\|(\nabla\mathcal{L}_n(\boldsymbol{X}) - \nabla\mathcal{L}_n(\boldsymbol{X}^*))\boldsymbol{\Delta}\|_F \leq \|\nabla\mathcal{L}_n(\boldsymbol{X}) - \nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_2 \cdot \|\boldsymbol{\Delta}\|_F,$$

the inequality in (iii) is because $\|\boldsymbol{M}\|_2 \leq \|\boldsymbol{M}\|_F$ for any matrix $\boldsymbol{M}$, and (iv) holds since $2ab \leq \beta a^2 + b^2/\beta$ is true for any $\beta > 0$, which upon setting $a = \|\nabla\mathcal{L}_n(\boldsymbol{X}) - \nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_F$, $b = \|\boldsymbol{\Delta}\|_F^2$, and $\beta = 1/(m+M)$ implies the inequality. Therefore, for the first two terms on the right-hand side of the equality in (C.2), we have thus far shown that

$$\begin{aligned}
\left\langle \nabla\mathcal{L}_n(\boldsymbol{X}) - \nabla\mathcal{L}_n(\boldsymbol{X}^*), \boldsymbol{X} - \boldsymbol{X}^* + \boldsymbol{\Delta}\boldsymbol{\Delta}^T \right\rangle \geq \frac{mM}{m+M}\|\boldsymbol{X} - \boldsymbol{X}^*\|_F^2 - \frac{m+M}{2}\|\boldsymbol{\Delta}\|_F^4 \\
+ \frac{1}{2(m+M)}\|\nabla\mathcal{L}_n(\boldsymbol{X}) - \nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_F^2 .
\end{aligned}$$

We approach the third term of the upper-bound in (C.2) as follows:

$$\begin{aligned}
\langle \nabla\mathcal{L}_n(\boldsymbol{X}^*), \boldsymbol{X} - \boldsymbol{X}^*\rangle &\geq -|\langle \nabla\mathcal{L}_n(\boldsymbol{X}^*), \boldsymbol{X} - \boldsymbol{X}^*\rangle| \\
&\geq -\|\boldsymbol{X} - \boldsymbol{X}^*\|_* \cdot \|\nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_2 \\
&\geq -\sqrt{2r}\,\|\boldsymbol{X} - \boldsymbol{X}^*\|_F \cdot \|\nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_2 \\
&\geq -\frac{mM}{2(m+M)}\|\boldsymbol{X} - \boldsymbol{X}^*\|_F^2 - \frac{(m+M)r}{Mm}\|\nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_2^2
\end{aligned}$$

where $\|\cdot\|_*$ denotes matrix nuclear norm, the third inequality leverages the facts that $\boldsymbol{X} - \boldsymbol{X}^*$ is of rank at most $2r$ as well as that $\|\boldsymbol{M}\|_* \leq \sqrt{s}\,\|\boldsymbol{M}\|_F$ for any rank-s matrix $\boldsymbol{M}$, and the last inequality is by using $ab \leq \frac{\beta}{2}a^2 + \frac{1}{2\beta}b^2$ with $a = \|\boldsymbol{X} - \boldsymbol{X}^*\|_F$, $b = \|\nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_2$, and $\beta = \frac{mM}{2(m+M)r}$.

For the last term of the upper-bound in (C.2), we derive the following bounds

$$\begin{aligned}\langle \nabla \mathcal{L}_n(\boldsymbol{X}^*), \boldsymbol{\Delta}\boldsymbol{\Delta}^T \rangle &\geq -|\langle \nabla \mathcal{L}_n(\boldsymbol{X}^*), \boldsymbol{\Delta}\boldsymbol{\Delta}^T \rangle| \\ &= -|\langle \nabla \mathcal{L}_n(\boldsymbol{X}^*)\boldsymbol{\Delta}, \boldsymbol{\Delta} \rangle| \\ &\geq -\|\nabla \mathcal{L}_n(\boldsymbol{X}^*)\boldsymbol{\Delta}\|_F \cdot \|\boldsymbol{\Delta}\|_F \\ &\geq -\|\nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_2 \cdot \|\boldsymbol{\Delta}\|_F^2 \\ &\geq \frac{-1}{2(m+M)} \|\nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_2^2 - \frac{m+M}{2} \|\boldsymbol{\Delta}\|_F^4.\end{aligned}$$

where the last inequality is again by applying $2ab \leq \beta a^2 + b^2/\beta$ with $a = \|\boldsymbol{\Delta}\|_F^2$, $b = \|\nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_2$, and $\beta = m + M$. Therefore we have shown that

$$\begin{aligned}2\langle \nabla \mathcal{L}_n(\boldsymbol{X})\, \boldsymbol{U}, \boldsymbol{U} - \boldsymbol{U}^* \boldsymbol{R}^* \rangle \geq{}& \frac{mM}{2(m+M)} \|\boldsymbol{X} - \boldsymbol{X}^*\|_F^2 - (m+M) \|\boldsymbol{\Delta}\|_F^4 \\ &+ \frac{1}{2(m+M)} \|\nabla \mathcal{L}_n(\boldsymbol{X}) - \nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_F^2 \\ &- \left[\frac{1}{2(m+M)} + \frac{(m+M)r}{mM}\right] \|\nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_2^2.\end{aligned}$$

Noticing that

$$\frac{1}{2(m+M)} + \frac{(m+M)r}{mM} \leq \frac{5r}{2m},$$

which holds since $m \leq M$ and $r \geq 1$, together with the fact that, by Lemma C.1.5, we have $\|\boldsymbol{X} - \boldsymbol{X}^*\|_F \geq \frac{\sigma_r^2(\boldsymbol{U}^*)}{2} \|\boldsymbol{\Delta}\|_F^2$, the yields that

$$\begin{aligned}2\langle \nabla \mathcal{L}_n(\boldsymbol{X})\, \boldsymbol{U}, \boldsymbol{U} - \boldsymbol{U}^* \boldsymbol{R}^* \rangle \geq{}& \frac{mM \cdot \sigma_r^2(\boldsymbol{U}^*)}{4(m+M)} \|\boldsymbol{\Delta}\|_F^2 - (m+M) \|\boldsymbol{\Delta}\|_F^4 \\ &+ \frac{1}{2(m+M)} \|\nabla \mathcal{L}_n(\boldsymbol{X}) - \nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_F^2 - \frac{5r}{2m} \|\nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_2^2.\end{aligned} \tag{C.3}$$

Assuming $\boldsymbol{U}$ satisfies $\mathrm{dist}^2(\boldsymbol{U}, \boldsymbol{U}^*) \leq \rho^2\, \sigma_r^2(\boldsymbol{U}^*)$, with $\rho^2 \leq mM/8(m+M)^2$, and given the notation $\mathrm{dist}(\boldsymbol{U}, \boldsymbol{U}^*) = \|\boldsymbol{U} - \boldsymbol{U}^* \boldsymbol{R}^*\|_F = \|\boldsymbol{\Delta}\|_F$, we get

$$(m+M)\|\boldsymbol{\Delta}\|_F^4 \leq (m+M)\rho^2 \sigma_r^2(\boldsymbol{U}^*) \|\boldsymbol{\Delta}\|_F^2 \leq \frac{mM \cdot \sigma_r^2(\boldsymbol{U}^*)}{8(m+M)} \|\boldsymbol{\Delta}\|_F^2.$$

By the relationship between the RSC and RSM constants, i.e. $m \leq M$, it follows that imposing $\rho^2 \leq m/32M$ ensures the condition $\rho^2 \leq mM/8(m+M)^2$. Therefore, the first two terms appearing on the right-hand side of (C.3) can be lower-bounded as

$$\frac{mM \cdot \sigma_r^2(\boldsymbol{U}^*)}{4(m+M)} \|\boldsymbol{\Delta}\|_F^2 - (m+M)\|\boldsymbol{\Delta}\|_F^4 \geq \frac{mM \cdot \sigma_r^2(\boldsymbol{U}^*)}{8(m+M)} \|\boldsymbol{\Delta}\|_F^2.$$

Incorporating this into (C.3) will complete the proof. □

Finally the last lemma in this section ensures the local smoothness of the loss function, when viewed as a function of $\boldsymbol{U}$.

Lemma C.1.7. *For any $\boldsymbol{U} \in \mathbb{R}^{d\times r}$, if we define $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{U}^T$ then we have that*

$$\|\nabla\mathcal{L}_n(\boldsymbol{X})\boldsymbol{U}\|_F^2 \leq 2\|\boldsymbol{U}\|_2^2\,\|\nabla\mathcal{L}_n(\boldsymbol{X}) - \nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_F^2 + 2r\|\boldsymbol{U}\|_2^2\,\|\nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_2^2.$$

Proof. The proof proceeds as follows

$$\begin{aligned}
\|\nabla\mathcal{L}_n(\boldsymbol{X})\boldsymbol{U}\|_F^2 &= \|\nabla\mathcal{L}_n(\boldsymbol{X})\boldsymbol{U} - \nabla\mathcal{L}_n(\boldsymbol{X}^*)\boldsymbol{U} + \nabla\mathcal{L}_n(\boldsymbol{X}^*)\boldsymbol{U}\|_F^2 \\
&\leq 2\,\|\nabla\mathcal{L}_n(\boldsymbol{X})\boldsymbol{U} - \nabla\mathcal{L}_n(\boldsymbol{X}^*)\boldsymbol{U}\|_F^2 + 2\,\|\nabla\mathcal{L}_n(\boldsymbol{X}^*)\boldsymbol{U}\|_F^2 \\
&\leq 2\|\boldsymbol{U}\|_2^2\,\|\nabla\mathcal{L}_n(\boldsymbol{X}) - \nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_F^2 + 2\|\boldsymbol{U}\|_F^2\|\nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_2^2 \\
&\leq 2\|\boldsymbol{U}\|_2^2\,\|\nabla\mathcal{L}_n(\boldsymbol{X}) - \nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_F^2 + 2r\|\boldsymbol{U}\|_2^2\|\nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_2^2,
\end{aligned}$$

where the first inequality is implied by that $\|\boldsymbol{A}+\boldsymbol{B}\|_F^2 \leq 2\|\boldsymbol{A}\|_F^2 + 2\|\boldsymbol{B}\|_F^2$ holds for any arbitrary pair $(\boldsymbol{A}, \boldsymbol{B})$ of same-sized matrices and the last inequality uses the rank$-r$ assumption on $\boldsymbol{U}$ to imply $\|\boldsymbol{U}\|_F^2 \leq r\,\|\boldsymbol{U}\|_2^2$. □

C.2 Symmetric Case

Proof of Lemma 4.3.1

First, we notice that

$$\begin{aligned}
\mathrm{dist}^2(\boldsymbol{U}_{t+1}, \boldsymbol{U}^*) &= \|\boldsymbol{U}_{t+1} - \boldsymbol{U}^*\boldsymbol{R}_{t+1}\|_F^2 \\
&= \min_{\boldsymbol{R}\in\mathcal{O}_r} \|\boldsymbol{U}_{t+1} - \boldsymbol{U}^*\boldsymbol{R}\|_F^2 \leq \|\boldsymbol{U}_{t+1} - \boldsymbol{U}^*\boldsymbol{R}_t\|_F^2,
\end{aligned}$$

where $\boldsymbol{R}_t = \mathrm{argmin}_{\boldsymbol{R}\in\mathcal{O}_r} \|\boldsymbol{U}_t - \boldsymbol{U}^*\boldsymbol{R}\|_F$. Invoking the inequality (C.1) in Lemma C.1.2, with setting $\boldsymbol{U}_2 = \boldsymbol{U}^*\boldsymbol{R}_t$, $\boldsymbol{U}_1 = \widetilde{\boldsymbol{U}}_{t+1}$, and $\alpha = \lambda\mu_t$, in there yields that

$$2\lambda\mu_t(\omega(\boldsymbol{U}^*) - \omega(\boldsymbol{U}_{t+1})) \geq \|\boldsymbol{U}_{t+1} - \widetilde{\boldsymbol{U}}_{t+1}\|_F^2 + \|\boldsymbol{U}_{t+1} - \boldsymbol{U}^*\boldsymbol{R}_t\|_F^2 - \|\widetilde{\boldsymbol{U}}_{t+1} - \boldsymbol{U}^*\boldsymbol{R}_t\|_F^2.$$

Upon re-arranging the terms of the above inequality and using the fact mentioned first in the proof, we obtain

$$\mathrm{dist}^2(\boldsymbol{U}_{t+1}, \boldsymbol{U}^*) \leq \|\widetilde{\boldsymbol{U}}_{t+1} - \boldsymbol{U}^*\boldsymbol{R}_t\|_F^2 + 2\lambda\mu_t(\omega(\boldsymbol{U}^*) - \omega(\boldsymbol{U}_{t+1})),$$

where to slightly simplify the analysis, we have dropped the negative term $-\|\boldsymbol{U}_{t+1} - \widetilde{\boldsymbol{U}}_{t+1}\|_F^2$ from the right-hand side of the inequality. To proceed, we bound the first term of the upper-bound by using the definition of $\widetilde{\boldsymbol{U}}_{t+1}$ as follows

$$
\begin{aligned}
\mathrm{dist}^2(\boldsymbol{U}_{t+1}, \boldsymbol{U}^*) &\leq \|\boldsymbol{U}_t - \mu_t \nabla \mathcal{L}_n(\boldsymbol{X}_t)\boldsymbol{U}_t - \boldsymbol{U}^*\boldsymbol{R}_t\|_F^2 - 2\lambda\mu_t \left(\omega(\boldsymbol{U}_{t+1}) - \omega(\boldsymbol{U}^*)\right) \\
&\overset{(i)}{=} \mathrm{dist}^2(\boldsymbol{U}_t, \boldsymbol{U}^*) - 2\mu_t \left\langle \nabla \mathcal{L}_n(\boldsymbol{X}_t)\boldsymbol{U}_t, \boldsymbol{U}_t - \boldsymbol{U}^*\boldsymbol{R}_t \right\rangle + \mu_t^2 \|\nabla \mathcal{L}_n(\boldsymbol{X}_t)\boldsymbol{U}_t\|_F^2 \\
&\quad - 2\lambda\mu_t \left(\omega(\boldsymbol{U}_{t+1}) - \omega(\boldsymbol{U}^*)\right) \\
&\overset{(ii)}{\leq} \mathrm{dist}^2(\boldsymbol{U}_t, \boldsymbol{U}^*) - 2\lambda\mu_t \left(\omega(\boldsymbol{U}_{t+1}) - \omega(\boldsymbol{U}^*)\right) \\
&\quad - \mu_t \left[\alpha_1 \mathrm{dist}^2(\boldsymbol{U}_t, \boldsymbol{U}^*) + \alpha_2 \|\nabla \mathcal{L}_n(\boldsymbol{X}_t) - \nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_F^2 - \alpha_3 \|\nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_2^2\right] \\
&\quad + 2\mu_t^2 \sigma_1^2(\boldsymbol{U}_t) \left[r \|\nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_2^2 + \|\nabla \mathcal{L}_n(\boldsymbol{X}_t) - \nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_F^2\right] \\
&\overset{(iii)}{=} (1 - \mu_t\alpha_1)\, \mathrm{dist}^2(\boldsymbol{U}_t, \boldsymbol{U}^*) - 2\lambda\mu_t \left(\omega(\boldsymbol{U}_{t+1}) - \omega(\boldsymbol{U}^*)\right) \\
&\quad + \left(2\mu_t^2 r\, \sigma_1^2(\boldsymbol{U}_t) + \mu_t\alpha_3\right) \|\nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_2^2 \\
&\quad + \left(2\mu_t^2\sigma_1^2(\boldsymbol{U}_t) - \mu_t\alpha_2\right) \|\nabla \mathcal{L}_n(\boldsymbol{X}) - \nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_F^2,
\end{aligned}
$$

where in (i) we are expanding the quadratic expression and replacing $\|\boldsymbol{U}_t - \boldsymbol{U}^*\boldsymbol{R}^*\|_F^2$ by $\mathrm{dist}^2(\boldsymbol{U}_t, \boldsymbol{U}^*)$, (ii) is by invoking Lemmata C.1.4 and C.1.7 and using the notation $\alpha_1 := mM\sigma_r^2(\boldsymbol{U}^*)/8(m+M)$, $\alpha_2 := 1/2(m+M)$, and $\alpha_3 := 5r/2m$, and (iii) is by simply rearranging the terms. Notice that the step-size assumption in (4.11), i.e.

$$
\mu_t \leq \frac{1}{4(m+M)\sigma_1^2(\boldsymbol{U}_t)} = \frac{\alpha_2}{2\sigma_1^2(\boldsymbol{U}_t)},
$$

implies that the term dependent on $\|\nabla \mathcal{L}(\boldsymbol{X}_t) - \nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_F^2$ in the above upper bound on $\mathrm{dist}(\boldsymbol{U}_{t+1}, \boldsymbol{U}^*)$ becomes negative and can be dropped. These simplifications, along with defining $\eta_t^2 := \left(2\mu_t^2 r\, \sigma_1^2(\boldsymbol{U}_t) + \alpha_3\mu_t\right) \|\nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_2^2$, yield

$$
\mathrm{dist}^2(\boldsymbol{U}_{t+1}, \boldsymbol{U}^*) + 2\lambda\mu_t \left(\omega(\boldsymbol{U}_{t+1}) - \omega(\boldsymbol{U}^*)\right) \leq (1 - \mu_t\alpha_1)\, \mathrm{dist}^2(\boldsymbol{U}_t, \boldsymbol{U}^*) + \eta_t^2.
$$

To complete the proof, we only need to simplify the expression for η_t^2 as follows

$$\begin{aligned}
\eta_t^2 &= \left(2\mu_t^2\, r\, \sigma_1^2(\boldsymbol{U}_t) + \mu_t \alpha_3\right) \|\nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_2^2 \\
&\overset{(i)}{=} \left(2\mu_t^2\, r\, \sigma_1^2(\boldsymbol{U}_t) + \frac{5r\mu_t}{2m}\right) \|\nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_2^2 \\
&\overset{(ii)}{\leq} \left[\frac{r}{8\sigma_1^2(\boldsymbol{U}_t)(m+M)^2} + \frac{5r}{8m\,\sigma_1^2(\boldsymbol{U}_t)(m+M)}\right] \|\nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_2^2 \\
&\overset{(iii)}{\leq} \left[\frac{r}{16m\,\sigma_1^2(\boldsymbol{U}_t)(m+M)} + \frac{5r}{8m\,\sigma_1^2(\boldsymbol{U}_t)(m+M)}\right] \|\nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_2^2 \\
&\overset{(iv)}{\leq} \left(\frac{1}{16} + \frac{5}{8}\right) \frac{1}{(1-\rho)^2} \cdot \frac{r}{m\,\sigma_1^2(\boldsymbol{U}^*)(m+M)} \|\nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_2^2 \\
&\overset{(v)}{\leq} \frac{r\, \|\nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_2^2}{m(m+M)\,\sigma_1^2(\boldsymbol{U}^*)},
\end{aligned}$$

where (i) is by the definition $\alpha_3 = 5r/2m$, (ii) is by assumption $\mu_t \leq 4(m+M)\sigma_1^2(\boldsymbol{U}_t)$, (iii) simply replaces $m+M$ by m leverages the result of Lemma C.1.6 to imply $\sigma_1(\boldsymbol{U}_t) \geq (1-\rho)\sigma_1(\boldsymbol{U}^*)$, the inequality in (iii) simply holds since $2m \leq m+M$, the inequality in (iv) exploits Lemma C.1.6 to imply $(1-\rho)\,\sigma_r(\boldsymbol{U}^*) \leq \sigma_1(\boldsymbol{U}_t)$, and finally (v) is true by the fact that $\rho \leq \sqrt{1/32}$, as stated in the Lemma.

Proof of Lemma 4.3.2

Upon incorporating Lemma C.1.3 into the result of Lemma 4.3.1, we will obtain that

$$\begin{aligned}
\operatorname{dist}^2(\boldsymbol{U}_{t+1}, \boldsymbol{U}^*) - 2\mu_t\, \tau \Psi(\mathcal{S}) \operatorname{dist}(\boldsymbol{U}_{t+1}, \boldsymbol{U}^*) &\leq \operatorname{dist}^2(\boldsymbol{U}_{t+1}, \boldsymbol{U}^*) + 2\mu_t \tau \left(\omega(\boldsymbol{U}_{t+1}) - \omega(\boldsymbol{U}^*)\right) \\
&\leq (1 - \mu_t \alpha_1) \operatorname{dist}^2(\boldsymbol{U}_t, \boldsymbol{U}^*) + \eta^2.
\end{aligned}$$

By finding the positive root of the quadratic equality associated with the last inequality and imposing $\operatorname{dist}(\boldsymbol{U}_{t+1}, \boldsymbol{U}^*)$ to be smaller than that, we derive the following inequality as the condition to make the inequality hold

$$\begin{aligned}
\operatorname{dist}(\boldsymbol{U}_{t+1}, \boldsymbol{U}^*) &\leq \mu_t\, \tau \Psi(\mathcal{S}) + \sqrt{\mu_t^2\, \tau^2 \Psi^2(\mathcal{S}) + (1 - \mu_t \alpha_1)\, \operatorname{dist}^2(\boldsymbol{U}_t, \boldsymbol{U}^*) + \eta^2} \\
&\overset{(i)}{\leq} 2\mu_t\, \tau \Psi(\mathcal{S}) + \sqrt{(1 - \mu_t \alpha_1) \operatorname{dist}^2(\boldsymbol{U}_t, \boldsymbol{U}^*) + \eta^2} \\
&\overset{(ii)}{\leq} \sqrt{1 - \mu_t \alpha_1} \cdot \operatorname{dist}(\boldsymbol{U}_t, \boldsymbol{U}^*) + 2\mu_t\, \tau \Psi(\mathcal{S}) + \eta,
\end{aligned}$$

where the inequalities in (i) and (ii) both utilize the fact that, for non-negative numbers a and b, it always holds $\sqrt{a^2+b^2} \leq a+b$. Here we are specializing to $a = \mu_t \tau \Psi(\mathcal{S})$ and

$$b = \sqrt{(1-\mu_t\alpha_1)\operatorname{dist}^2(\boldsymbol{U}_t, \boldsymbol{U}^*) + \eta^2}$$

to infer (i) and to $a = \sqrt{1-\mu_t\alpha_1} \cdot \operatorname{dist}(\boldsymbol{U}_t, \boldsymbol{U}^*)$ and $b = \eta$ to yield the inequality (ii).

C.3 Non-PSD Case

Before beginning the proofs and in order to ease the presentation, we introduce a few pieces of notation that will be frequently used throughout the section. First, the *lifted* factor matrices $\boldsymbol{W}$ and $\boldsymbol{W}^*$ are defined as follows

$$\boldsymbol{W} := \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} \in \mathbb{R}^{(d_1+d_2)\times r}, \quad \boldsymbol{W}^* := \begin{bmatrix} \boldsymbol{U}^* \\ \boldsymbol{V}^* \end{bmatrix} \in \mathbb{R}^{(d_1+d_2)\times r},$$

where $(\boldsymbol{U}, \boldsymbol{V}) \in \mathbb{R}^{d_1\times r} \times \mathbb{R}^{d_2 \times r}$ is an arbitrary pair of factors and $(\boldsymbol{U}^*, \boldsymbol{V}^*) \in \mathcal{X}^*$ denotes an equally-footed factorization of $\boldsymbol{X}^*$. Given such lifted matrices, we notice that

$$\begin{aligned} \operatorname{dist}(\boldsymbol{W}; \boldsymbol{W}^*) &:= \min_{\boldsymbol{R}\in\mathcal{O}_r} \|\boldsymbol{W} - \boldsymbol{W}^*\boldsymbol{R}\|_F \\ &= \min_{\boldsymbol{R}\in\mathcal{O}_r} \left\| \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix} - \begin{bmatrix} \boldsymbol{U}^*\boldsymbol{R} \\ \boldsymbol{V}^*\boldsymbol{R} \end{bmatrix} \right\|_F \\ &= \operatorname{dist}(\boldsymbol{U}, \boldsymbol{V}; \boldsymbol{U}^*, \boldsymbol{V}^*). \end{aligned} \tag{C.4}$$

Similar to [18], another set of stacked matrices that will be useful in our analysis is

$$\boldsymbol{Y} := \begin{bmatrix} \boldsymbol{U} \\ -\boldsymbol{V} \end{bmatrix} \in \mathbb{R}^{(d_1+d_2)\times r}, \text{ and } \boldsymbol{Y}^* := \begin{bmatrix} \boldsymbol{U}^* \\ -\boldsymbol{V}^* \end{bmatrix} \in \mathbb{R}^{(d_1+d_2)\times r}.$$

Assuming $\boldsymbol{R}^* \in \mathcal{O}_r$ denotes the orthonormal matrix attaining the minimum value of the optimization problem (C.4), the following error matrices often appear in our analysis

$$\boldsymbol{\Delta}_U := \boldsymbol{U} - \boldsymbol{U}^*\boldsymbol{R}^*, \quad \boldsymbol{\Delta}_V := \boldsymbol{V} - \boldsymbol{V}^*\boldsymbol{R}^*, \quad \boldsymbol{\Delta}_W := \boldsymbol{W} - \boldsymbol{W}^*\boldsymbol{R}^*, \quad \boldsymbol{\Delta}_Y := \boldsymbol{Y} - \boldsymbol{Y}^*\boldsymbol{R}^*.$$

We note that, by defining $\boldsymbol{R}^*$ as the minimizer in (C.4), we have $\|\boldsymbol{\Delta}_W\|_F = \operatorname{dist}(\boldsymbol{W}; \boldsymbol{W}^*)$. Moreover, the matrix $\boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{V}^T\boldsymbol{V} \in \mathbb{R}^{r\times r}$, which arises in the regularized estimator analysis, will be compactly denoted by $\boldsymbol{Z}$.

For any $t > 0$, $\boldsymbol{R}_t$ will denote the optimal rotation matrix corresponding to the t-th iteration of the proximal descent method, i.e.

$$\boldsymbol{R}_t := \underset{\boldsymbol{R}\in\mathcal{O}_r}{\operatorname{argmin}} \left\| \begin{bmatrix} \boldsymbol{U}_t \\ \boldsymbol{V}_t \end{bmatrix} - \begin{bmatrix} \boldsymbol{U}^*\boldsymbol{R} \\ \boldsymbol{V}^*\boldsymbol{R} \end{bmatrix} \right\|_F . \tag{C.5}$$

Finally, we often prefer to simply denote $\operatorname{dist}(\boldsymbol{U}_t, \boldsymbol{V}_t; \boldsymbol{U}^*, \boldsymbol{V}^*)$ by d_t, represent the partial derivatives of the regularization term $g(\boldsymbol{Z}_t)$, where $\boldsymbol{Z}_t := \boldsymbol{U}_t^T\boldsymbol{U}_t - \boldsymbol{V}_t^T\boldsymbol{V}_t$, by the following compact notations

$$\nabla_U g(\boldsymbol{Z}_t) := 2\boldsymbol{U}_t \cdot \nabla g(\boldsymbol{Z}_t)$$
$$\nabla_V g(\boldsymbol{Z}_t) := -2\boldsymbol{V}_t \cdot \nabla g(\boldsymbol{Z}_t),$$

and use $M_{\max} := \max\{M, M_g\}$ and $m_{\min} := \min\{m, m_g\}$.

Since the results for non-PSD matrix factorization are essentially derived by generalizing those for PSD matrices, we first demonstrate how some of the auxiliary Lemmata that were stated earlier can be extended into the general non-PSD case. In particular, the following Lemma, which can be viewed as an extension of Lemma C.1.4 for PSD matrices, will be crucial in the proofs of this section. Intuitively speaking, the Lemma guarantees that moving along the negative gradient direction of the smooth component of the regularized objective function will reduce the distance to the optimal point.

Lemma C.3.1. *Assume the pair $(\boldsymbol{U}, \boldsymbol{V}) \in \mathbb{R}^{d_1\times r} \times \mathbb{R}^{d_2\times r}$ is such that*

$$\textit{dist}(\boldsymbol{U}, \boldsymbol{V}; \boldsymbol{U}^*, \boldsymbol{V}^*) \leq \rho\sqrt{\sigma_r(\boldsymbol{X}^*)},$$

where $\rho^2 = m_{\min}/68M_{\max}$, and that $\lambda = 1/8$ in the statement of problem (4.5)*. Then, under Assumptions 1 and 2, the following inequality holds true*

$$\begin{aligned} \langle \nabla_U \mathcal{L}_n(\boldsymbol{X}), \boldsymbol{\Delta}_U\rangle + \langle \nabla_V \mathcal{L}_n(\boldsymbol{X}), \boldsymbol{\Delta}_V\rangle + \lambda\langle \nabla_U g(\boldsymbol{Z}), \boldsymbol{\Delta}_U\rangle + \lambda\langle \nabla_V g(\boldsymbol{Z}), \boldsymbol{\Delta}_V\rangle \\ \geq \frac{m_{\min}}{32}\sigma_r(\boldsymbol{X}^*)\|\boldsymbol{\Delta}_W\|_F^2 + \frac{1}{4M}\|\nabla\mathcal{L}_n(\boldsymbol{X}) - \nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_F^2 \\ + \frac{1}{32\,M_g}\|\nabla g(\boldsymbol{Z})\|_F^2 - \frac{5r}{2m}\|\nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_2^2 . \end{aligned} \tag{C.6}$$

The next Lemma, which is proved near the end of this section, is analogous to Lemma C.1.7 for PSD matrices and crucial for ensuring the local smoothness of the regularized loss function, when viewed as a function of $(\boldsymbol{U}, \boldsymbol{V})$.

Lemma C.3.2. *For any* $(\boldsymbol{U}, \boldsymbol{V}) \in \mathbb{R}^{d_1 \times r} \times \mathbb{R}^{d_2 \times r}$, *let* $\boldsymbol{W} := \begin{bmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{bmatrix}$, $\boldsymbol{X} := \boldsymbol{U}\boldsymbol{V}^T$, *and* $\boldsymbol{Z} := \boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{V}^T\boldsymbol{V}$. *It can be shown that*

$$\begin{aligned} \|\nabla_U \mathcal{L}_n(\boldsymbol{X}) + \lambda \nabla_U g(\boldsymbol{Z})\|_F^2 + \|\nabla_V \mathcal{L}_n(\boldsymbol{X}) + \lambda \nabla_V g(\boldsymbol{Z})\|_F^2 \\ \leq 8\|\boldsymbol{W}\|_2^2 \cdot \|\nabla \mathcal{L}_n(\boldsymbol{X}) - \nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_F^2 \\ + 4\|\boldsymbol{W}\|_2^2 \cdot \left(r\|\nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_2^2 + 4\lambda^2 \|\nabla g(\boldsymbol{Z})\|_F^2\right). \end{aligned} \tag{C.7}$$

With the auxiliary Lemmata C.3.1 and C.3.2 stated, the proof of the first result claimed for the general non-PSD factorization case can be detailed as follows.

Proof of Lemma 4.4.1

The proof mimics the steps of the proving Lemma 4.3.1 for PSD matrices. Given the definition of the orthonormal matrices $\boldsymbol{R}_{t+1}$ and $\boldsymbol{R}_t$, as in (C.5), we have

$$\begin{aligned} \operatorname{dist}^2(\boldsymbol{U}_{t+1}, \boldsymbol{V}_{t+1}; \boldsymbol{U}^*, \boldsymbol{V}^*) &= \left\| \begin{bmatrix} \boldsymbol{U}_{t+1} \\ \boldsymbol{V}_{t+1} \end{bmatrix} - \begin{bmatrix} \boldsymbol{U}^*\boldsymbol{R}_{t+1} \\ \boldsymbol{V}^*\boldsymbol{R}_{t+1} \end{bmatrix} \right\|_F^2 \\ &\leq \left\| \begin{bmatrix} \boldsymbol{U}_{t+1} \\ \boldsymbol{V}_{t+1} \end{bmatrix} - \begin{bmatrix} \boldsymbol{U}^*\boldsymbol{R}_{t} \\ \boldsymbol{V}^*\boldsymbol{R}_{t} \end{bmatrix} \right\|_F^2 \\ &= \|\boldsymbol{U}_{t+1} - \boldsymbol{U}^*\boldsymbol{R}_t\|_F^2 + \|\boldsymbol{V}_{t+1} - \boldsymbol{V}^*\boldsymbol{R}_t\|_F^2. \end{aligned} \tag{C.8}$$

We begin the proof by twice invoking Lemma C.1.2. First, setting $\boldsymbol{U}_2 = \boldsymbol{U}^*\boldsymbol{R}_t$, $\boldsymbol{U}_1 = \widetilde{\boldsymbol{U}}_{t+1}$, and $\alpha = \tau\mu_t$, in the statement of the inequality (C.1) of the Lemma yields

$$\|\boldsymbol{U}_{t+1} - \boldsymbol{U}^*\boldsymbol{R}_t\|_F^2 \leq \|\widetilde{\boldsymbol{U}}_{t+1} - \boldsymbol{U}^*\boldsymbol{R}_t\|_F^2 - \|\boldsymbol{U}_{t+1} - \widetilde{\boldsymbol{U}}_{t+1}\|_F^2 + 2\tau\mu_t(\omega_1(\boldsymbol{U}^*) - \omega_1(\boldsymbol{U}_{t+1})),$$

where we have also leveraged Assumption 3 to imply $\omega_1(\boldsymbol{U}^*\boldsymbol{R}_t) = \omega_1(\boldsymbol{U}^*)$ and the definition of $\boldsymbol{U}_{t+1}$ which gives $\boldsymbol{U}_{t+1} = \operatorname{prox}_{\omega_1}(\widetilde{\boldsymbol{U}}_{t+1}; \tau\,\mu_t)$. Second time, we set $\boldsymbol{V} = \boldsymbol{V}^*\boldsymbol{R}_t$, $\widetilde{\boldsymbol{V}} = \widetilde{\boldsymbol{V}}_{t+1}$, and $\alpha = \tau\mu_t$ to obtain

$$\|\boldsymbol{V}_{t+1} - \boldsymbol{V}^*\boldsymbol{R}_t\|_F^2 \leq \|\widetilde{\boldsymbol{V}}_{t+1} - \boldsymbol{V}^*\boldsymbol{R}_t\|_F^2 - \|\boldsymbol{V}_{t+1} - \widetilde{\boldsymbol{V}}_{t+1}\|_F^2 + 2\tau\mu_t(\omega_2(\boldsymbol{V}^*) - \omega_2(\boldsymbol{V}_{t+1})).$$

Incorporating the last two inequalities into (C.8) yields

$$\begin{aligned} d_{t+1}^2 \leq \|\widetilde{\boldsymbol{U}}_{t+1} - \boldsymbol{U}^*\boldsymbol{R}_t\|_F^2 + \|\widetilde{\boldsymbol{V}}_{t+1} - \boldsymbol{V}^*\boldsymbol{R}_t\|_F^2 \\ - 2\tau\mu_t\left(\omega_1(\boldsymbol{U}_{t+1}) - \omega_1(\boldsymbol{U}^*) + \omega_2(\boldsymbol{V}_{t+1}) - \omega_2(\boldsymbol{V}^*)\right), \end{aligned} \tag{C.9}$$

where, to slightly simplify the analysis, we have dropped the terms $-\|\boldsymbol{U}_{t+1} - \widetilde{\boldsymbol{U}}_{t+1}\|_F^2$ and $-\|\boldsymbol{V}_{t+1} - \widetilde{\boldsymbol{V}}_{t+1}\|_F^2$ from the right-hand side of the inequality and used the notation $d_{t+1} = \mathrm{dist}(\boldsymbol{U}_{t+1}, \boldsymbol{V}_{t+1}; \boldsymbol{U}^*, \boldsymbol{V}^*)$.

To proceed, we replace $\widetilde{\boldsymbol{U}}_{t+1}$ and $\widetilde{\boldsymbol{V}}_{t+1}$ in (C.9) with their definitions as given by the update rule of the proximal descent method in Algorithm 6 and write

$$
\begin{aligned}
d_{t+1}^2 \leq\ & \|\boldsymbol{U}_t - \mu_t \nabla_U \mathcal{L}_n(\boldsymbol{X}_t) - \mu_t \lambda \nabla_U g(\boldsymbol{Z}_t) - \boldsymbol{U}^* \boldsymbol{R}_t\|_F^2 && \text{(C.10)}\\
&+ \|\boldsymbol{V}_t - \mu_t \nabla_V \mathcal{L}_n(\boldsymbol{X}_t) - \mu_t \lambda \nabla_V g(\boldsymbol{Z}_t) - \boldsymbol{V}^* \boldsymbol{R}_t\|_F^2 && \text{(C.11)}\\
&- 2\tau\mu_t \left(\omega_1(\boldsymbol{U}_{t+1}) - \omega_1(\boldsymbol{U}^*) + \omega_2(\boldsymbol{V}_{t+1}) - \omega_2(\boldsymbol{V}^*)\right)\\
&= d_t^2 + \mu_t^2 \|\nabla_U \mathcal{L}_n(\boldsymbol{X}_t) + \lambda \nabla_U g(\boldsymbol{Z}_t)\|_F^2 + \mu_t^2 \|\nabla_V \mathcal{L}_n(\boldsymbol{X}_t) + \lambda \nabla_V g(\boldsymbol{Z}_t)\|_F^2 && \text{(C.12)}\\
&- 2\mu_t \left\langle \nabla_U \mathcal{L}_n(\boldsymbol{X}_t), \boldsymbol{U}_t - \boldsymbol{U}^* \boldsymbol{R}_t \right\rangle - 2\mu_t \left\langle \nabla_V \mathcal{L}_n(\boldsymbol{X}_t), \boldsymbol{V}_t - \boldsymbol{V}^* \boldsymbol{R}_t \right\rangle && \text{(C.13)}\\
&- 2\lambda\mu_t \left\langle \nabla_U g(\boldsymbol{Z}_t), \boldsymbol{U}_t - \boldsymbol{U}^* \boldsymbol{R}_t \right\rangle - 2\lambda\mu_t \left\langle \nabla_V g(\boldsymbol{Z}_t), \boldsymbol{V}_t - \boldsymbol{V}^* \boldsymbol{R}_t \right\rangle && \text{(C.14)}\\
&- 2\tau\mu_t \left(\omega_1(\boldsymbol{U}_{t+1}) - \omega_1(\boldsymbol{U}^*) + \omega_2(\boldsymbol{V}_{t+1}) - \omega_2(\boldsymbol{V}^*)\right)
\end{aligned}
$$

where $\boldsymbol{X}_t = \boldsymbol{U}_t \boldsymbol{V}_t^T$, $\boldsymbol{Z}_t = \boldsymbol{U}_t^T \boldsymbol{U}_t - \boldsymbol{V}_t^T \boldsymbol{V}_t$, and the equality follows by expanding the quadratic expressions in (C.10) and (C.11) and noticing that

$$
d_t^2 = \|\boldsymbol{U}_t - \boldsymbol{U}^* \boldsymbol{R}_t\|_F^2 + \|\boldsymbol{V}_t - \boldsymbol{V}^* \boldsymbol{R}_t\|_F^2 .
$$

Incorporating Lemma C.3.1 and Lemma C.3.2, with $\boldsymbol{U} = \boldsymbol{U}_t$ and $\boldsymbol{V} = \boldsymbol{V}_t$, into the last upper bound derived for d_{t+1} we get

$$
\begin{aligned}
d_{t+1}^2 \leq\ & d_t^2 \cdot \left[1 - \frac{\mu_t\, m_{\min}}{16} \sigma_r(\boldsymbol{X}^*)\right]\\
&+ \|\nabla \mathcal{L}_n(\boldsymbol{X}_t) - \nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_F^2 \cdot \left[8\mu_t^2 \|\boldsymbol{W}_t\|_2^2 - \frac{\mu_t}{2M}\right]\\
&+ \|\nabla g(\boldsymbol{Z}_t)\|_F^2 \cdot \left[\frac{\mu_t^2}{4} \|\boldsymbol{W}_t\|_2^2 - \frac{\mu_t}{16\, M_g}\right]\\
&+ \|\nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_2^2 \cdot \left[4r\mu_t^2 \|\boldsymbol{W}_t\|_2^2 + \frac{5r\mu_t}{m}\right]\\
&- 2\tau\mu_t \cdot \left[\omega_1(\boldsymbol{U}_{t+1}) - \omega_1(\boldsymbol{U}^*) + \omega_2(\boldsymbol{V}_{t+1}) - \omega_2(\boldsymbol{V}^*)\right],
\end{aligned}
$$

where we have rearranged the terms to obtain the above form of the resulting upper bound. Upon setting

$$
\mu_t \leq \frac{1}{16 M_{\max} \|\boldsymbol{W}_t\|_2^2},
$$

where $M_{\max} := \max\{M, M_g\}$, the scalings corresponding to the terms depending on $\|\nabla \mathcal{L}_n(\boldsymbol{X}_t) - \nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_F^2$ and $\|\nabla g(\boldsymbol{Z}_t)\|_F^2$ will become negative. Therefore, those terms can be removed from the upper bound to obtain

$$\begin{aligned} d_{t+1}^2 \leq d_t^2 \cdot \Big[1 - \frac{\mu_t\, m_{\min}}{16}\sigma_r(\boldsymbol{X}^*)\Big] \\ + \|\nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_2^2 \cdot \left[4r\mu_t^2\, \|\boldsymbol{W}_t\|_2^2 + \frac{5r\mu_t}{m}\right] \\ - 2\tau\mu_t \cdot \left[\omega_1(\boldsymbol{U}_{t+1}) - \omega_1(\boldsymbol{U}^*) + \omega_2(\boldsymbol{V}_{t+1}) - \omega_2(\boldsymbol{V}^*)\right]. \end{aligned}$$

Next, we can upper bound the expression in brackets that is multiplied by $\|\nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_2^2$ in the above via utilizing the condition earlier imposed on μ_t as follows

$$\begin{aligned} 4r\mu_t^2\, \|\boldsymbol{W}_t\|_2^2 + \frac{5r\mu_t}{m} &\leq \frac{r}{64M_{\max}^2\|\boldsymbol{W}_t\|_2^2} + \frac{5r}{16mM_{\max}\|\boldsymbol{W}_t\|_2^2} \\ &\overset{(i)}{\leq} \frac{21\, r}{64\, mM_{\max}\|\boldsymbol{W}_t\|_2^2} \\ &\overset{(ii)}{\leq} \frac{21\, r}{64\, mM_{\max}(1-\rho)^2\|\boldsymbol{W}^*\|_2^2} \\ &\overset{(iii)}{=} \frac{21\, r}{128\, mM_{\max}(1-\rho)^2\|\boldsymbol{X}^*\|_2} \\ &\leq \frac{r}{4\, mM_{\max}\|\boldsymbol{X}^*\|_2}, \end{aligned}$$

where (i) is by that $M_{\max}^2 \geq mM_{\max}$, (ii) utilizes the result of Lemma C.1.6 (upon setting $\boldsymbol{U}$ to $\boldsymbol{W}$ in there) to imply $\|\boldsymbol{W}_t\|_2 \geq (1-\rho)\|\boldsymbol{W}^*\|_2$, the equality in (iii) leverages the fact that $\|\boldsymbol{W}^*\|_2^2 = 2\|\boldsymbol{X}^*\|_2$, which holds since $(\boldsymbol{U}^*, \boldsymbol{V}^*) \in \mathcal{X}^*$ is an equally-footed factorization, and finally the last inequality uses the assumption of the Lemma on ρ to imply $\rho \leq 1/68$. This completes the proof of the Lemma.

Proof of Lemma 4.4.2

We begin the proof by utilizing Lemma C.1.3 to imply

$$\omega_1(\boldsymbol{U}_{t+1}) - \omega_1(\boldsymbol{U}^*) \geq -\Psi(\mathcal{S}_1)\|\boldsymbol{U}_{t+1} - \boldsymbol{U}^*\boldsymbol{R}_{t+1}\|_F,$$

where $\Psi(\mathcal{S}_1)$ is the subspace compatibility constant of $\mathcal{S}_1 \subseteq \mathbb{R}^{d_1 \times r}$ with respect to the regularizer ω_1. The same Lemma can be applied to the other regularizer ω_2 to obtain

$$\omega_2(\boldsymbol{V}_{t+1}) - \omega_2(\boldsymbol{V}^*) \geq -\Psi(\mathcal{S}_2)\|\boldsymbol{V}_{t+1} - \boldsymbol{V}^*\boldsymbol{R}_{t+1}\|_F,$$

where $\Psi(\mathcal{S}_2)$ denotes the subspace compatibility constant of $\mathcal{S}_2 \subseteq \mathbb{R}^{d_2 \times r}$ with respect to ω_2. Adding up the two inequalities would imply that

$$\begin{aligned}
\omega_1(\boldsymbol{U}_{t+1}) - \omega_1(\boldsymbol{U}^*) + \omega_2(\boldsymbol{V}_{t+1}) - \omega_2(\boldsymbol{V}^*) & \\
&\geq -\Psi(\mathcal{S}_1)\|\boldsymbol{U}_{t+1} - \boldsymbol{U}^*\boldsymbol{R}_{t+1}\|_F - \Psi(\mathcal{S}_2)\|\boldsymbol{V}_{t+1} - \boldsymbol{V}^*\boldsymbol{R}_{t+1}\|_F \\
&\geq -\left(\Psi(\mathcal{S}_1) + \Psi(\mathcal{S}_2)\right)\|\boldsymbol{W}_{t+1} - \boldsymbol{W}^*\boldsymbol{R}_{t+1}\|_F \\
&= -\left(\Psi(\mathcal{S}_1) + \Psi(\mathcal{S}_2)\right)\, d_{t+1},
\end{aligned}$$

where the second inequality is implied by that

$$\begin{aligned}
\|\boldsymbol{W}_{t+1} - \boldsymbol{W}^*\boldsymbol{R}_{t+1}\|_F &= \left\| \begin{bmatrix} \boldsymbol{U}_{t+1} - \boldsymbol{U}^*\boldsymbol{R}_{t+1} \\ \boldsymbol{V}_{t+1} - \boldsymbol{V}^*\boldsymbol{R}_{t+1} \end{bmatrix} \right\|_F \\
&\geq \max\left\{\|\boldsymbol{U}_{t+1} - \boldsymbol{U}^*\boldsymbol{R}_{t+1}\|_F, \|\boldsymbol{V}_{t+1} - \boldsymbol{V}^*\boldsymbol{R}_{t+1}\|_F\right\}.
\end{aligned}$$

Using this inequality together with the result of Lemma 4.4.1 leads to the following

$$d_{t+1}^2 - 2\mu_t\tau\left(\Psi(\mathcal{S}_1) + \Psi(\mathcal{S}_2)\right) d_{t+1} \leq (1 - \mu_t\,\alpha)\; d_t^2 + \eta_t^2. \tag{C.15}$$

Defining $b := \mu_t\tau\left(\Psi(\mathcal{S}_1) + \Psi(\mathcal{S}_2)\right)$, $c_1 := \sqrt{1-\mu_t\,\alpha}\cdot d_t$, and $c_2 := \eta_t$, the quadratic equality (in terms of d_{t+1}) associated with the above inequality takes the following form

$$d_{t+1}^2 - 2b\, d_{t+1} - (c_1^2 + c_2^2) = 0.$$

The positive root of the above equality is then given by $d_{t+1}^* = b + \sqrt{b^2 + c_1^2 + c_2^2}$. Any d_{t+1} satisfying the inequality in (C.15) has to necessarily be smaller than d_{t+1}^*, which in turn meets the following set of inequalities

$$\begin{aligned}
d_{t+1}^* &= b + \sqrt{b^2 + c_1^2 + c_2^2} \\
&\leq 2b + \sqrt{c_1^2 + c_2^2} \\
&\leq 2b + c_1 + c_2 \\
&= \sqrt{1-\mu_t\alpha}\cdot d_t + 2\mu_t\tau\left(\Psi(\mathcal{S}_1) + \Psi(\mathcal{S}_2)\right) + \eta_t
\end{aligned}$$

where the second and third inequalities hold because $\sqrt{a_1^2 + a_2^2} \leq a_1 + a_2$ for any non-negative numbers a_1 and a_2 and the last equality follows by substituting b, c_1, and c_2 with their definitions.

Proof of Theorem 4.4.1

The proof simply involves ensuring that the conditions of Lemma 4.4.2 are met so that its recursive application is possible. First, we focus on $T = 1$ and notice that, by Lemma C.1.6, we have $\|\boldsymbol{W}_0\|_2 \leq (1+\rho)\|\boldsymbol{W}^*\|_2$, and therefore the step-size condition of Lemma 4.4.2 is satisfied because

$$\mu \leq \frac{1}{32M_{\max}\,(1+\rho)^2\|\boldsymbol{X}^*\|_2} = \frac{1}{16M_{\max}\,(1+\rho)^2\|\boldsymbol{W}^*\|_2^2} \leq \frac{1}{16M_{\max}\|\boldsymbol{W}_0\|_2^2},$$

where the equality in here holds due to the fact that $\|\boldsymbol{W}^*\|_2^2 = 2\|\boldsymbol{X}^*\|_2$ for any $(\boldsymbol{U}^*, \boldsymbol{V}^*) \in \mathcal{X}^*$. Hence, by Lemma 4.4.2 we have

$$d_1 \leq \gamma\, d_0 + 2\mu\tau\left[\Psi_{\omega_1}(\mathcal{S}_1) + \Psi_{\omega_2}(\mathcal{S}_2)\right] + \eta \leq \rho\sqrt{\sigma_r(\boldsymbol{X}^*)},$$

where the second inequality utilizes the Theorem assumptions on the statistical error $\epsilon_{\text{stat}} = 2\mu\tau\left[\Psi_{\omega_1}(\mathcal{S}_1) + \Psi_{\omega_2}(\mathcal{S}_2)\right] + \eta$. Now, since we have shown that $d_1 \leq \rho\sqrt{\sigma_r(\boldsymbol{X}^*)}$, we can go ahead by applying Lemma 4.4.2 for $T = 2$. Since the step-size choice in Theorem statement satisfies the Lemma condition, we obtain

$$\begin{aligned} d_2 &\leq \gamma\, d_1 + \epsilon_{\text{stat}} \\ &\leq \gamma\,(\gamma\, d_0 + \epsilon_{\text{stat}}) + \epsilon_{\text{stat}} \\ &= \gamma^2\, d_0 + (\gamma + 1)\,\epsilon_{\text{stat}}. \end{aligned}$$

Recursively applying the same inequalities completes the proof.

Proof of Lemma C.3.1

The proof takes advantage of the following two Lemmata (both proved later in the current section), which separately deal with the terms dependent on gradients of $\mathcal{L}_n$ and g in the expression (C.6). The first Lemma, resembling Lemma C.1.4 of section C.1, will be helpful to show a sufficient descent condition with respect to the loss function $\mathcal{L}_n$.

Lemma C.3.3. *Let $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^T$ be an arbitrary rank-r matrix, where $\boldsymbol{U} \in \mathbb{R}^{d_1 \times r}$ and*

$\boldsymbol{V} \in \mathbb{R}^{d_2 \times r}$. *Then, under Assumption 1 we will have that*

$$\begin{aligned}\langle \nabla_U \mathcal{L}_n(\boldsymbol{X}), \boldsymbol{\Delta}_U \rangle + \langle \nabla_V \mathcal{L}_n(\boldsymbol{X}), \boldsymbol{\Delta}_V \rangle \geq \frac{mM}{2(m+M)} \|\boldsymbol{U}\boldsymbol{V}^T - \boldsymbol{U}^*\boldsymbol{V}^{*T}\|_F^2 \\ + \frac{1}{2(m+M)} \|\nabla \mathcal{L}_n(\boldsymbol{X}) - \nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_F^2 \\ - (m+M)\|\boldsymbol{\Delta}_W\|_F^4 - \frac{5r}{2m} \|\nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_2^2 .\end{aligned}$$

The second Lemma also states a sufficient descent condition, but with respect to the regularizer g.

Lemma C.3.4. *Let* $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^T$ *be an arbitrary rank-r matrix, where* $\boldsymbol{U} \in \mathbb{R}^{d_1 \times r}$ *and* $\boldsymbol{V} \in \mathbb{R}^{d_2 \times r}$. *Denoting* $\boldsymbol{Z} = \boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{V}^T\boldsymbol{V}$, *then under Assumption 2 we will have that*

$$\begin{aligned}\langle \nabla_U g(\boldsymbol{Z}), \boldsymbol{\Delta}_U \rangle + \langle \nabla_V g(\boldsymbol{Z}), \boldsymbol{\Delta}_V \rangle \geq \frac{m_g M_g}{m_g + M_g} \|\boldsymbol{Z}\|_F^2 + \frac{1}{2(m_g + M_g)} \|\nabla g(\boldsymbol{Z})\|_F^2 \\ - \frac{m_g + M_g}{2} \|\boldsymbol{\Delta}_W\|_F^4.\end{aligned}$$

Adding up the results of the above two Lemmata, after setting $\lambda = \frac{1}{8}$, yields

$$\begin{aligned}A := \langle \nabla_U \mathcal{L}_n(\boldsymbol{U}\boldsymbol{V}^T), \boldsymbol{\Delta}_U \rangle + \langle \nabla_V \mathcal{L}_n(\boldsymbol{U}\boldsymbol{V}^T), \boldsymbol{\Delta}_V \rangle & \quad \text{(C.16)} \\ + \lambda \langle \nabla_U g(\boldsymbol{Z}), \boldsymbol{\Delta}_U \rangle + \lambda \langle \nabla_V g(\boldsymbol{Z}), \boldsymbol{\Delta}_V \rangle & \\ \geq \frac{m}{4} \|\boldsymbol{U}\boldsymbol{V}^T - \boldsymbol{U}^*\boldsymbol{V}^{*T}\|_F^2 + \frac{m_g}{16} \|\boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{V}^T\boldsymbol{V}\|_F^2 & \quad \text{(C.17)} \\ - (2M + \frac{M_g}{8}) \|\boldsymbol{\Delta}_W\|_F^4 + \frac{1}{4M} \|\nabla \mathcal{L}_n(\boldsymbol{X}) - \nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_F^2 & \quad \text{(C.18)} \\ - \frac{5r}{2m} \|\nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_2^2 + \frac{1}{32\, M_g} \|\nabla g(\boldsymbol{Z})\|_F^2, & \quad \text{(C.19)}\end{aligned}$$

where we also took advantage of that $M_g \geq m_g$ and $M \geq m$ to slightly simplify the scalings. The following Lemma (proved later in this section) is then used to manage the second term in (C.17).

Lemma C.3.5. *For any* $(\boldsymbol{U}^*, \boldsymbol{V}^*) \in \mathcal{X}^*$ *and* $(\boldsymbol{U}, \boldsymbol{V}) \in \mathbb{R}^{d_1 \times r} \times \mathbb{R}^{d_2 \times r}$, *it holds that*

$$\|\boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{V}^T\boldsymbol{V}\|_F^2 \geq \|\boldsymbol{U}\boldsymbol{U}^T - \boldsymbol{U}^*\boldsymbol{U}^{*T}\|_F^2 + \|\boldsymbol{V}\boldsymbol{V}^T - \boldsymbol{V}^*\boldsymbol{V}^{*T}\|_F^2 - 2\|\boldsymbol{U}\boldsymbol{V}^T - \boldsymbol{U}^*\boldsymbol{V}^{*T}\|_F^2.$$

Using this Lemma, and defining $m_{\min} := \min\{m, m_g\}$ as the minimum of (restricted)

strong convexity constants, leads to

$$\begin{aligned}
&\frac{m}{4}\|\boldsymbol{U}\boldsymbol{V}^T - \boldsymbol{U}^*\boldsymbol{V}^{*T}\|_F^2 + \frac{m_g}{16}\|\boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{V}^T\boldsymbol{V}\|_F^2 \\
&\geq \frac{m_{\min}}{16}\left[\|\boldsymbol{U}\boldsymbol{U}^T - \boldsymbol{U}^*\boldsymbol{U}^{*T}\|_F^2 + \|\boldsymbol{V}\boldsymbol{V}^T - \boldsymbol{V}^*\boldsymbol{V}^{*T}\|_F^2 + 2\|\boldsymbol{U}\boldsymbol{V}^T - \boldsymbol{U}^*\boldsymbol{V}^{*T}\|_F^2\right] \\
&= \frac{m_{\min}}{16}\|\boldsymbol{W}\boldsymbol{W}^T - \boldsymbol{W}^*\boldsymbol{W}^{*T}\|_F^2 \\
&\geq \frac{m_{\min}}{16}\sigma_r(\boldsymbol{X}^*)\|\boldsymbol{\Delta}_W\|_F^2
\end{aligned}$$

where, to obtain the equality, we used the definitions of $\boldsymbol{W}$ and $\boldsymbol{W}^*$ to compactly express the lower-bound and then, to achieve the last inequality, utilized Lemma C.1.5 of the former appendix section (after setting $\boldsymbol{U} = \boldsymbol{W}$ and $\boldsymbol{U}^* = \boldsymbol{W}^*$ in there) to get

$$\|\boldsymbol{W}\boldsymbol{W}^T - \boldsymbol{W}^*\boldsymbol{W}^{*T}\|_F \geq \frac{\sigma_r(\boldsymbol{W}^*)}{\sqrt{2}} \cdot \operatorname{dist}(\boldsymbol{W}; \boldsymbol{W}^*) = \sqrt{\sigma_r(\boldsymbol{X}^*)} \cdot \operatorname{dist}(\boldsymbol{W}; \boldsymbol{W}^*),$$

where the equality follows from the equal-footedness assumption on $(\boldsymbol{U}^*, \boldsymbol{V}^*) \in \mathcal{X}^*$ which implies $\sigma_r(\boldsymbol{W}^*) = \sqrt{2\sigma_r(\boldsymbol{X}^*)}$. For the first term in (C.18), using the assumption $\|\boldsymbol{\Delta}_W\|_F \leq \rho\sqrt{\sigma_r(\boldsymbol{X}^*)}$, along with defining $M_{\max} := \max\{M, M_g\}$, will obtain that

$$(2M + \frac{M_g}{8})\|\boldsymbol{\Delta}_W\|_F^4 \leq \frac{17\, M_{\max}}{8}\rho^2\sigma_r(\boldsymbol{X}^*)\|\boldsymbol{\Delta}_W\|_F^2 \leq \frac{m_{\min}}{32}\sigma_r(\boldsymbol{X}^*)\|\boldsymbol{\Delta}_W\|_F^2,$$

where the last inequality holds due to the assumption on ρ, i.e. that $\rho^2 \leq m_{\min}/68\, M_{\max}$. Plugging all these into the expression of the lower bound for A in (C.16) will give us

$$\begin{aligned}
A \geq\ & \frac{m_{\min}}{32}\sigma_r(\boldsymbol{X}^*)\|\boldsymbol{\Delta}_W\|_F^2 + \frac{1}{4M}\,\|\nabla\mathcal{L}_n(\boldsymbol{X}) - \nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_F^2 \\
&+ \frac{1}{32\, M_g}\|\nabla g(\boldsymbol{Z})\|_F^2 - \frac{5r}{2m}\,\|\nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_2^2\,,
\end{aligned}$$

which completes the proof.

Proof of Lemma C.3.2

First, we show how the first term in (C.7) could be bounded:

$$\begin{aligned}
\|\nabla_U \mathcal{L}_n(\boldsymbol{U}\boldsymbol{V}^T) + \lambda \nabla_U g(\boldsymbol{Z})\|_F^2 &= \|\nabla \mathcal{L}_n(\boldsymbol{U}\boldsymbol{V}^T)\boldsymbol{V} + 2\lambda \boldsymbol{U}\nabla g(\boldsymbol{Z})\|_F^2 \\
&\leq 2\|\nabla \mathcal{L}_n(\boldsymbol{X})\boldsymbol{V}\|_F^2 + 8\,\lambda^2 \|\boldsymbol{U}\nabla g(\boldsymbol{Z})\|_F^2 \\
&= 2\,\|\nabla \mathcal{L}_n(\boldsymbol{X})\boldsymbol{V} - \nabla \mathcal{L}_n(\boldsymbol{X}^*)\boldsymbol{V} + \nabla \mathcal{L}_n(\boldsymbol{X}^*)\boldsymbol{V}\|_F^2 \\
&+ 8\,\lambda^2 \|\boldsymbol{U}\nabla g(\boldsymbol{Z})\|_F^2 \\
&\leq 4\,\|\nabla \mathcal{L}_n(\boldsymbol{X})\boldsymbol{V} - \nabla \mathcal{L}_n(\boldsymbol{X}^*)\boldsymbol{V}\|_F^2 + 4\,\|\nabla \mathcal{L}_n(\boldsymbol{X}^*)\boldsymbol{V}\|_F^2 \\
&+ 8\,\lambda^2 \|\boldsymbol{U}\nabla g(\boldsymbol{Z})\|_F^2 \\
&\leq 4\|\boldsymbol{V}\|_2^2\,\|\nabla \mathcal{L}_n(\boldsymbol{X}) - \nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_F^2 \\
&+ 4\|\boldsymbol{V}\|_F^2 \|\nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_2^2 + 8\,\lambda^2 \|\boldsymbol{U}\|_2^2\,\|\nabla g(\boldsymbol{Z})\|_F^2,
\end{aligned}$$

where to derive the first two inequalities we have leveraged $\|\boldsymbol{A}+\boldsymbol{B}\|_F^2 \leq 2\|\boldsymbol{A}\|_F^2 + 2\|\boldsymbol{B}\|_F^2$, which holds for any arbitrary same-sized matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ and the last inequality is true since $\|\boldsymbol{A}\boldsymbol{B}\|_F \leq \|\boldsymbol{A}\|_2 \cdot \|\boldsymbol{B}\|_F$. By applying the same steps to the second term in (C.7) we obtain

$$\begin{aligned}
\|\nabla_V \mathcal{L}_n(\boldsymbol{U}\boldsymbol{V}^T) + \lambda \nabla_V g(\boldsymbol{Z})\|_F^2 &\leq 4\|\boldsymbol{U}\|_2^2\,\|\nabla \mathcal{L}_n(\boldsymbol{X}) - \nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_F^2 \\
&+ 4\|\boldsymbol{U}\|_F^2 \|\nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_2^2 + 8\,\lambda^2 \|\boldsymbol{V}\|_2^2\,\|\nabla g(\boldsymbol{Z})\|_F^2.
\end{aligned}$$

Adding up the last two upper bounds and using the fact that $\|\boldsymbol{U}\|_2, \|\boldsymbol{V}\|_2 \leq \|\boldsymbol{W}\|_2$ would then complete the proof:

$$\begin{aligned}
\|\nabla_U \mathcal{L}_n(\boldsymbol{U}\boldsymbol{V}^T) + \lambda \nabla_U g(\boldsymbol{Z})\|_F^2 + \|\nabla_V \mathcal{L}_n(\boldsymbol{U}\boldsymbol{V}^T) + \lambda \nabla_V g(\boldsymbol{Z})\|_F^2 \\
&\leq 4\left(\|\boldsymbol{U}\|_2^2 + \|\boldsymbol{V}\|_2^2\right)\|\nabla \mathcal{L}_n(\boldsymbol{X}) - \nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_F^2 \\
&+ 4\left(\|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2\right)\|\nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_2^2 \\
&+ 8\,\lambda^2\left(\|\boldsymbol{U}\|_2^2 + \|\boldsymbol{V}\|_2^2\right)\|\nabla g(\boldsymbol{Z})\|_F^2 \\
&\leq 8\|\boldsymbol{W}\|_2^2 \cdot \|\nabla \mathcal{L}_n(\boldsymbol{X}) - \nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_F^2 \\
&+ 4\|\boldsymbol{W}\|_F^2 \cdot \|\nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_2^2 + 16\,\lambda^2 \|\boldsymbol{W}\|_2^2 \cdot \|\nabla g(\boldsymbol{Z})\|_F^2 \\
&\leq 4\|\boldsymbol{W}\|_2^2 \cdot \Big(2\,\|\nabla \mathcal{L}_n(\boldsymbol{X}) - \nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_F^2 \\
&+ r\|\nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_2^2 + 4\lambda^2 \|\nabla g(\boldsymbol{Z})\|_F^2\Big),
\end{aligned}$$

where the last inequality is valid since $\|\boldsymbol{W}\|_F \leq \sqrt{r}\|\boldsymbol{W}\|_2$, which in turn leverages the rank-r assumption on $\boldsymbol{W}$.

Proof of Lemma C.3.3

We proceed by merely adapting the proof of Lemma C.1.4 to the general non-PSD case. First, we note that

$$\begin{aligned}
&\langle \nabla_U \mathcal{L}_n(\boldsymbol{U}\boldsymbol{V}^T), \boldsymbol{\Delta}_U \rangle + \langle \nabla_V \mathcal{L}_n(\boldsymbol{U}\boldsymbol{V}^T), \boldsymbol{\Delta}_V \rangle \\
&\qquad = \langle \nabla \mathcal{L}_n(\boldsymbol{U}\boldsymbol{V}^T)\boldsymbol{V}, \boldsymbol{\Delta}_U \rangle + \langle \nabla \mathcal{L}_n(\boldsymbol{U}\boldsymbol{V}^T)^T\boldsymbol{U}, \boldsymbol{\Delta}_V \rangle \qquad \text{(C.20)} \\
&\qquad = \langle \nabla \mathcal{L}_n(\boldsymbol{U}\boldsymbol{V}^T), \boldsymbol{\Delta}_U\boldsymbol{V}^T \rangle + \langle \nabla \mathcal{L}_n(\boldsymbol{U}\boldsymbol{V}^T), \boldsymbol{U}\boldsymbol{\Delta}_V^T \rangle \\
&\qquad = \langle \nabla \mathcal{L}_n(\boldsymbol{U}\boldsymbol{V}^T), \boldsymbol{U}\boldsymbol{V}^T - \boldsymbol{U}^*\boldsymbol{V}^{*T} + \boldsymbol{\Delta}_U\boldsymbol{\Delta}_V^T \rangle \qquad \text{(C.21)} \\
&\qquad = \langle \nabla \mathcal{L}_n(\boldsymbol{X}), \boldsymbol{X} - \boldsymbol{X}^* + \boldsymbol{\Delta}_U\boldsymbol{\Delta}_V^T \rangle \qquad \text{(C.22)} \\
&\qquad = \langle \nabla \mathcal{L}_n(\boldsymbol{X}) - \nabla \mathcal{L}_n(\boldsymbol{X}^*), \boldsymbol{X} - \boldsymbol{X}^* \rangle \qquad \text{(C.23)} \\
&\qquad + \langle \nabla \mathcal{L}_n(\boldsymbol{X}) - \nabla \mathcal{L}_n(\boldsymbol{X}^*), \boldsymbol{\Delta}_U\boldsymbol{\Delta}_V^T \rangle \\
&\qquad + \langle \nabla \mathcal{L}_n(\boldsymbol{X}^*), \boldsymbol{X} - \boldsymbol{X}^* \rangle + \langle \nabla \mathcal{L}_n(\boldsymbol{X}^*), \boldsymbol{\Delta}_U\boldsymbol{\Delta}_V^T \rangle
\end{aligned}$$

where $\boldsymbol{\Delta}_U := \boldsymbol{U} - \boldsymbol{U}^*\boldsymbol{R}^*$, $\boldsymbol{\Delta}_V := \boldsymbol{V} - \boldsymbol{V}^*\boldsymbol{R}^*$, (C.20) follows by the definitions of the partial gradients $\nabla_U \mathcal{L}_n(\boldsymbol{U}\boldsymbol{V}^T)$ and $\nabla_V \mathcal{L}_n(\boldsymbol{U}\boldsymbol{V}^T)$, the equality in (C.21) holds since it can be easily shown that

$$\boldsymbol{\Delta}_U\boldsymbol{V}^T + \boldsymbol{U}\boldsymbol{\Delta}_V^T = \boldsymbol{U}\boldsymbol{V}^T - \boldsymbol{U}^*\boldsymbol{V}^{*T} + \boldsymbol{\Delta}_U\boldsymbol{\Delta}_V^T,$$

and (C.22) is obtained by replacing $\boldsymbol{U}\boldsymbol{V}^T$ and $\boldsymbol{U}^*\boldsymbol{V}^{*T}$ with $\boldsymbol{X}$ and $\boldsymbol{X}^*$, respectively. For the first term of the summation beginning on (C.23), we take advantage of the fact that $\mathcal{L}_n$ meets RSC and RSM conditions to yield (via utilizing Theorem 2.1.11 in [14] and adapting it to the set of rank-r matrices)

$$\begin{aligned}
\langle \nabla \mathcal{L}_n(\boldsymbol{X}) - \nabla \mathcal{L}_n(\boldsymbol{X}^*), \boldsymbol{X} - \boldsymbol{X}^* \rangle \geq \frac{mM}{m+M}\|\boldsymbol{X} - \boldsymbol{X}^*\|_F^2 \\
+ \frac{1}{m+M}\|\nabla \mathcal{L}_n(\boldsymbol{X}) - \nabla \mathcal{L}_n(\boldsymbol{X}^*)\|_F^2.
\end{aligned}$$

For the second term, we can apply the following chain of inequalities

$$\begin{aligned}
\langle \nabla\mathcal{L}_n(\boldsymbol{X}) - \nabla\mathcal{L}_n(\boldsymbol{X}^*), \boldsymbol{\Delta}_U\boldsymbol{\Delta}_V^T\rangle &\geq -|\langle \nabla\mathcal{L}_n(\boldsymbol{X}) - \nabla\mathcal{L}_n(\boldsymbol{X}^*), \boldsymbol{\Delta}_U\boldsymbol{\Delta}_V^T\rangle| \\
&= -|\langle (\nabla\mathcal{L}_n(\boldsymbol{X}) - \nabla\mathcal{L}_n(\boldsymbol{X}^*))\boldsymbol{\Delta}_V, \boldsymbol{\Delta}_U\rangle| \\
&\overset{(i)}{\geq} -\|(\nabla\mathcal{L}_n(\boldsymbol{X}) - \nabla\mathcal{L}_n(\boldsymbol{X}^*))\boldsymbol{\Delta}_V\|_F \cdot \|\boldsymbol{\Delta}_U\|_F \\
&\overset{(ii)}{\geq} -\|\nabla\mathcal{L}_n(\boldsymbol{X}) - \nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_2 \cdot \|\boldsymbol{\Delta}_V\|_F \cdot \|\boldsymbol{\Delta}_U\|_F \\
&\overset{(iii)}{\geq} -\|\nabla\mathcal{L}_n(\boldsymbol{X}) - \nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_F \cdot \|\boldsymbol{\Delta}_W\|_F^2 \\
&\overset{(iv)}{\geq} -\frac{1}{2(m+M)}\|\nabla\mathcal{L}_n(\boldsymbol{X}) - \nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_F^2 \\
&\quad - \frac{m+M}{2}\|\boldsymbol{\Delta}_W\|_F^4,
\end{aligned}$$

where (i) is by Cauchy-Schwartz inequality, (ii) follows from that by the definition of matrix spectral norm we can say

$$\|(\nabla\mathcal{L}_n(\boldsymbol{X}) - \nabla\mathcal{L}_n(\boldsymbol{X}^*))\,\boldsymbol{\Delta}_V\|_F \leq \|\nabla\mathcal{L}_n(\boldsymbol{X}) - \nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_2 \cdot \|\boldsymbol{\Delta}_V\|_F,$$

(iii) leverages the fact that for $\boldsymbol{\Delta}_W = \begin{bmatrix} \boldsymbol{\Delta}_U \\ \boldsymbol{\Delta}_V \end{bmatrix}$ we have $\|\boldsymbol{\Delta}_W\|_F^2 = \|\boldsymbol{\Delta}_U\|_F^2 + \|\boldsymbol{\Delta}_V\|_F^2$, and (iv) holds because $ab \leq \frac{\beta}{2}a^2 + \frac{1}{2\beta}b^2$ is true for any $\beta > 0$, which upon setting $a = \|\nabla\mathcal{L}_n(\boldsymbol{X}) - \nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_F$, $b = \|\boldsymbol{\Delta}_W\|_F^2$, and $\beta = \frac{1}{m+M}$ implies the inequality. Therefore, we have so far shown that

$$\begin{aligned}
\left\langle \nabla\mathcal{L}_n(\boldsymbol{X}) - \nabla\mathcal{L}_n(\boldsymbol{X}^*), \boldsymbol{X} - \boldsymbol{X}^* + \boldsymbol{\Delta}_U\boldsymbol{\Delta}_V^T\right\rangle &\geq \frac{mM}{m+M}\,\|\boldsymbol{X} - \boldsymbol{X}^*\|_F^2 - \frac{m+M}{2}\|\boldsymbol{\Delta}_W\|_F^4 \\
&\quad + \frac{1}{2(m+M)}\,\|\nabla\mathcal{L}_n(\boldsymbol{X}) - \nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_F^2\,.
\end{aligned}$$

We approach the third term of the upper-bound in (C.23) as follows:

$$\begin{aligned}
\langle \nabla\mathcal{L}_n(\boldsymbol{X}^*), \boldsymbol{X} - \boldsymbol{X}^*\rangle &\geq -|\langle \nabla\mathcal{L}_n(\boldsymbol{X}^*), \boldsymbol{X} - \boldsymbol{X}^*\rangle| \\
&\geq -\|\boldsymbol{X} - \boldsymbol{X}^*\|_* \cdot \|\nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_2 \\
&\geq -\sqrt{2r}\,\|\boldsymbol{X} - \boldsymbol{X}^*\|_F \cdot \|\nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_2 \\
&\geq -\frac{mM}{2(m+M)}\,\|\boldsymbol{X} - \boldsymbol{X}^*\|_F^2 - \frac{(m+M)r}{Mm}\,\|\nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_2^2
\end{aligned}$$

where $\|\cdot\|_*$ denotes matrix nuclear norm, the third inequality leverages the facts that $\boldsymbol{X}-\boldsymbol{X}^*$ is of rank at most $2r$ as well as that $\|\boldsymbol{M}\|_* \leq \sqrt{s}\,\|\boldsymbol{M}\|_F$ holds for any arbitrary rank-s matrix $\boldsymbol{M}$, and the last inequality is by using the inequality $ab \leq \frac{\beta}{2}a^2 + \frac{1}{2\beta}b^2$ with $a = \|\boldsymbol{X}-\boldsymbol{X}^*\|_F$, $b = \|\nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_2$, and $\beta = \frac{mM}{2(m+M)r}$.

Finally, similar steps as the ones used for the second term of the upper-bound in (C.23) can be applied to the last term as follows

$$\begin{aligned}
\langle \nabla\mathcal{L}_n(\boldsymbol{X}^*), \boldsymbol{\Delta}_U\boldsymbol{\Delta}_V^T\rangle &\geq -|\langle \nabla\mathcal{L}_n(\boldsymbol{X}^*), \boldsymbol{\Delta}_U\boldsymbol{\Delta}_V^T\rangle| \\
&\geq -\|\nabla\mathcal{L}_n(\boldsymbol{X}^*)\boldsymbol{\Delta}_V\|_F \cdot \|\boldsymbol{\Delta}_U\|_F \\
&\geq -\|\nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_2 \cdot \|\boldsymbol{\Delta}_U\|_F \cdot \|\boldsymbol{\Delta}_V\|_F \\
&\geq -\|\nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_2 \cdot \|\boldsymbol{\Delta}_W\|_F^2 \\
&\geq -\frac{1}{2(m+M)}\|\nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_2^2 - \frac{m+M}{2}\|\boldsymbol{\Delta}_W\|_F^4.
\end{aligned}$$

where the last inequality is by again applying $ab \leq \frac{\beta}{2}a^2 + \frac{1}{2\beta}b^2$ with $a = \|\boldsymbol{\Delta}_W\|_F^2$, $b = \|\nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_2$, and $\beta = m+M$. Therefore, we have

$$\begin{aligned}
\langle \nabla_U\mathcal{L}_n(\boldsymbol{U}\boldsymbol{V}^T), \boldsymbol{\Delta}_U\rangle + \langle \nabla_V\mathcal{L}_n(\boldsymbol{U}\boldsymbol{V}^T), \boldsymbol{\Delta}_V\rangle \geq{}& \frac{mM}{2(m+M)}\|\boldsymbol{X}-\boldsymbol{X}^*\|_F^2 - (m+M)\|\boldsymbol{\Delta}_W\|_F^4 \\
&+ \frac{1}{2(m+M)}\|\nabla\mathcal{L}_n(\boldsymbol{X}) - \nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_F^2 \\
&- \left[\frac{1}{2(m+M)} + \frac{(m+M)r}{mM}\right]\|\nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_F^2.
\end{aligned}$$

Noticing that $\frac{1}{2(m+M)} + \frac{(m+M)r}{mM} \leq \frac{5r}{2m}$, which holds since $m \leq M$ and $r \geq 1$, will obtain the claim of the Lemma.

Proof of Lemma C.3.4

Given that

$$\nabla_U g(\boldsymbol{Z}) = 2\boldsymbol{U}\,\nabla g(\boldsymbol{Z}) \quad \text{and} \quad \nabla_V g(\boldsymbol{Z}) = -2\boldsymbol{V}\,\nabla g(\boldsymbol{Z}),$$

where $\boldsymbol{Z} = \boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{V}^T\boldsymbol{V}$, the proof follows the steps of a similar analysis in [18], which is brought here for the sake of completeness. First, notice that

$$\begin{aligned}
\langle \nabla_U g(\boldsymbol{Z}), \boldsymbol{\Delta}_U \rangle + \langle \nabla_V g(\boldsymbol{Z}), \boldsymbol{\Delta}_V \rangle &= 2\langle \boldsymbol{U}\nabla g(\boldsymbol{Z}), \boldsymbol{U} - \boldsymbol{U}^*\boldsymbol{R}^* \rangle - 2\langle \boldsymbol{V}\nabla g(\boldsymbol{Z}), \boldsymbol{V} - \boldsymbol{V}^*\boldsymbol{R}^* \rangle \\
&= 2\langle \nabla g(\boldsymbol{Z}), \boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{U}^T\boldsymbol{U}^*\boldsymbol{R}^* - \boldsymbol{V}^T\boldsymbol{V} + \boldsymbol{V}^T\boldsymbol{V}^*\boldsymbol{R}^* \rangle \\
&\overset{(i)}{=} 2\langle \nabla g(\boldsymbol{Z}), \boldsymbol{Y}^T(\boldsymbol{W} - \boldsymbol{W}^*\boldsymbol{R}^*) \rangle \\
&\overset{(ii)}{=} \langle \nabla g(\boldsymbol{Z}), \boldsymbol{Y}^T\boldsymbol{W} \rangle \qquad \text{(C.24)} \\
&\quad + \langle \nabla g(\boldsymbol{Z}), \boldsymbol{Y}^T\boldsymbol{W} - 2\boldsymbol{Y}^T\boldsymbol{W}^*\boldsymbol{R}^* + \boldsymbol{R}^*\boldsymbol{Y}^{*T}\boldsymbol{W}^*\boldsymbol{R}^* \rangle \\
&\overset{(iii)}{=} \langle \nabla g(\boldsymbol{Z}), \boldsymbol{Y}^T\boldsymbol{W} \rangle + \langle \nabla g(\boldsymbol{Z}), \boldsymbol{\Delta}_Y^T\boldsymbol{\Delta}_W \rangle, \qquad \text{(C.25)}
\end{aligned}$$

where (i) is by the definitions of $\boldsymbol{Y}$, $\boldsymbol{W}$, and $\boldsymbol{W}^*$ as introduced in the beginning of the current appendix sectio, (ii) is by that

$$\boldsymbol{Y}^{*T}\boldsymbol{W}^* = \boldsymbol{U}^{*T}\boldsymbol{U}^* - \boldsymbol{V}^{*T}\boldsymbol{V}^* = \boldsymbol{0},$$

which holds since $(\boldsymbol{U}^*, \boldsymbol{V}^*) \in \mathcal{X}^*$, and (iii) utilizes the definitions of $\boldsymbol{\Delta}_Y$ and $\boldsymbol{\Delta}_W$. Using the assumption that g is M_g-smooth and m_g-strongly convex enables us to lower bound the first term of the above summation as follows

$$\begin{aligned}
\langle \nabla g(\boldsymbol{Z}), \boldsymbol{Y}^T\boldsymbol{W} \rangle &= \langle \nabla g(\boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{V}^T\boldsymbol{V}), \boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{V}^T\boldsymbol{V} \rangle \\
&= \langle \nabla g(\boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{V}^T\boldsymbol{V}) - \nabla g(\boldsymbol{0}), \boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{V}^T\boldsymbol{V} - \boldsymbol{0} \rangle \\
&\geq \frac{m_g M_g}{m_g + M_g} \|\boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{V}^T\boldsymbol{V}\|_F^2 + \frac{1}{m_g + M_g} \|\nabla g(\boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{V}^T\boldsymbol{V})\|_F^2
\end{aligned}$$

where the second equality leverages the assumption that $\nabla g(\boldsymbol{0}) = \boldsymbol{0}$ and the inequality is via utilizing Theorem 2.1.11 in [14]. For the second term in the right-hand side expression of (C.25) we have

$$\begin{aligned}
\langle \nabla g(\boldsymbol{Z}), \boldsymbol{\Delta}_Y^T\boldsymbol{\Delta}_W \rangle &\geq -|\langle \nabla g(\boldsymbol{Z}), \boldsymbol{\Delta}_Y^T\boldsymbol{\Delta}_W \rangle| \\
&= -|\langle \boldsymbol{\Delta}_Y \nabla g(\boldsymbol{Z}), \boldsymbol{\Delta}_W \rangle| \\
&\overset{(i)}{\geq} -\|\boldsymbol{\Delta}_W\|_F \cdot \|\boldsymbol{\Delta}_Y \nabla g(\boldsymbol{Z})\|_F \\
&\geq -\|\boldsymbol{\Delta}_W\|_F \cdot \|\boldsymbol{\Delta}_Y\|_F \cdot \|\nabla g(\boldsymbol{Z})\|_2 \\
&\overset{(ii)}{\geq} -\|\boldsymbol{\Delta}_W\|_F^2 \cdot \|\nabla g(\boldsymbol{Z})\|_F \\
&\overset{(iii)}{\geq} -\frac{m_g + M_g}{2} \|\boldsymbol{\Delta}_W\|_F^4 - \frac{1}{2(m_g + M_g)} \|\nabla g(\boldsymbol{Z})\|_F^2
\end{aligned}$$

where (i) is by Cauchy-Schwartz inequality, (ii) uses $\|\boldsymbol{\Delta}_W\|_F = \|\boldsymbol{\Delta}_Y\|_F$ and that $\|\nabla g(\boldsymbol{Z})\|_F \geq \|\nabla g(\boldsymbol{Z})\|_2$, and (iii) holds since $ab \leq \frac{\beta}{2}a^2 + \frac{1}{2\beta}b^2$ for any $\beta > 0$, which after setting $a = \|\nabla g(\boldsymbol{Z})\|_F$, $b = \|\boldsymbol{\Delta}_W\|_F^2$, and $\beta = \frac{1}{m_g + M_g}$ implies the inequality. The Lemma will then follow by summing up the last two inequalities that were derived.

Proof of Lemma C.3.5

The proof is taken from [18] and is included here for the sake of completeness. It proceeds as follows

$$\begin{aligned}
\|\boldsymbol{U}^T\boldsymbol{U} - \boldsymbol{V}^T\boldsymbol{V}\|_F^2 &= \|\boldsymbol{U}^T\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}^T\boldsymbol{V}\|_F^2 - 2\langle \boldsymbol{U}^T\boldsymbol{U}, \boldsymbol{V}^T\boldsymbol{V}\rangle \\
&= \|\boldsymbol{U}\boldsymbol{U}^T\|_F^2 + \|\boldsymbol{V}\boldsymbol{V}^T\|_F^2 - 2\langle \boldsymbol{U}\boldsymbol{V}^T, \boldsymbol{U}\boldsymbol{V}^T\rangle \\
&= \langle \boldsymbol{W}\boldsymbol{W}^T, \boldsymbol{Y}\boldsymbol{Y}^T\rangle \\
&= \langle \boldsymbol{W}\boldsymbol{W}^T - \boldsymbol{W}^*\boldsymbol{W}^{*T}, \boldsymbol{Y}\boldsymbol{Y}^T - \boldsymbol{Y}^*\boldsymbol{Y}^{*T}\rangle \\
&+ \langle \boldsymbol{W}\boldsymbol{W}^T, \boldsymbol{Y}^*\boldsymbol{Y}^{*T}\rangle + \langle \boldsymbol{W}^*\boldsymbol{W}^{*T}, \boldsymbol{Y}\boldsymbol{Y}^T - \boldsymbol{Y}^*\boldsymbol{Y}^{*T}\rangle \\
&\overset{(i)}{=} \langle \boldsymbol{W}\boldsymbol{W}^T - \boldsymbol{W}^*\boldsymbol{W}^{*T}, \boldsymbol{Y}\boldsymbol{Y}^T - \boldsymbol{Y}^*\boldsymbol{Y}^{*T}\rangle \\
&+ \langle \boldsymbol{W}\boldsymbol{W}^T, \boldsymbol{Y}^*\boldsymbol{Y}^{*T}\rangle + \langle \boldsymbol{W}^*\boldsymbol{W}^{*T}, \boldsymbol{Y}\boldsymbol{Y}^T\rangle \\
&= \langle \boldsymbol{W}\boldsymbol{W}^T - \boldsymbol{W}^*\boldsymbol{W}^{*T}, \boldsymbol{Y}\boldsymbol{Y}^T - \boldsymbol{Y}^*\boldsymbol{Y}^{*T}\rangle + \|\boldsymbol{Y}^{*T}\boldsymbol{W}\|_F^2 + \|\boldsymbol{Y}^T\boldsymbol{W}^*\|_F^2 \\
&\geq \langle \boldsymbol{W}\boldsymbol{W}^T - \boldsymbol{W}^*\boldsymbol{W}^{*T}, \boldsymbol{Y}\boldsymbol{Y}^T - \boldsymbol{Y}^*\boldsymbol{Y}^{*T}\rangle \\
&= \|\boldsymbol{U}\boldsymbol{U}^T - \boldsymbol{U}^*\boldsymbol{U}^{*T}\|_F^2 + \|\boldsymbol{V}\boldsymbol{V}^T - \boldsymbol{V}^*\boldsymbol{V}^{*T}\|_F^2 - 2\|\boldsymbol{U}\boldsymbol{V}^T - \boldsymbol{U}^*\boldsymbol{V}^{*T}\|_F^2
\end{aligned}$$

where (i) is by that since $(\boldsymbol{U}^*, \boldsymbol{V}^*)$ is an equally-footed pair of factors we have

$$\langle \boldsymbol{W}^*\boldsymbol{W}^{*T}, \boldsymbol{Y}^*\boldsymbol{Y}^{*T}\rangle = \|\boldsymbol{Y}^{*T}\boldsymbol{W}^*\|_F^2 = \|\boldsymbol{U}^{*T}\boldsymbol{U}^* - \boldsymbol{V}^{*T}\boldsymbol{V}^*\|_F^2 = 0.$$

C.4 Proof of Theorem Implications

Proof of Lemma 4.5.1

Defining $\xi(\boldsymbol{\Sigma}) := \sup_{\|\boldsymbol{u}\|_2=1, \|\boldsymbol{v}\|_2=1} \mathrm{var}(\boldsymbol{u}^T\boldsymbol{A}\boldsymbol{v})$, where $\mathrm{vec}(\boldsymbol{A}) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, the Lemma is an immediate implication of Lemma 7 in [123], which is stated in the below:

Lemma C.4.1. *Assume the measurement matrices $\boldsymbol{A}_i \in \mathbb{R}^{d_1 \times d_2}$, for $1 \leq i \leq n$, are i.i.d. samples from the $\boldsymbol{\Sigma}$-ensemble Gaussian distribution. Then, there exist universal*

positive constants (c_0, c_1) *such that for any* $\boldsymbol{X} \in \mathbb{R}^{d_1 \times d_2}$ *we have*

$$\frac{\|\mathcal{A}(\boldsymbol{X})\|_2^2}{2n} \geq \frac{\lambda_{\min}(\boldsymbol{\Sigma})}{4} \|\boldsymbol{X}\|_F^2 - c_1 \xi(\boldsymbol{\Sigma}) \left(\frac{d_1 + d_2}{n}\right) \|\boldsymbol{X}\|_*^2, \quad \textit{and}$$
$$\frac{\|\mathcal{A}(\boldsymbol{X})\|_2^2}{2n} \leq \lambda_{\max}(\boldsymbol{\Sigma}) \|\boldsymbol{X}\|_F^2 + c_1 \xi(\boldsymbol{\Sigma}) \left(\frac{d_1 + d_2}{n}\right) \|\boldsymbol{X}\|_*^2,$$

where $\|\boldsymbol{X}\|_*$ *denotes the nuclear norm of* $\boldsymbol{X}$*, with probability at least* $1 - \exp(-c_0\, n)$*.*

For the least squares loss function, $\mathcal{L}_n(\boldsymbol{X}) = \frac{1}{2n}\|\mathcal{A}(\boldsymbol{X}) - \boldsymbol{y}\|_2^2$, and two arbitrary rank-r matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, if we define the function $\tau(\boldsymbol{X}_2, \boldsymbol{X}_1) := \mathcal{L}_n(\boldsymbol{X}_2) - \mathcal{L}_n(\boldsymbol{X}_1) - \langle \nabla \mathcal{L}_n(\boldsymbol{X}_1), \boldsymbol{X}_2 - \boldsymbol{X}_1 \rangle$, then it holds that

$$\tau(\boldsymbol{X}_2, \boldsymbol{X}_1) = \frac{\|\mathcal{A}(\boldsymbol{X}_2 - \boldsymbol{X}_1)\|_2^2}{2n}.$$

The first inequality of the above Lemma would then imply

$$\begin{aligned}\tau(\boldsymbol{X}_2, \boldsymbol{X}_1) &\geq \frac{\lambda_{\min}(\boldsymbol{\Sigma})}{4} \|\boldsymbol{X}_2 - \boldsymbol{X}_1\|_F^2 - c_1 \xi(\boldsymbol{\Sigma}) \left(\frac{d_1 + d_2}{n}\right) \|\boldsymbol{X}_2 - \boldsymbol{X}_1\|_*^2 \\ &\geq \left[\frac{\lambda_{\min}(\boldsymbol{\Sigma})}{4} - 2c_1 r\, \xi(\boldsymbol{\Sigma}) \left(\frac{d_1 + d_2}{n}\right)\right] \|\boldsymbol{X}_2 - \boldsymbol{X}_1\|_F^2,\end{aligned}$$

where the second inequality leverages $\|\boldsymbol{X}_1 - \boldsymbol{X}_2\|_* \leq \sqrt{2r}\, \|\boldsymbol{X}_1 - \boldsymbol{X}_2\|_F$, which is true by the fact that $\boldsymbol{X}_1 - \boldsymbol{X}_2$ is of rank at most $2r$. Then, by requiring

$$n \geq \frac{16 c_1\, \xi(\boldsymbol{\Sigma})}{\lambda_{\min}(\boldsymbol{\Sigma})} (r d_1 + r d_2), \tag{C.26}$$

we can ensure $\tau(\boldsymbol{X}_2, \boldsymbol{X}_1) \geq \frac{\lambda_{\min}(\boldsymbol{\Sigma})}{8} \|\boldsymbol{X}_1 - \boldsymbol{X}_2\|_F^2$, therefore the RSC condition holds with $m = \lambda_{\min}(\boldsymbol{\Sigma})/4$. Similarly, the second inequality of the above Lemma, along with the stated relation between nuclear and Frobenius norms of $\boldsymbol{X}_2 - \boldsymbol{X}_1$, could be leveraged to imply

$$\tau(\boldsymbol{X}_2, \boldsymbol{X}_1) \leq \left[\lambda_{\max}(\boldsymbol{\Sigma}) + 2c_1 r\, \xi(\boldsymbol{\Sigma}) \left(\frac{d_1 + d_2}{n}\right)\right] \|\boldsymbol{X}_2 - \boldsymbol{X}_1\|_F^2.$$

Upon imposing that

$$n \geq \frac{2 c_1\, \xi(\boldsymbol{\Sigma})}{\lambda_{\max}(\boldsymbol{\Sigma})} (r d_1 + r d_2),$$

we are guaranteed that $\tau(\boldsymbol{X}_2, \boldsymbol{X}_1) \leq 2\lambda_{\max}(\boldsymbol{\Sigma}) \|\boldsymbol{X}_2 - \boldsymbol{X}_1\|_F^2$, which in turn ensures the RSM assumption with $M = 4\lambda_{\max}(\boldsymbol{\Sigma})$. Enforcing the condition in (C.26) will verufy both requirements on n, since $\lambda_{\max}(\boldsymbol{\Sigma}) \geq \lambda_{\min}(\boldsymbol{\Sigma})$.

Proof of Lemma 4.5.3

Notice that, upon defining the loss function $\mathcal{L}_n(\boldsymbol{X}) := \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{\Sigma}_n\|_F^2$, we have that $\nabla\mathcal{L}_n(\boldsymbol{X}^*) = \boldsymbol{X}^* - \boldsymbol{\Sigma}_n$. Moreover, let $\boldsymbol{D} \in \mathbb{R}^{n\times d}$ denote the data matrix whose rows contain the (transposed) data vectors $\boldsymbol{x}_i \in \mathbb{R}^d$, for $i = 1, 2, \cdots, n$. It is easy to show that, given the spiked model of the population covariance matrix $\boldsymbol{\Sigma}$, the data matrix can be decomposed as follows

$$\boldsymbol{D} = \boldsymbol{D}_{lc} + \boldsymbol{N} = \boldsymbol{P}\sqrt{\boldsymbol{\Lambda}^*}\,\boldsymbol{Q}^{*T} + \boldsymbol{N},$$

where $\boldsymbol{D}_{lc}$ corresponds to the r leading components, while $\boldsymbol{N}$ is associated with the remaining weaker ones and has i.i.d. $\mathcal{N}(0, \nu)$ entries. In the second equality of the above expression, $\boldsymbol{Q}^*$ and $\boldsymbol{\Lambda}^*$ are the eigenvectors and eigenvalues of $\boldsymbol{X}^*$, respectively, and $\boldsymbol{P}$ is the $n \times r$ random effects matrix with i.i.d. $\mathcal{N}(0, 1)$ entries (see [94] for similar representations).

Now, since the sample covariance matrix can be expressed as $\boldsymbol{\Sigma}_n = \frac{1}{n}\boldsymbol{D}^T\boldsymbol{D}$, the above representation of data matrix leads to the following

$$\begin{aligned}\boldsymbol{\Sigma}_n &= \frac{1}{n}\boldsymbol{D}_{lc}^T\boldsymbol{D}_{lc} + \frac{1}{n}\boldsymbol{N}^T\boldsymbol{N} \\ &= \boldsymbol{Q}^*\left(\frac{1}{n}\sqrt{\boldsymbol{\Lambda}^*}\boldsymbol{P}^T\boldsymbol{P}\sqrt{\boldsymbol{\Lambda}^*}\right)\boldsymbol{Q}^{*T} + \frac{1}{n}\boldsymbol{N}^T\boldsymbol{N}.\end{aligned}$$

Getting back to the proof, note that characterizing the statistical error of the proximal descent algorithm amounts to bounding $\|\nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_2 = \|\boldsymbol{\Sigma}_n - \boldsymbol{X}^*\|_2$, which can be further upper-bounded via the triangle inequality as follows

$$\begin{aligned}\|\nabla\mathcal{L}_n(\boldsymbol{X}^*)\|_2 &= \|\boldsymbol{X}^* - \boldsymbol{\Sigma}_n\|_2 \\ &= \left\|\left(\boldsymbol{X}^* - \frac{1}{n}\boldsymbol{D}_{lc}^T\boldsymbol{D}_{lc}\right) - \frac{1}{n}\boldsymbol{N}^T\boldsymbol{N}\right\|_2 \\ &= \left\|\boldsymbol{Q}^*\left(\boldsymbol{\Lambda}^* - \frac{1}{n}\sqrt{\boldsymbol{\Lambda}^*}\boldsymbol{P}^T\boldsymbol{P}\sqrt{\boldsymbol{\Lambda}^*}\right)\boldsymbol{Q}^{*T} - \frac{1}{n}\boldsymbol{N}^T\boldsymbol{N}\right\|_2 \\ &\leq \left\|\boldsymbol{Q}^*\left(\boldsymbol{\Lambda}^* - \frac{1}{n}\sqrt{\boldsymbol{\Lambda}^*}\boldsymbol{P}^T\boldsymbol{P}\sqrt{\boldsymbol{\Lambda}^*}\right)\boldsymbol{Q}^{*T}\right\|_2 + \frac{1}{n}\left\|\boldsymbol{N}^T\boldsymbol{N}\right\|_2. \qquad \text{(C.27)}\end{aligned}$$

Letting $\boldsymbol{T} := \boldsymbol{\Lambda}^* - \frac{1}{n}\sqrt{\boldsymbol{\Lambda}^*}\boldsymbol{P}^T\boldsymbol{P}\sqrt{\boldsymbol{\Lambda}^*}$, by the joint k-sparsity assumption on $\boldsymbol{Q}^*$ we have

$$\|\boldsymbol{Q}^*\boldsymbol{T}\boldsymbol{Q}^{*T}\|_2 = \sup_{\boldsymbol{u}:\|\boldsymbol{u}\|_2\leq 1}\boldsymbol{u}^T\boldsymbol{Q}^*\boldsymbol{T}\boldsymbol{Q}^{*T}\boldsymbol{u} = \sup_{\boldsymbol{u}:\,\|\boldsymbol{u}\|_2\leq 1,\,\|\boldsymbol{u}\|_0\leq k}\boldsymbol{u}^T\boldsymbol{Q}^*\boldsymbol{T}\boldsymbol{Q}^{*T}\boldsymbol{u}.$$

Then, the following fact (implied by Lemma 15 of [149]) is helpful towards our proof

$$\Pr\left(\sup_{\boldsymbol{u}:\,\|\boldsymbol{u}\|_2\leq 1,\,\|\boldsymbol{u}\|_0\leq k}\left|\frac{1}{n}\|\boldsymbol{D}_{lc}\,\boldsymbol{u}\|_2^2-\boldsymbol{u}^T\boldsymbol{X}^*\boldsymbol{u}\right|\geq t\right)$$
$$\leq 2\exp\left(-cn\min\left\{\frac{t^2}{\sigma_1^2(\boldsymbol{X}^*)},\frac{t}{\sigma_1(\boldsymbol{X}^*)}\right\}+k\log d\right),$$

where $c>0$ is a universal constant. Applying this inequality with

$$t=c'\sigma_1(\boldsymbol{X}^*)\max\left\{\frac{k\log d}{n},\sqrt{\frac{k\log d}{n}}\right\},$$

where $c'=\frac{8}{c}$, then yields that $\|\boldsymbol{Q}^*\boldsymbol{T}\boldsymbol{Q}^{*T}\|_2\leq t$, with probability at least $1-2d^{-4}$. For the second term in (C.27), we make use of the following result, which is stated as Corollary 5.35 in [150]:

Corollary C.4.1. *Let $\boldsymbol{A}$ be an $n\times d$ matrix whose entries are i.i.d. standard normal variables. Then for every $t\geq 0$, with probability at least $1-2\exp(-t^2/2)$ one has*

$$\|\boldsymbol{A}\|_2\leq\sqrt{n}+\sqrt{d}+t.$$

Upon setting $t=\sqrt{d}$ and accounting for the fact that the entries of $\boldsymbol{N}$ have variance σ^2, we obtain

$$\frac{1}{\sqrt{n}}\|\boldsymbol{N}\|_2\leq\nu\left(1+2\sqrt{\frac{d}{n}}\right),$$

with probability at least $1-2\exp(-d/2)$. Assuming $n\geq d$, the above inequality implies $\frac{1}{n}\|\boldsymbol{N}\boldsymbol{N}^T\|_2\leq\nu^2\left(1+3\sqrt{\frac{d}{n}}\right)$.

Proof of Correctness of Algorithm 9

First, we show that if $0<\tau<\|\boldsymbol{U}\|_{2,1}$ then the answer to

$$\widetilde{\boldsymbol{U}}=\underset{\check{\boldsymbol{U}}\in\mathbb{R}^{d\times r}}{\operatorname{argmin}}\frac{1}{2}\|\boldsymbol{U}-\check{\boldsymbol{U}}\|_F^2+\tau\|\check{\boldsymbol{U}}\|_{2,\infty}\tag{C.28}$$

is non-zero. Defining $f(\check{\boldsymbol{U}}):=\frac{1}{2}\|\boldsymbol{U}-\check{\boldsymbol{U}}\|_F^2+\tau\|\check{\boldsymbol{U}}\|_{2,\infty}$ we notice that $f(\boldsymbol{0})=\frac{1}{2}\|\boldsymbol{U}\|_F^2$. Furthermore, let us define

$$\boldsymbol{U}^*:=\underset{\check{\boldsymbol{U}}\in\mathbb{R}^{d\times r}:\|\check{\boldsymbol{U}}\|_{2,\infty}\leq 1}{\operatorname{argmax}}\langle\boldsymbol{U},\check{\boldsymbol{U}}\rangle=\underset{\check{\boldsymbol{U}}\in\mathbb{R}^{d\times r}}{\operatorname{argmax}}\quad\langle\boldsymbol{U},\frac{\check{\boldsymbol{U}}}{\|\check{\boldsymbol{U}}\|_{2,\infty}}\rangle.$$

Then, for a positive scalar $t > 0$, we can expand $f(t\boldsymbol{U}^*)$ simply as follows

$$\begin{aligned} f(t\,\boldsymbol{U}^*) &= \frac{1}{2}\|\boldsymbol{U} - t\boldsymbol{U}^*\|_F^2 + \tau t\|\boldsymbol{U}^*\|_{2,\infty} \\ &\overset{(i)}{=} \frac{1}{2}\|\boldsymbol{U}\|_F^2 + \frac{t^2}{2}\|\boldsymbol{U}^*\|_F^2 - t\langle \boldsymbol{U}, \boldsymbol{U}^*\rangle + \tau t \\ &\overset{(ii)}{=} f(\boldsymbol{0}) + \frac{t^2}{2}\|\boldsymbol{U}^*\|_F^2 + t\left(\tau - \langle \boldsymbol{U}, \boldsymbol{U}^*\rangle\right) \\ &\overset{(iii)}{=} f(\boldsymbol{0}) + \frac{t^2}{2}\|\boldsymbol{U}^*\|_F^2 + t\left(\tau - \|\boldsymbol{U}\|_{2,1}\right), \end{aligned}$$

where (i) leverages the fact that $\|\boldsymbol{U}^*\|_{2,\infty} = 1$, (ii) holds because $f(\boldsymbol{0}) = \frac{1}{2}\|\boldsymbol{U}\|_F^2$, and finally (iii) is implied by the duality of the $\ell_{2,1}$ and $\ell_{2,\infty}$ norms. Since $\tau - \|\boldsymbol{U}\|_{2,1} < 0$, by choosing the positive scalar t small enough we can then ensure that $f(t\boldsymbol{U}^*) < f(\boldsymbol{0})$.

We continue by casting the problem in (C.28) as a constrained optimization

$$\begin{aligned} \min_{\check{\boldsymbol{U}} \in \mathbb{R}^{d\times r},\, t\in\mathbb{R}} \quad & \frac{1}{2}\|\boldsymbol{U} - \check{\boldsymbol{U}}\|_F^2 + \tau t, \\ \text{subject to} \quad & \|\check{\boldsymbol{U}}_{i*}\|_2 \le t, \;\; \text{for } 1 \le i \le d. \end{aligned} \tag{C.29}$$

The augmented Lagrangian [151] for the above constrained problem has the form of

$$L(\check{\boldsymbol{U}}, t, \boldsymbol{\varrho}) = \frac{1}{2}\|\boldsymbol{U} - \check{\boldsymbol{U}}\|_F^2 + \tau t + \sum_{i=1}^{d} \rho_i(\|\check{\boldsymbol{U}}_{i*}\|_2 - t),$$

where $\boldsymbol{\varrho} \in \mathbb{R}^d$ denotes the vector of Lagrange variables corresponding to the constraints of problem (C.29). Writing the KKT conditions [151] for the optimal primal-dual variables of this problem, which we denote by $(\widetilde{\boldsymbol{U}}, \widetilde{t}, \widetilde{\boldsymbol{\varrho}})$, leads to the following conditions

1. $\widetilde{\boldsymbol{U}}_{i*} = \left(1 - \frac{\widetilde{\rho}_i}{\|\boldsymbol{U}_{i*}\|_2}\right)_+ \boldsymbol{U}_{i*}$ for $i = 1, 2, \cdots, d$ [1]
2. $\sum_{i=1}^{d} \widetilde{\rho}_i = \tau$
3. $\widetilde{\rho}_i \ge 0$, for $i = 1, 2, \cdots, d$

[1] Notice that the first-order optimality condition $\nabla_{\boldsymbol{U}} L(\widetilde{\boldsymbol{U}}, \widetilde{t}, \widetilde{\boldsymbol{\varrho}}) = \boldsymbol{0}$ implies that for $i = 1, 2, \cdots, d$

$$\begin{cases} \widetilde{\boldsymbol{U}}_{i*} = \boldsymbol{0}, \;\; \text{if } \|\boldsymbol{U}_{i*}\|_2 \le \widetilde{\rho}_i \\ \left(1 + \frac{\widetilde{\rho}_i}{\|\widetilde{\boldsymbol{U}}_{i*}\|_2}\right)\widetilde{\boldsymbol{U}}_{i*} = \boldsymbol{U}_{i*}, \;\; \text{otherwise.} \end{cases}$$

Attempting to write both cases in the above together, while noticing that in the second case, i.e. when $\|\boldsymbol{U}_{i*}\|_2 > \widetilde{\rho}_i$, the gradient condition implies $\frac{\widetilde{\boldsymbol{U}}_{i*}}{\|\widetilde{\boldsymbol{U}}_{i*}\|_2} = \frac{\boldsymbol{U}_{i*}}{\|\boldsymbol{U}_{i*}\|_2}$, leads to the listed condition.

4. $\widetilde{t} \geq \|\widetilde{\boldsymbol{U}}_{i*}\|_2$ for $i = 1, 2, \cdots, d$

5. If $\widetilde{\rho}_i > 0$ for some $1 \leq i \leq d$, then $\|\widetilde{\boldsymbol{U}}_{i*}\|_2 = \widetilde{t}$.

6. If $\widetilde{t} > \|\widetilde{\boldsymbol{U}}_{i*}\|_2$ for some $1 \leq i \leq d$, then $\widetilde{\rho}_i = 0$.

Now, assume $\widetilde{\rho}_i > 0$ for some $1 \leq i \leq d$, then condition (5) in the above implies that $\|\widetilde{\boldsymbol{U}}_{i*}\|_2 = \widetilde{t}$. So, by leveraging condition (1) we obtain the following expression for $\widetilde{t}$

$$\widetilde{t} = \left(1 - \frac{\widetilde{\rho}_i}{\|\boldsymbol{U}_{i*}\|_2}\right)_+ \|\boldsymbol{U}_{i*}\|_2 = \left(\|\boldsymbol{U}_{i*}\|_2 - \widetilde{\rho}_i\right)_+ . \tag{C.30}$$

Since, as shown in the beginning of the proof, we have $\widetilde{\boldsymbol{U}} \neq \boldsymbol{0}$ when $0 < \tau < \|\boldsymbol{U}\|_{2,1}$, therefore, by condition (4) we have that $\widetilde{t} > 0$, and so the above equality in (C.30) leads to $\widetilde{t} = \|\boldsymbol{U}_{i*}\|_2 - \widetilde{\rho}_i$, which in turn implies $\widetilde{\rho}_i = \|\boldsymbol{U}_{i*}\|_2 - \widetilde{t}$. This means that either $\widetilde{\rho}_i = 0$ or $\widetilde{\rho}_i = \|\boldsymbol{U}_{i*}\|_2 - \widetilde{t}$. Both cases can be compactly expressed as $\widetilde{\rho}_i = \left(\|\boldsymbol{U}_{i*}\|_2 - \widetilde{t}\right)_+$. The second condition, along with this expression for entries of $\widetilde{\boldsymbol{\varrho}}$, leads to the following

$$\sum_{i=1}^{d} \left(\|\boldsymbol{U}_{i*}\|_2 - \widetilde{t}\right)_+ = \tau,$$

which can be solved via bisection, over the interval of $[0, \|\boldsymbol{U}\|_{2,\infty}]$, to find the optimal value $\widetilde{t}$. Having found $\widetilde{t}$, the second step is to compute the components of $\boldsymbol{\varrho}$ via $\widetilde{\rho}_i = \left(\|\boldsymbol{U}_{i*}\|_2 - \widetilde{t}\right)_+$. Finally, the rows of $\widetilde{\boldsymbol{U}}$ can be found through condition (1) as follows

$$\widetilde{\boldsymbol{U}}_{i*} = \left(1 - \frac{\widetilde{\rho}_i}{\|\boldsymbol{U}_{i*}\|_2}\right)_+ \boldsymbol{U}_{i*}.$$