ON THE LOCAL CORRECTNESS OF ℓ_1-MINIMIZATION FOR DICTIONARY LEARNING ALGORITHM

BY

QUAN GENG

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2011

Urbana, Illinois

Adviser:

Professor Pramod Viswanath

ABSTRACT

The idea that many classes of signals can be represented by linear combination of a small set of atoms of a dictionary has had a great impact on various signal processing applications, e.g., image compression, super resolution imaging and robust face recognition. For practical problems such a sparsifying dictionary is usually unknown ahead of time, and many heuristics have been proposed to learn an efficient dictionary from the given data. However, there is little theory explaining the empirical success of these heuristic methods. In this work, we prove that under mild conditions, the dictionary learning problem is actually locally well-posed: the desired solution is a local optimum of the ℓ_1-norm minimization problem. More precisely, let $\mathbf{A} \in \mathbb{R}^{m \times n}$ be an incoherent (and possibly overcomplete, i.e., $m < n$) dictionary, the coefficients $\mathbf{X} \in \mathbb{R}^{n \times p}$ follow a random sparse model, and $\mathbf{Y} = \mathbf{AX}$ be the observed data; then with high probability $(\mathbf{A}, \mathbf{X})$ is a local optimum of the ℓ_1-minimization problem:

$$\underset{\mathbf{A}', \mathbf{X}'}{\text{minimize}} \ \|\mathbf{X}'\|_1 \quad \text{s.t.} \quad \mathbf{Y} = \mathbf{A}'\mathbf{X}', \ \|\mathbf{A}'_i\|_2 = 1 \ \forall\, i,$$

provided the number of samples $p = \Omega(n^3 k)$. This is the first result showing that the dictionary learning problem is locally solvable for an overcomplete dictionary. Our analysis draws on tools developed for the matrix completion problem. In particular, inspired by David Gross's golfing scheme, we derive relaxed optimality conditions and construct dual variables to certify the local optimality conditions.

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION 1
1.1 Background 1
1.2 Organization 3
1.3 Notation 4

CHAPTER 2 MODEL AND MAIN RESULT 5
2.1 Model 5
2.2 Simulation 7
2.3 Main Result 8

CHAPTER 3 LOCAL PROPERTY AND LINEARIZED SUB-PROBLEM 10
3.1 Local Property 10
3.2 Optimality Conditions of Linearized Subproblem 13

CHAPTER 4 CERTIFICATION PROCESS 16
4.1 Markov Process Construction 18
4.2 Analysis 19

CHAPTER 5 BALANCEDNESS PROPERTY 26

CHAPTER 6 CONCLUSION 35

APPENDIX A MATRIX NORM BOUND BY MUTUAL CO-HERENCE 37

APPENDIX B MATRIX CHERNOFF BOUND AND KAHANE-KHINTCHINE INEQUALITY 39

APPENDIX C A DECOUPLING LEMMA 42

APPENDIX D PROOF OF LEMMA 5.2 45

APPENDIX E PROOF OF LEMMA 5.3 48
E.1 Proof of Lemma 5.3 48
E.2 Proof of Lemma E.2 52
E.3 Proof of Lemma E.3 53

REFERENCES 58

CHAPTER 1

INTRODUCTION

1.1 Background

In the past few decades, much progress in signal processing has been driven by the goal of finding effective signal representations. Several well known and powerful bases have been developed to represent audio and image signals, for example, DFT, DCT and wavelet, which help discover the structures of many classes of signals and lead to various useful applications, including the successful practical image coding standards [1]. Hand design of bases to effectively represent signals has been a paradigm in the signal processing field.

However, there is a different idea which suggests that instead of designing bases for each class of signals we encounter, we may learn an effective signal representation from large sets of data. This idea can be illustrated by the well known principal component analysis (PCA) and Karhunen-Loève decomposition. In PCA, by doing eigenvalue decomposition of the data covariance matrix, we can find the most important components in the data, which help to understand the underlying data structures. The idea of learning bases from the data is very appealing, compared to manually designing bases for each class of signals we may encounter. On the one hand, it is inefficient and even not possible to design bases for each class of signals. On the other hand, as the dimension of data goes higher and higher, it may be beyond our human intuition and capability to design effective representations for high dimensional data.

In the past ten years, motivated by this promise, researchers have devoted great effort to developing an automatic procedure to find effective signal representation from sample data. In particular, much effort has been focused on the sparse linear representation. A signal $\mathbf{y} \in \mathbb{R}^m$ has a sparse representation in terms of the dictionary of $\mathbf{A} = [\mathbf{A}_1, \ldots, \mathbf{A}_n] \in \mathbb{R}^{m \times n}$ if $\mathbf{y} = \mathbf{A}\mathbf{x}$, for some sparse coefficient vector $\mathbf{x} \in \mathbb{R}^n$. Sparse linear representation has been a dominant idea in the field of signal processing recently. This is because in many practical problems, signals have the property of sparsity or near sparsity [2]. In addition, due to the fundamental theoretical results in compressed sensing [3], it is well known how to efficiently represent and recover sparse signals. One basic result in compressed sensing can be understood as: If a signal $\mathbf{y} \in \mathbb{R}^m$ has a sparse representation under the dictionary $\mathbf{A}$, where $\mathbf{A}$ satisfies restricted isometry property (RIP) [3], then we can find the sparse coefficients exactly by solving an ℓ_1-minimization problem:

$$\text{minimize} \quad \|\mathbf{x}\|_1 \quad \text{s.t.} \quad \mathbf{y} = \mathbf{A}\mathbf{x}. \tag{1.1}$$

Compressed sensing theory gives us a powerful tool to study sparse signal representations: once a signal has a sparse representation under an known dictionary $\mathbf{A}$, there exists an efficient algorithm which can guarantee to recover it.

However, when one is given a new class of signals, it may not be clear under what bases the signals can be sparsely represented. A popular heuristic is to find a dictionary to represent these signals as sparsely as possible. Specifically, we are studying the following model problem, which is referred to as "dictionary learning":

> Given samples $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_p] \in \mathbb{R}^{m \times p}$ all of which can be sparsely represented in terms of some unknown dictionary $\mathbf{A}$ (i.e., $\mathbf{Y} = \mathbf{A}\mathbf{X}$, for some $\mathbf{X}$ with sparse columns), recover $\mathbf{A}$ and $\mathbf{X}$.

Many heuristic algorithms have been proposed to solve this problem [2],[4]. Although they demonstrated great empirical performance in various applications, their success is largely a mystery and there is little theory available to explain when and why they succeed.

In this work we take a step towards such a theory. In particular, we study the following model approach to dictionary learning:

$$\underset{\mathbf{A},\mathbf{X}}{\text{minimize}} \ \|\mathbf{X}\|_1 \quad \text{subject to} \quad \mathbf{Y} = \mathbf{A}\mathbf{X}, \ \|\mathbf{A}_i\|_2 = 1 \, \forall \, i, \tag{1.2}$$

where $\|\mathbf{X}\|_1 = \sum_{ij} |X_{ij}|$. This model, as a natural abstraction of many heuristic algorithms, was first proposed in the work of Gribonval and Schnass

[5]. They show that if the dictionary $\mathbf{A}$ is a square matrix and the coefficient $\mathbf{X}$ follows Gaussian-Bernoulli random model, $(\mathbf{A}, \mathbf{X})$ is a local optimum with high probability, provided the sample number $p = O(n \log n)$. In our work, we do not restrict the dictionary to be square and it can be overcomplete, i.e., $\mathbf{A}$ can have more columns than rows.

Before we state our main result in Chapter 2, let us discuss some properties of the above optimization problem. First, notice that although the objective function is convex, both of the constraints are not. Therefore it is a nonconvex problem. Second, the problem has an exponential number of optimal solutions. Indeed, let $\mathbf{\Pi} \in \mathbb{R}^{n \times n}$ be a permutation matrix and $\mathbf{\Sigma} \in \mathbb{R}^{n \times n}$ be a diagonal matrix of signs. Suppose $(\mathbf{A}, \mathbf{X})$ is an optimal solution; then $(\mathbf{A\Pi\Sigma}, \mathbf{\Sigma\Pi}^*\mathbf{X})$ also solves the optimization problem. Because of this "permutation and sign" ambiguity, the problem has at least $2^n n!$ solutions. The nonconvexity and the exponential number of equivalent solutions make it very difficult to solve and analyze the problem, and it seems nothing rigorous can be said about the problem.

However, we will show that with high probability this problem is solvable, at least locally, if the dictionary $\mathbf{A}$ satisfies some incoherence property, the coefficients $\mathbf{X}$ follow a random sparse model and the number of samples is $\Omega(n^3 k)$. Intriguingly, simulation results (see Chapter 2) even suggest global correct recovery: no matter which initial point one chooses, a local algorithm always converges to the desired solution (of course, up to the "permutation and sign" ambiguity), if the problem is well-structured.

1.2 Organization

The goal of our work is to show why the dictionary learning problem is locally solvable. In Chapter 2, we describe the model in detail and state our main result. The rest of the thesis is dedicated to the proof of the result. In Chapter 3, we show that the local correctness of the ℓ_1-minimization (1.2) is equivalent to the global optimality over the tangent space of the constraint of (1.2) and derive relaxed KKT optimality conditions. Chapter 4 proves the optimality conditions indeed hold with high probability by constructing the dual variables explicitly using a Markov process. The success of the construction relies on a certain balancedness property of the linearized subproblem at the optimal point, which is stated and proved in Chapter 5. Chapter 6 concludes the thesis and outlines several directions for future work.

1.3 Notation

For matrices, $\mathbf{X}^*$ denotes the transpose of $\mathbf{X}$, $\|\mathbf{X}\|$ denotes the matrix operator norm and $\|\mathbf{X}\|_F = \sqrt{\mathrm{tr}[\mathbf{X}^*\mathbf{X}]}$ denotes the Frobenius norm. By slight abuse of notation, $\|\mathbf{X}\|_1$ and $\|\mathbf{X}\|_\infty$ will denote the ℓ_1 and ℓ_∞ norms of the matrix, viewed as a large vector:

$$\|\mathbf{X}\|_1 = \sum_{ij} |\mathbf{X}_{ij}|, \qquad \|\mathbf{X}\|_\infty = \max_{ij} |X_{ij}|. \tag{1.3}$$

For vector $\mathbf{x}$, $\|\mathbf{x}\|$ denotes the ℓ_2-norm of $\mathbf{x}$. $\|\mathbf{x}\|_1$ and $\|\mathbf{x}\|_\infty$ will denote the usual ℓ_1-norm and ℓ_∞-norm, respectively. The symbols $\mathbf{e}_1, \dots, \mathbf{e}_d$ will denote the standard basis vectors for $\mathbb{R}^d$. Throughout, the symbols $C_1, C_2, \dots,$ $c_1, c_2, \dots$ refer to numerical constants. When used in different sections, they need not refer to the same constant. For a linear subspace $V \subset \mathbb{R}^d$, $\mathbf{P}_V \in \mathbb{R}^{d\times d}$ denotes the projection matrix onto V. For matrices, $\mathbf{A} \otimes \mathbf{B}$ denotes the Kronecker product between matrices $\mathbf{A}$ and $\mathbf{B}$.

CHAPTER 2

MODEL AND MAIN RESULT

2.1 Model

As introduced in Chapter 1, our goal is to study under what conditions the dictionary learning problem can be solved, at least locally, via ℓ_1-minimization. The behavior of the algorithm obviously depends on the property of $\mathbf{A}$ and the sparse coefficient matrix $\mathbf{X}$. In this paper, we consider the dictionary $\mathbf{A}$ whose column has a unit ℓ_2-norm and take the simple assumption that the columns of $\mathbf{A}$ are well-spread, i.e., the mutual coherence [3] of $\mathbf{A}$

$$\mu(\mathbf{A}) = \max_{i \neq j} |\langle \mathbf{A}_i, \mathbf{A}_j \rangle| \tag{2.1}$$

is small. Classical results [2] show that if a known dictionary $\mathbf{A}$ has low coherence, then we can recover any sparse representation with sparsity size up to $1/2\mu(\mathbf{A})$ by solving a ℓ_1-minimization problem:

$$\|\mathbf{x}_0\|_0 < \tfrac{1}{2}(1 + 1/\mu(\mathbf{A})) \tag{2.2}$$

$$\Longrightarrow \quad \mathbf{x}_0 = \arg\min \|\mathbf{x}\|_1 \quad \text{s.t.} \quad \mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{x}_0. \tag{2.3}$$

Although pessimistic compared to surprisingly good empirical performance, this result is powerful since the assumption on low coherence of $\mathbf{A}$ is very simple and reasonable. On the other hand, there is less prior to believe a dictionary satisfies the more powerful restricted isometry property [3].

Next we model the sparse coefficients of $\mathbf{X}$. As we described before, we want to show that if $\mathbf{X}$ is sufficiently sparse, then $\mathbf{A}$ can be locally recovered by solving an ℓ_1-minimization problem. However, sparsity itself is not enough to guarantee the success of recovery. Suppose for some i, $\mathbf{X}_{ij} = 0$ for all j, then by no means can one recover the ith column of $\mathbf{A}$. Therefore, in this work we assume $\mathbf{X}$ not only is sparse but also has a random sparsity pattern.

Formally, we assume the sparsity pattern of $\mathbf{X}$ is random and the nonzero entries are i.i.d. Gaussian.

We will use the Gaussian-random-sparsity model for the coefficient matrix $\mathbf{X}$. We assume $x_j \in \mathbb{R}^n$ is generated in the following way: first choose k out of its n entries uniformly, and then set each of these k entries to be i.i.d. Gaussian, while setting the remaining $n-k$ entries to be zero. More precisely, $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_p] \in \mathbb{R}^{m\times p}$ are generated i.i.d by $\mathbf{y}_j = \mathbf{A}\mathbf{x}_j$, where $\mathbf{x}_j \in \mathbb{R}^n$ satisfies a Gaussian-random-sparsity model:

$$\Omega_j \;\sim\; \mathrm{uni}\binom{[n]}{k} \tag{2.4}$$

and

$$\mathbf{x}_j = \mathbf{P}_{\Omega_j}\mathbf{v}_j, \tag{2.5}$$

where

$$v_{ij} \;\sim_{i.i.d.}\; \mathcal{N}\left(0, \sigma^2\right), \qquad \sigma = \sqrt{n/kp},^{1} \tag{2.6}$$

and $\mathbf{P}_{\Omega_j}$ is the projection matrix onto Ω_j.

In dictionary learning problem, what we observe is neither $\mathbf{A}$ nor $\mathbf{X}$, but rather their product $\mathbf{Y} = \mathbf{A}\mathbf{X}$. Given $\mathbf{Y}$, there is a corresponding constraint manifold which is

$$\mathcal{M} = \{(\mathbf{A}, \mathbf{X}) \mid \mathbf{A}\mathbf{X} = \mathbf{Y},\ \|\mathbf{A}_i\|_2 = 1\ \forall i\} \subset \mathbb{R}^{m\times n} \times \mathbb{R}^{n\times p}. \tag{2.7}$$

Inspired by the success of ℓ_1-minimization to recover sparse signals in compressed sensing, the following approach based on ℓ_1-minimization over the constraint manifold has been proposed by Gribonval and Schnass [5] to solve the dictionary learning problem:

$$\underset{\mathbf{A},\mathbf{X}}{\text{minimize}}\ \|\mathbf{X}\|_1 \quad \text{subject to} \quad \mathbf{Y} = \mathbf{A}\mathbf{X},\ \|\mathbf{A}_i\|_2 = 1\ \forall i, \tag{2.8}$$

where $\|\mathbf{X}\|_1 = \sum_{ij} |X_{ij}|$.

This model is a natural abstraction of many heuristic algorithms, and [5] shows that if the dictionary $\mathbf{A}$ is a square matrix and the coefficient $\mathbf{X}$ follows Gaussian-Bernoulli random model, $(\mathbf{A}, \mathbf{X})$ is a local optimum with

[1] Here, σ can be any positive number, and it will not affect the correctness of our theory. The reason we choose $\sigma = \sqrt{n/kp}$ is for notational convenience: when p is large, then the spectrum norm of $\|\mathbf{X}\|$ will be approximately one.

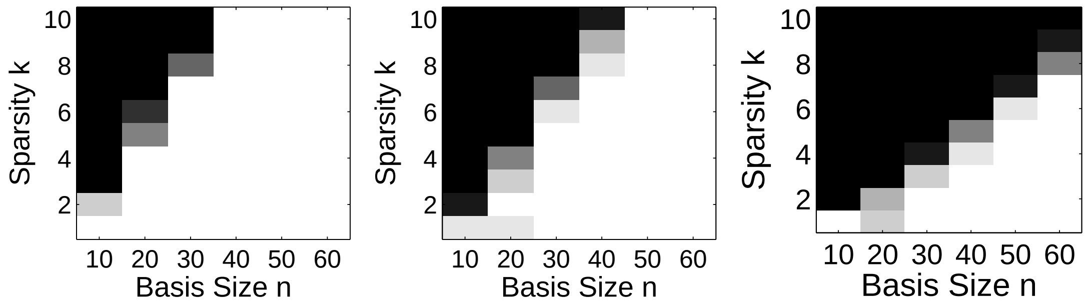

Figure 2.1: **Phase transitions in dictionary recovery?** We synthesize sample data with varying sparsity level and problem size to test whether locally minimizing the ℓ_1-norm correctly recovers the dictionary. Specifically, the dictionary $\mathbf{A} \in \mathbb{R}^{m \times n}$, sparse coefficients $\mathbf{X} \in \mathbb{R}^{n \times p}$, where $m = n \times \delta$ and $p = 5n \log n$. Left: $\delta = 1$. Middle: $\delta = 0.8$. Right: $\delta = 0.6$. If the relative error $\|\hat{\mathbf{A}} - \mathbf{A}\|_F / \|\mathbf{A}\|_F$ in the recovered $\hat{\mathbf{A}}$ is smaller than 10^{-5}, the trial is judged success. We average over 10 trials; white corresponds to success in all trials, black to failure in all trials. The problems are solved using an algorithm based on sequential convex optimization and augmented Lagrangian multiplier method [6].

high probability, provided the sample number $p = O(n \log n)$. In our work, we do not restrict the dictionary to be square and it can be overcomplete, i.e., $\mathbf{A}$ can have more columns than rows.

2.2 Simulation

Before formally stating our main theoretical result, we show some very intriguing simulation results in Figure 2.1. The simulation results indicate that if $\mathbf{A}$ is incoherent, $\mathbf{X}$ follows the above Gaussian that under the above Gaussian-random-sparsity model and $p = \Omega(n \log n)$, the dictionary learning problem is indeed solvable, not only locally.

In Figure 2.1, we synthesize sample data with varying sparsity level and problem size to test whether locally minimizing the ℓ_1-norm correctly recovers the dictionary. For each problem instance, we start with a random initial point in the constraint manifold and then use sequential convex optimization to find a local optimal point of (2.8). Empirically, when the problem is well-structured ($\mathbf{X}$ is sufficiently sparse), the local algorithm always correctly recovers the dictionary $\mathbf{A}$ up to sign and permutation ambiguity from any random initial points; otherwise, when $\mathbf{X}$ is not sufficiently sparse, the algorithm fails. Further, the phase transition in Figure 2.1 is fairly sharp.

The simulation results show that there exist important classes of dictionary learning problems which are solvable, and when the problem is well structured even global recovery is achievable.

2.3 Main Result

Although the empirical results are very intriguing, the aforementioned difficulties of nonconvexity and sign-permutation ambiguity make it very difficult to analyze the correctness of the ℓ_1-minimization approach for dictionary learning. However, Gribonval and Schnass [5] made the first step towards developing a theory for the correctness of dictionary learning algorithm. They show that if the dictionary $\mathbf{A}$ is a square matrix and the coefficient $\mathbf{X}$ follows Gaussian-Bernoulli random model, $(\mathbf{A}, \mathbf{X})$ is a local optimum of (2.8) with high probability, provided the sample number $p = O(n \log n)$. In our work, we show that the same result can be extended to wider classes of dictionary, including the overcomplete dictionary, which has more columns than rows.

Our main result is that provided the number of samples $p = \Omega(n^3 k)$, with high probability the desired solution $(\mathbf{A}, \mathbf{X})$ is a local optimum of the ℓ_1-minimization problem over the manifold $\mathcal{M}$ defined in 2.7. More precisely,

Theorem 2.1. *There exist numerical constants $C_1, C_2, C_3 > 0$, such that if $\mathbf{x} = (\mathbf{A}, \mathbf{X})$ satisfy the probability model (2.4)-(2.6) with*

$$k \;\leq\; \min\{C_1/\mu(\mathbf{A}), C_2 n\}, \tag{2.9}$$

then $\mathbf{x}$ is a local minimum of the ℓ_1-norm over $\mathcal{M}$, with probability at least

$$1 - C_3 \|\mathbf{A}\|^2 n^{3/2} k^{1/2} p^{-1/2} (\log p). \tag{2.10}$$

This result implies when given polynomially many samples, the dictionary learning problem is actually locally well-posed and a local algorithm can hope to recover the desired dictionary. Our result on the local correctness of ℓ_1-minimization for dictionary learning is the first result suggesting correct recovery is possible via ℓ_1-minimization for non-square matrices.

The rest of the thesis is devoted to the proof of Theorem 2.1. In Chapter 3 we will show that the local correctness of the ℓ_1-minimization (2.8) is equivalent to the global optimality over the tangent space of $\mathcal{M}$, and derive relaxed KKT optimality conditions. Chapter 4 proves the optimality conditions indeed hold with high probability by constructing the dual variables explicitly using a Markov process. The success of the construction relies on a certain balancedness property of the linearized subproblem at the optimal point, which is formally stated and proved in Chapter 5.

CHAPTER 3

LOCAL PROPERTY AND LINEARIZED SUBPROBLEM

3.1 Local Property

As described in the previous chapter, a key role in studying the problem (2.8) will be played by the tangent space of manifold $\mathcal{M}$. The tangent space at the point $(\mathbf{A}, \mathbf{X})$ is characterized by

$$\{(\boldsymbol{\Delta}_A, \boldsymbol{\Delta}_X) \mid \boldsymbol{\Delta}_A \mathbf{X} + \mathbf{A} \boldsymbol{\Delta}_X = \mathbf{0}, \langle \mathbf{A}_i, \boldsymbol{\Delta}_{Ai} \rangle = 0 \; \forall i\}, \tag{3.1}$$

where the first equation comes from differentiating the bilinear constraint $\mathbf{Y} = \mathbf{A}\mathbf{X}$, while the second comes from differentiating the constraint $\|\mathbf{A}_i\|^2 = 1$.

The minimization of the ℓ_1-norm of $\mathbf{X}$ over the above tangent space is the linearized subproblem:

$$\begin{aligned} \underset{\Delta_A, \Delta_X}{\text{minimize}} \quad & \|\mathbf{X} + \boldsymbol{\Delta}_X\|_1 \\ \text{s.t.} \quad & \boldsymbol{\Delta}_A \mathbf{X} + \mathbf{A} \boldsymbol{\Delta}_X = \mathbf{0}, \\ & \langle \mathbf{A}_i, \boldsymbol{\Delta}_{Ai} \rangle = 0, \; \forall i. \end{aligned} \tag{3.2}$$

Intuitively, the local property of the objective function of the main optimization problem (2.8) should be related to its behavior over the tangent space at the desired solution $(\mathbf{A}, \mathbf{X})$. Indeed, Lemma 3.1 shows that if $(\mathbf{A}, \mathbf{X})$ is locally optimal over the constraint manifold $\mathcal{M}$ in (2.8), then $(\mathbf{A}, \mathbf{X})$ is also a local optimal point over the tangent space (3.1). Since the ℓ_1-minimization over the tangent space is a convex optimization problem, local optimum over the tangent space is also global optimum. Further, the converse is also true; i.e., global optimum over the tangent space implies the local optimum over the constraint manifold $\mathcal{M}$.

Lemma 3.1. *Suppose that* $(\mathbf{\Delta}_A, \mathbf{\Delta}_X) = (\mathbf{0}, \mathbf{0})$ *is the unique optimal solution to* (3.2). *Then* $(\mathbf{A}, \mathbf{X})$ *is a local minimum of (2.8). Conversely, if* $(\mathbf{A}, \mathbf{X})$ *is a local minimum of (2.8), then* $(\mathbf{\Delta}_A, \mathbf{\Delta}_X) = (\mathbf{0}, \mathbf{0})$ *is an optimal solution to* (3.2).

Proof. Let $T_{\mathbf{x}}\mathcal{M}$ denote the tangent space of $\mathcal{M}$ at the point $\mathbf{x} = (\mathbf{A}, \mathbf{X})$. For any $\delta \in T_{\mathbf{x}}\mathcal{M}$, let $\mathbf{x}_\delta : (-\epsilon, \epsilon) \to \mathcal{M}$ be the geodesic satisfying $\mathbf{x}_\delta(0) = \mathbf{x}$ and $\dot{\mathbf{x}}_\delta(0) = \delta$. Then

$$\mathbf{x}_\delta(t) = \mathbf{x} + t\delta + O(t^2). \tag{3.3}$$

We first prove that global optimality over the tangent space is necessary. To simplify the notation, let $f(\cdot)$ denote the ℓ_1-norm, which is the objective function of (2.8). Suppose there exists $\delta \in T_{\mathbf{x}}\mathcal{M}$ with $f(\mathbf{x}+\delta) < f(\mathbf{x})$. Set $\tau = f(\mathbf{x}) - f(\mathbf{x}+\delta) > 0$. By convexity of ℓ_1-norm, for $\eta \in [0, 1]$,

$$f(\mathbf{x} + \eta\delta) \leq f(\mathbf{x}) - \eta\tau. \tag{3.4}$$

But,

$$\begin{aligned} f(\mathbf{x}_\delta(t)) &= f(\mathbf{x} + \eta\delta + (\mathbf{x}_\delta(t) - (\mathbf{x} + \eta\delta))) && (3.5)\\ &\leq f(\mathbf{x} + \eta\delta) + L\|\mathbf{x}_\delta(t) - \mathbf{x} - \eta\delta\|_2 && (3.6)\\ &\leq f(\mathbf{x}) - \eta\tau + L\|(t-\eta)\delta\|_2 + O(Lt^2). && (3.7) \end{aligned}$$

When t is sufficiently small and letting $\eta = t$, this value is strictly smaller than $f(\mathbf{x})$.

Conversely, suppose that $\delta = \mathbf{0}$ is the unique minimizer of $f(\mathbf{x}+\delta)$ over $\delta \in T_{\mathbf{x}}\mathcal{M}$. We will show that this minimizer is strongly unique, i.e., $\exists\, \beta > 0$ such that

$$f(\mathbf{x} + \delta) \geq f(\mathbf{x}) + \beta\|\delta\| \quad \forall\, \delta \in T_{\mathbf{x}}\mathcal{M}. \tag{3.8}$$

To see this, notice that if we write $\mathbf{x} = (\mathbf{A}, \mathbf{X})$ and $\delta = (\mathbf{\Delta}_A, \mathbf{\Delta}_X)$, then $f(\mathbf{x}) = \|\mathbf{X}\|_1$. Hence, if we set $r_0 = \min\{|X_{ij}| \mid X_{ij} \neq 0\} > 0$, whenever $\|\mathbf{\Delta}_X\|_\infty < r_0$ and $t < 1$, we have

$$\begin{aligned} \|\mathbf{X} + t\mathbf{\Delta}_X\|_1 &= \|\mathbf{X}\|_1 + t\langle \mathbf{\Sigma}, \mathbf{\Delta}_X\rangle + t\|\mathcal{P}_{\Omega^c}\mathbf{\Delta}_X\|_1 && (3.9)\\ &= \|\mathbf{X}\|_1 + t\langle \mathbf{\Sigma} + \operatorname{sign}(\mathcal{P}_{\Omega^c}\mathbf{\Delta}_X), \mathbf{\Delta}_X\rangle. && (3.10) \end{aligned}$$

Set $\beta(\delta) \doteq \langle \mathbf{\Sigma} + \operatorname{sign}(\mathcal{P}_{\Omega^c}\mathbf{\Delta}_X), \mathbf{\Delta}_X\rangle$, and notice that β is a continuous function of δ. Let

$$\beta^\star = \inf_{\delta \in T_{\mathbf{x}}\mathcal{M}, \|\delta\|=r_0/2} \beta(\delta) \geq 0. \tag{3.11}$$

Then for all $\delta \in T_{\mathbf{x}}\mathcal{M}$ with $\|\delta\| \leq r_0/2$, we have

$$f(\mathbf{x}+\delta) \geq f(\mathbf{x}) + (2\beta^\star/r_0)\|\delta\|. \tag{3.12}$$

Moreover, by convexity of f, the same bound holds for all $\delta \in T_{\mathbf{x}}\mathcal{M}$ (regardless of $\|\delta\|$). It remains to show that $\beta^\star$ is strictly larger than zero. On the contrary, suppose $\beta^\star = 0$. Since the infimum in (3.11) is taken over a compact set, it is achieved by some $\delta^\star \in T_{\mathbf{x}}\mathcal{M}$. Hence, if $\beta^\star = 0$, $f(\mathbf{x}+\delta^\star) = f(\mathbf{x})$, contradicting the uniqueness of the minimizer $\mathbf{x}$. This establishes (3.8).

Hence, continuing forward, we have

$$f(\mathbf{x}_\delta(\eta)) \geq f(\mathbf{x}) + \eta\beta\|\delta\| - O(\beta\eta^2). \tag{3.13}$$

For η sufficiently small, the right-hand side is strictly greater than $f(\mathbf{x})$, and thus $\mathbf{x} = (\mathbf{A}, \mathbf{X})$ is local minimum of (2.8). □

Due to Lemma 3.1, to prove the local correctness of dictionary learning problem we only need to analyze the linearized subproblem (3.2). Since it is a ℓ_1-minimization problem with linear constraints, it would be appealing to use the tools developed for compressed sensing to prove the unique optimality. However, unfortunately, our linearized subproblem does not have the restricted isometry property (RIP), which is a dominant tool in compressed sensing. At a high level, RIP states that there is no sparse vector near the null space of constraint matrix. In our case the constraint matrix is specified by equality constraints in (3.2). Let $\mathbf{\Pi} \in \mathbb{R}^{n \times n}$ be a permutation matrix with no fixed point and set $\mathbf{\Delta}_A = -\mathbf{A}\mathbf{\Pi}, \Delta_X = \mathbf{\Pi}\mathbf{X}$.

Then, it is easy to see that $\mathbf{\Delta}_A\mathbf{X} + \mathbf{A}\mathbf{\Delta}_X = \mathbf{0}$. Moreover, for each i,

$$\langle \mathbf{A}_i, \mathbf{\Delta}_{Ai} \rangle = -\langle \mathbf{A}_i, \mathbf{A}_{\pi(i)} \rangle \approx 0, \tag{3.14}$$

which follows the incoherence property of $\mathbf{A}$. In the ideal case when $\mathbf{A}$ is an orthogonal matrix, $\langle \mathbf{A}_i, \mathbf{\Delta}_{Ai} \rangle$ is exactly zero. Thus we have constructed a sparse vector which lies in the null space of the constraint matrix.

The absence of RIP makes the linearized subproblem very difficult to attack by directly using tools in compressed sensing. Instead, our analysis is drawn on tools developed for the matrix completion problem [7], where RIP also fails. Motivated by the success of David Gross's golfing scheme for matrix completion [8], we derive relaxed optimality conditions of the linearized subproblem (3.2) in the next section and describe how to construct the dual variables to certify these conditions in Chapter 4.

3.2 Optimality Conditions of Linearized Subproblem

To show $(\mathbf{\Delta}_A, \mathbf{\Delta}_X) = (\mathbf{0}, \mathbf{0})$ is an optimal solution of (3.2), it is easy to derive the corresponding KKT conditions for the convex optimization problem (3.2). The KKT conditions say that $(\mathbf{0}, \mathbf{0})$ is the optimal solution if and only if there exist two dual variables, a matrix $\mathbf{\Lambda} \in \mathbb{R}^{m \times p}$ (corresponding to the constraint $\mathbf{\Delta}_A \mathbf{X} + \mathbf{A}\mathbf{\Delta}_X = \mathbf{0}$) and a diagonal matrix $\mathbf{\Gamma} \in \mathbb{R}^{n \times n}$ (corresponding to the constraint $\langle \mathbf{A}_i, \mathbf{\Delta}_{Ai} \rangle = 0$) satisfying

$$\mathbf{A}^* \mathbf{\Lambda} \in \partial \| \cdot \|_1 (\mathbf{X}) \tag{3.15}$$

$$\mathbf{\Lambda} \mathbf{X}^* = \mathbf{A}\mathbf{\Gamma}, \tag{3.16}$$

where $\partial \| \cdot \|_1$ denotes the subgradient of ℓ_1-norm function.

The first constraint simply asserts that each column $\mathbf{x}_j$ of $\mathbf{X}$ is the minimum ℓ_1-norm solution to $\mathbf{A}\mathbf{x} = \mathbf{y}_j$. Specifically, let $\Omega = \text{support}(\mathbf{X})$ and $\mathbf{\Sigma} = \text{sign}(\mathbf{X})$. Then (3.15) holds if and only if $\exists\, \mathbf{w}_1, \ldots, \mathbf{w}_p \in \mathbb{R}^m$ such that

$$\mathbf{A}^* \lambda_j = \mathbf{\Sigma}_j + \mathbf{w}_j, \quad \mathbf{P}_{\Omega_j} \mathbf{w}_j = 0, \; \|\mathbf{w}_j\|_\infty \le 1. \tag{3.17}$$

This constraint is quite familiar from the duality of ℓ_1-minimization in compressed sensing literature [3].

On the other hand, the second constraint (3.16) is less familiar. It essentially says that we cannot locally improve the objective function by changing the bases in $\mathbf{A}$. Since it states each column in $\mathbf{\Lambda}\mathbf{X}^*$ is proportional to the corresponding column in $\mathbf{A}$, it can be equivalently expressed as $\Phi[\mathbf{\Lambda}\mathbf{X}^*] = \mathbf{0}$, where

$$\Phi[\mathbf{M}] := [(\mathbf{I} - \mathbf{A}_1 \mathbf{A}_1^*)\mathbf{M}_1 \mid \cdots \mid (\mathbf{I} - \mathbf{A}_n \mathbf{A}_n^*)\mathbf{M}_n]. \tag{3.18}$$

It requires all λ_j to satisfy the equality constraint exactly, which makes it potentially more difficult to satisfy than (3.15).

Inspired by David Gross's golfing scheme for matrix completion [8], in Lemma 3.2 we trade off between the two constraints (3.15) and (3.16), and show that by tightening one constraint and relaxing the other one, under some balancedness property we can still guarantee the optimality of solution $(\mathbf{0},\mathbf{0})$.

Lemma 3.2. *Let* $\mathbf{A}$ *be a matrix with no* k*-sparse vectors in its null space. Suppose that there exists* $\alpha > 0$ *such that for all pairs* $(\boldsymbol{\Delta}_A, \boldsymbol{\Delta}_X)$ *satisfying* (3.1)

$$\|\mathcal{P}_{\Omega^c}\boldsymbol{\Delta}_X\|_F \;\geq\; \alpha\|\boldsymbol{\Delta}_A\|_F. \tag{3.19}$$

Then if there exists $\boldsymbol{\Lambda} \in \mathbb{R}^{m\times p}$ *such that*

$$\mathcal{P}_\Omega[\mathbf{A}^*\boldsymbol{\Lambda}] = \boldsymbol{\Sigma}, \quad \|\mathcal{P}_{\Omega^c}[\mathbf{A}^*\boldsymbol{\Lambda}]\|_\infty \leq 1/2, \tag{3.20}$$

and

$$\|\Phi[\boldsymbol{\Lambda}\mathbf{X}^*]\|_F < \frac{\alpha}{2}, \tag{3.21}$$

we conclude that $(\boldsymbol{\Delta}_A, \boldsymbol{\Delta}_X) = (\mathbf{0},\mathbf{0})$ *is the unique optimal solution to* (3.2).

Proof. Consider any feasible $(\boldsymbol{\Delta}_A, \boldsymbol{\Delta}_X) \neq (\mathbf{0},\mathbf{0})$. Choose $\mathbf{H} \in \partial\|\cdot\|_1(\mathbf{X})$ such that $\langle \mathbf{H}, \mathcal{P}_{\Omega^c}\boldsymbol{\Delta}_X\rangle = \|\mathcal{P}_{\Omega^c}\boldsymbol{\Delta}_X\|_1$, and notice that $\mathcal{P}_\Omega\mathbf{H} = \boldsymbol{\Sigma}$. Then

$$\|\mathbf{X} + \boldsymbol{\Delta}_X\|_1 \geq \|\mathbf{X}\|_1 + \langle \mathbf{H}, \boldsymbol{\Delta}_X\rangle. \tag{3.22}$$

Notice that since $(\boldsymbol{\Delta}_A, \boldsymbol{\Delta}_X)$ is feasible,

$$\begin{aligned}
\langle \boldsymbol{\Delta}_A, \boldsymbol{\Lambda}\mathbf{X}^*\rangle + \langle \boldsymbol{\Delta}_X, \mathbf{A}^*\boldsymbol{\Lambda}\rangle &= \langle \boldsymbol{\Delta}_A\mathbf{X}, \boldsymbol{\Lambda}\rangle + \langle \mathbf{A}\boldsymbol{\Delta}_X, \boldsymbol{\Lambda}\rangle && (3.23)\\
&= \langle \boldsymbol{\Delta}_A\mathbf{X} + \mathbf{A}\boldsymbol{\Delta}_X, \boldsymbol{\Lambda}\rangle && (3.24)\\
&= \langle \mathbf{0}, \boldsymbol{\Lambda}\rangle = 0. && (3.25)
\end{aligned}$$

Hence,

$$\begin{aligned}
&\|\mathbf{X} + \boldsymbol{\Delta}_X\|_1 && (3.26)\\
&\geq \|\mathbf{X}\|_1 + \langle \mathbf{H}, \boldsymbol{\Delta}_X\rangle - \langle \mathbf{A}^*\boldsymbol{\Lambda}, \boldsymbol{\Delta}_X\rangle - \langle \boldsymbol{\Lambda}\mathbf{X}^*, \boldsymbol{\Delta}_A\rangle && (3.27)\\
&= \|\mathbf{X}\|_1 + \langle \mathbf{H} - \mathbf{A}^*\boldsymbol{\Lambda}, \boldsymbol{\Delta}_X\rangle - \langle \boldsymbol{\Lambda}\mathbf{X}^*, \boldsymbol{\Delta}_A\rangle && (3.28)\\
&= \|\mathbf{X}\|_1 + \langle \mathcal{P}_\Omega[\mathbf{H} - \mathbf{A}^*\boldsymbol{\Lambda}], \mathcal{P}_\Omega\boldsymbol{\Delta}_X\rangle + \langle \mathcal{P}_{\Omega^c}[\mathbf{H} - \mathbf{A}^*\boldsymbol{\Lambda}], \mathcal{P}_{\Omega^c}\boldsymbol{\Delta}_X\rangle - \langle \Phi[\boldsymbol{\Lambda}\mathbf{X}^*], \boldsymbol{\Delta}_A\rangle && (3.29)
\end{aligned}$$

$$= \|\mathbf{X}\|_1 + \langle \mathcal{P}_{\Omega^c}[\mathbf{H} - \mathbf{A}^*\mathbf{\Lambda}], \mathcal{P}_{\Omega^c}\mathbf{\Delta}_X \rangle - \langle \Phi[\mathbf{\Lambda}\mathbf{X}^*], \mathbf{\Delta}_A \rangle \tag{3.30}$$

$$\geq \|\mathbf{X}\|_1 + \|\mathcal{P}_{\Omega^c}\mathbf{\Delta}_X\|_1/2 - \|\mathbf{\Delta}_A\|_F \|\Phi[\mathbf{\Lambda}\mathbf{X}^*]\|_F \tag{3.31}$$

$$\geq \|\mathbf{X}\|_1 + \left(\tfrac{1}{2} - \alpha^{-1}\|\Phi[\mathbf{\Lambda}\mathbf{X}^*]\|_F\right) \|\mathcal{P}_{\Omega^c}\mathbf{\Delta}_X\|_1. \tag{3.32}$$

In (3.29), we have used the fact that since $\mathbf{\Delta}_A$ is feasible, each column of $\mathbf{\Delta}_A$ is orthogonal to the corresponding column of $\mathbf{A}$, and thus $\Phi[\mathbf{\Delta}_A] = \mathbf{\Delta}_A$. Furthermore, it can be easily verified that Φ is self-adjoint, and thus $\langle \mathbf{\Lambda}\mathbf{X}^*, \Phi[\mathbf{\Delta}_A] \rangle = \langle \Phi[\mathbf{\Lambda}\mathbf{X}^*], \mathbf{\Delta}_A \rangle$. In (3.30), we have used that since $\mathbf{H} \in \partial\|\cdot\|_1$, $\mathcal{P}_\Omega \mathbf{H} = \mathbf{\Sigma} = \mathbf{P}_\Omega[\mathbf{A}^*\mathbf{\Lambda}]$.

The right-hand side of (3.32) is strictly greater than $\|\mathbf{X}\|_1$ provided that (i) $\|\Phi[\mathbf{\Lambda}\mathbf{X}^*]\|_F < \alpha/2$ and (ii) $\mathcal{P}_{\Omega^c}\mathbf{\Delta}_X \neq \mathbf{0}$. Condition (i) is simply (3.21), and (ii) is implied by the assumption (3.19) and our assumption on the nullspace of $\mathbf{A}$. Indeed, suppose $\mathcal{P}_{\Omega^c}\mathbf{\Delta}_X = \mathbf{0}$, then by (3.19), $\mathbf{\Delta}_A = \mathbf{0}$, and thus $\mathbf{A}\mathbf{\Delta}_X = 0$. Since $\mathcal{P}_{\Omega^c}\mathbf{\Delta}_X = \mathbf{0}$, each column of $\mathbf{\Delta}_X$ has sparsity size at most k. Due to the assumption that there are no k sparse vectors in the null space of $\mathbf{A}$, we have $\mathbf{\Delta}_X = 0$. But this contradicts the assumption $(\mathbf{\Delta}_A, \mathbf{\Delta}_X) \neq (\mathbf{0}, \mathbf{0})$. Therefore, $\mathcal{P}_{\Omega^c}\mathbf{\Delta}_X \neq \mathbf{0}$. This completes the proof of Lemma 3.2. □

The remainder of the argument is to show the condition (3.19) indeed holds with high probability and we can construct the dual variables $\mathbf{\Lambda} \in \mathbb{R}^{m \times p}$ to satisfy conditions (3.20) and (3.21) with high probability under the random model described in Section 2.1, provided the number of samples is large enough.

CHAPTER 4

CERTIFICATION PROCESS

In this chapter, we prove that provided the number of samples $p = \Omega(n^3 k)$, the desired dual variable $\mathbf{\Lambda} \in \mathbb{R}^{m \times p}$ in Lemma 3.2 indeed exists with high probability by explicitly constructing $\mathbf{\Lambda}$ using a Markov process.

Before describing the construction of dual variables in detail, we first state our main result on the existence of the desired dual variable.

Theorem 4.1. *There exist numerical constants $C_1, C_2, C_3 > 0$ such that if*

$$k \;\le\; \min\left\{ \frac{C_1}{\mu(\mathbf{A})}, C_2 n \right\}, \tag{4.1}$$

then for any $\alpha > 0$, there exists $\mathbf{\Lambda} \in \mathbb{R}^{m \times p}$ simultaneously satisfying the following three properties:

$$\mathcal{P}_\Omega[\mathbf{A}^*\mathbf{\Lambda}] \;=\; \mathbf{\Sigma}, \tag{4.2}$$

$$\|\mathcal{P}_{\Omega^c}[\mathbf{A}^*\mathbf{\Lambda}]\|_\infty \;\le\; 1/2, \tag{4.3}$$

$$\|\Phi\left[\mathbf{\Lambda}\mathbf{X}^*\right]\|_F \;<\; \alpha/2, \tag{4.4}$$

with probability at least

$$1 - C_3 \alpha^{-1} n^{3/2} k^{1/2} p^{-1/2} (\log p). \tag{4.5}$$

Due to Theorem 4.1, if $p = \Omega(n^3 k)$, then the probability of the existence of dual variable $\mathbf{\Lambda} \in \mathbb{R}^{m \times p}$ satisfying (4.2), (4.3) and (4.4) can be arbitrarily close to one, as n goes to infinity.

We will repeatedly apply the following lemma to prove Theorem 4.1.

Lemma 4.2. *For any given $p > 0$ and letting $\mathbf{x}_1, \ldots, \mathbf{x}_p$ be independent and identically distributed random vectors with $\mathbf{x}_j = \mathbf{P}_{\Omega_j}\mathbf{v}_j$, where the $\Omega_j \subset [n]$ are uniform random subsets of size k and $\mathbf{v}_j \sim_{i.i.d.} \mathcal{N}(0, n/kp)$, then* *there exists a positive integer* $t_\star \in [\lfloor (p-1)/2 \rfloor, p]$ *and a sequence of vectors* $\lambda_1, \dots, \lambda_{t_\star}$ *depending only on* $\mathbf{x}_1, \dots, \mathbf{x}_{t_\star}$ *such that*

$$\begin{aligned} \mathbf{P}_{\Omega_j} \mathbf{A}^* \lambda_j &= \operatorname{sign}(\mathbf{x}_j), \quad j = 1, \dots, t_\star && (4.6) \\ \|\mathbf{P}_{\Omega_j^c} \mathbf{A}^* \lambda_j\|_\infty &\leq 1/2, \qquad j = 1, \dots, t_\star && (4.7) \\ \mathbb{E}\Big[\Big\| \Phi[\sum_{j=1}^{t_\star} \lambda_j \mathbf{x}_j^*] \Big\|_F \Big] &\leq C n^{3/2} k^{1/2} p^{-1/2}, && (4.8) \end{aligned}$$

where C *is a numerical constant.*

Proof of Theorem 4.1. Choose $t_1 = t_\star \in [\lfloor (p-1)/2 \rfloor, p]$ according to Lemma 4.2, and let $\lambda_1, \dots, \lambda_{t_1}$ be the corresponding dual vectors indicated by Lemma 4.2. Then we have

$$\mathbb{E}\Big[\Big\| \Phi[\sum_{j=1}^{t_1} \lambda_j \mathbf{x}_j^*] \Big\|_F \Big] \leq C n^{3/2} k^{1/2} p^{-1/2}. \tag{4.9}$$

Moreover, if $p \geq 3$, $p - t_1 \leq 3p/4$. Consider the i.i.d. random vectors

$$\left(\tfrac{p}{p-t_1}\right)^{1/2} \mathbf{x}_{t_1+1}, \ \dots, \ \left(\tfrac{p}{p-t_1}\right)^{1/2} \mathbf{x}_p, \tag{4.10}$$

and they also satisfy the hypotheses of Lemma 4.2. Hence, we can apply Lemma 4.2 again: there exists $\delta \in [\lfloor (p - t_1 - 1)/2 \rfloor, p - t_1]$ and corresponding certificates $\lambda_{t_1+1}, \dots, \lambda_{t_1+\delta}$ satisfying (4.6) and (4.7), such that if we set $t_2 = t_1 + \delta$, we have

$$\begin{aligned} \mathbb{E}\Big[\Big\| \Phi[\sum_{j=t_1+1}^{t_2} \lambda_j \mathbf{x}_j^*] \Big\|_F \Big] &\leq \left(\frac{p - t_1}{p}\right)^{1/2} C n^{3/2} k^{1/2} (p - t_1)^{-1/2} \\ &= C n^{3/2} k^{1/2} p^{-1/2}. \end{aligned}$$

After applying Lemma 4.2 twice, we only need to certify the remaining at most $p - t_2 \leq \max((3/4)^2 p, 2)$ vectors $\mathbf{x}_{t_2+1}, \dots, \mathbf{x}_p$. Repeat this construction $O(\log p)$ times, and we will get a sequence of dual certificates $\lambda_1, \dots, \lambda_p$ satisfying (4.6)-(4.7), with

$$\mathbb{E}\Big[\Big\| \Phi[\sum_{j=1}^{p} \lambda_j \mathbf{x}_j^*] \Big\|_F \Big] \leq C' (\log p) n^{3/2} k^{1/2} p^{-1/2}.$$

Then by applying Markov inequality we obtain the desired probability in Theorem 4.1. $\square$

4.1 Markov Process Construction

In this section we show how to construct the random sequences of $\lambda_1, \dots, \lambda_{t_\star}$ as described in Lemma 4.2. Let $\mathbf{\Theta}_j \in \mathbb{R}^{m \times m}$ denote the projection matrix onto the orthogonal complement of the range of $\mathbf{A}_{\Omega_j}$:

$$\mathbf{\Theta}_j = \mathbf{I} - \mathbf{A}_{\Omega_j}(\mathbf{A}^*_{\Omega_j}\mathbf{A}_{\Omega_j})^{-1}\mathbf{A}^*_{\Omega_j}, \tag{4.11}$$

and let $\mathbf{Q}_j$ denote the residual at time j:

$$\mathbf{Q}_j \doteq \sum_{l=1}^{j} \Phi[\lambda_l \mathbf{x}^*_l]. \tag{4.12}$$

Let $\sigma_j = \mathrm{sign}(\mathbf{x}_j(\Omega_j))$ and set

$$\zeta_j = \begin{cases} \frac{1}{4}\frac{\mathbf{\Theta}_j\mathbf{Q}_{j-1}\mathbf{x}_j}{\|\mathbf{\Theta}_j\mathbf{Q}_{j-1}\mathbf{x}_j\|} & \mathbf{\Theta}_j\mathbf{Q}_{j-1}\mathbf{x}_j \neq \mathbf{0} \\ \mathbf{0} & \text{else} \end{cases} \tag{4.13}$$

$$\lambda_j^{LS} = \mathbf{A}_{\Omega_j}(\mathbf{A}^*_{\Omega_j}\mathbf{A}_{\Omega_j})^{-1}\sigma_j, \tag{4.14}$$

$$\lambda_j = \lambda_j^{LS} - \zeta_j. \tag{4.15}$$

While the above construction process looks quite complicated, the rationale for the construction is fairly simple: at each step we construct a dual certificate λ_j to deflate the residual $\mathbf{Q}_j$ as much as possible and at the same time satisfy the constraints $\mathbf{A}^*_{\Omega_j}\lambda_j = \sigma_j$ and $\|\mathbf{A}^*_{\Omega_j^c}\lambda_j\|_\infty \leq 1/2$. The least square solution λ_j^{LS} [1] helps us to locate the dual certificate. In particular, when the dictionary $\mathbf{A}$ satisfies low coherence property, $\lambda_j = \lambda_j^{LS} - \zeta_j$ will satisfy the two constraints (4.6) and (4.7). Indeed,

$$\mathbf{A}^*_{\Omega_j}\lambda_j = \mathbf{A}^*_{\Omega_j}(\lambda_j^{LS} - \zeta_j) = \mathbf{A}^*_{\Omega_j}\lambda_j^{LS} = \sigma_j, \tag{4.16}$$

[1] λ_j^{LS} is the least square solution to $\mathbf{A}^*_{\Omega_j}\lambda = \sigma_j$.

where we use the fact $\mathbf{A}_{\Omega_j}^* \mathbf{\Theta}_j = \mathbf{0}$. And for each $i \notin \Omega_j$,

$$|\mathbf{A}_i^* \lambda_j^{LS}| = |\mathbf{A}_i^* \mathbf{A}_{\Omega_j} (\mathbf{A}_{\Omega_j}^* \mathbf{A}_{\Omega_j})^{-1} \sigma_j| \le \|\mathbf{A}_i^* \mathbf{A}_{\Omega_j}\|_2 \|(\mathbf{A}_{\Omega_j}^* \mathbf{A}_{\Omega_j})^{-1}\| \|\sigma_j\|_2. \quad (4.17)$$

Since σ_j is a vector of size k with each component being either 1 or -1, $\|\sigma_j\|_2 = \sqrt{k}$. Under the assumption $k\mu(\mathbf{A}) < 1/2$, a standard argument in Appendix A shows that $\|(\mathbf{A}_{\Omega_j}^* \mathbf{A}_{\Omega_j})^{-1}\| \le 2$. In addition, since $\mathbf{A}_i^* \mathbf{A}_{\Omega_j}$ is a vector of length k with entries bounded by $\mu(\mathbf{A})$, $\|\mathbf{A}_i^* \mathbf{A}_{\Omega_j}\|_2 \le \mu(\mathbf{A})\sqrt{k}$. Combining these bounds, we have

$$\|\mathbf{A}_{\Omega_j^c}^* \lambda_j^{LS}\|_\infty \ = \ \max_{i \notin \Omega_j} |\mathbf{A}_i^* \lambda_j^{LS}| \ \le \ 2k\mu(\mathbf{A}). \quad (4.18)$$

Hence, further assuming $k\mu(\mathbf{A}) < 1/8$, we upper bound $\|\mathbf{A}_{\Omega_j^c}^* \lambda_j^{LS}\|_\infty$ by $1/4$ and thus

$$\begin{aligned} \left\|\mathbf{A}_{\Omega_j^c}^* \lambda_j\right\|_\infty &\le \left\|\mathbf{A}_{\Omega_j^c}^* \lambda_j^{LS}\right\|_\infty + \left\|\mathbf{A}_{\Omega_j^c}^* \zeta_j\right\|_\infty & (4.19) \\ &\le \frac{1}{4} + \max_i \|\mathbf{A}_i\|_2 \|\zeta_j\|_2 & (4.20) \\ &\le \frac{1}{2}, & (4.21) \end{aligned}$$

where in (4.20) we have used the fact that $\|\mathbf{A}_i\|_2 = 1$ and $\|\zeta_j\|_2 \le \frac{1}{4}$.

The term ζ_j is chosen to deflate the residual $\mathbf{Q}_j$ as much as possible. Indeed, ζ_j is a scaled version of the solution to the optimization problem

$$\underset{\zeta}{\text{minimize}} \ \|\mathbf{Q}_{j-1} + \zeta \mathbf{x}_j^*\|_F \quad \text{subject to} \quad \mathbf{A}_{\Omega_j}^* \zeta = \mathbf{0}. \quad (4.22)$$

The reason we make ζ_j have a small ℓ_2-norm is to guarantee (4.6) and (4.7) will not be violated. As shown in Section 4.2, ζ_j can successfully control the norm of the residual $\mathbf{Q}_j$.

4.2 Analysis

Our next task is to analyze the growth of the residual $\|\mathbf{Q}_j\|_F$ and show it is indeed very small with high probability. Let $\{\mathcal{F}_j\}_{1 \le j \le p}$ be the the natural filtration with respect to $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p$. Specifically, $\mathcal{F}_j$ is the σ-

algebra generated by $\Omega_1, \ldots, \Omega_j$ and $v_1, \ldots, v_j$, and we have

$$\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}_p. \tag{4.23}$$

Proof of Lemma 4.2. Since by definition $\mathbf{Q}_j = \mathbf{Q}_{j-1} + \Phi[\lambda_j \mathbf{x}_j^*]$,

$$\begin{aligned} &\mathbb{E}\left[\|\mathbf{Q}_j\|_F^2 \mid \mathcal{F}_{j-1}\right] \\ &= \|\mathbf{Q}_{j-1}\|_F^2 + \mathbb{E}\left[\|\Phi[\lambda_j \mathbf{x}_j^*]\|_F^2 \mid \mathcal{F}_{j-1}\right] + 2\mathbb{E}\left[\left\langle \mathbf{Q}_{j-1}, \Phi[\lambda_j \mathbf{x}_j^*]\right\rangle \mid \mathcal{F}_{j-1}\right]. \end{aligned} \tag{4.24}$$

We will show that there exist $\varepsilon(p) > 0$ and $\tau(p)$ to upper bound the last two terms on the right-hand side of (4.24) by

$$\mathbb{E}\left[\left\langle \mathbf{Q}_{j-1}, \Phi[\lambda_j \mathbf{x}_j^*]\right\rangle \mid \mathcal{F}_{j-1}\right] \;\leq\; -\varepsilon(p) \times \|\mathbf{Q}_{j-1}\|_F \tag{4.25}$$

$$\mathbb{E}\left[\|\Phi[\lambda_j \mathbf{x}_j^*]\|_F^2 \mid \mathcal{F}_{j-1}\right] \;\leq\; \tau(p). \tag{4.26}$$

Plugging into (4.24) and taking the expectation of both sides gives

$$\mathbb{E}[\|\mathbf{Q}_j\|_F^2] \leq \mathbb{E}[\|\mathbf{Q}_{j-1}\|_F^2] - 2\varepsilon(p)\mathbb{E}[\|\mathbf{Q}_{j-1}\|_F] + \tau(p). \tag{4.27}$$

Summing from $j = 1, \ldots, p$ and using the fact that $\mathbf{Q}_0 = \mathbf{0}$, we have

$$\mathbb{E}[\|\mathbf{Q}_p\|_F^2] \quad \leq \quad p\tau(p) - 2\varepsilon(p) \sum_{j=1}^{p-1} \mathbb{E}[\|\mathbf{Q}_j\|_F]. \tag{4.28}$$

In Sections 4.2.1, 4.2.2 and 4.2.3, we show that the quantities $\varepsilon(p)$ and $\tau(p)$ satisfy the following bounds:

$$\varepsilon(p) \;\geq\; C_1\sqrt{k/np}, \quad \text{and} \quad \tau(p) \;\leq\; C_2 nk/p. \tag{4.29}$$

Taking these bounds as given, by Jensen's inequality and (4.28) we get

$$\mathbb{E}[\|\mathbf{Q}_1\|_F] \;\leq\; (\mathbb{E}[\|\mathbf{Q}_1\|_F^2])^{1/2} \;\leq\; \sqrt{\tau(p)}, \tag{4.30}$$

and hence Lemma 4.2 is verified in the case $p = 1$. On the other hand, if $p > 1$, then since the left-hand side of (4.28) is nonnegative, we have

$$\frac{1}{p-1} \sum_{j=1}^{p-1} \mathbb{E}[\|\mathbf{Q}_j\|_F] \leq \frac{p}{p-1} \frac{\tau(p)}{2\,\varepsilon(p)} \leq \tau(p)/\varepsilon(p). \tag{4.31}$$

It is easy to recognize that the left-hand side of this inequality is an average. We claim there exists at least one $t_\star \in [\lfloor (p-1)/2 \rfloor, p]$ such that

$$\mathbb{E}[\|\mathbf{Q}_{t_\star}\|_F] \;\le\; 2\tau(p)/\varepsilon(p). \tag{4.32}$$

Indeed, suppose $\mathbb{E}[\|\mathbf{Q}_t\|_F] \;>\; 2\tau(p)/\varepsilon(p)$ for all $t \in [\lfloor (p-1)/2 \rfloor, p]$, then

$$\begin{aligned} \frac{1}{p-1}\sum_{j=1}^{p-1}\mathbb{E}[\|\mathbf{Q}_j\|_F] &\ge \frac{1}{p-1}\sum_{j=\lfloor (p-1)/2\rfloor}^{p-1}\mathbb{E}[\|\mathbf{Q}_j\|_F] && (4.33)\\ &> \frac{1}{p-1}\frac{p-1}{2}2\tau(p)/\varepsilon(p) && (4.34)\\ &= \tau(p)/\varepsilon(p), && (4.35) \end{aligned}$$

contradicting (4.31).

Combining (4.29) and (4.32) we establish Lemma 4.2. □

What remains to do is show that the bounds in (4.29) indeed hold. Splitting $\mathbb{E}\left[\langle \mathbf{Q}_{j-1}, \lambda_j \mathbf{x}_j^* \rangle \mid \mathcal{F}_{j-1}\right]$ as

$$\begin{aligned} &\mathbb{E}\left[\langle \mathbf{Q}_{j-1}, \lambda_j \mathbf{x}_j^* \rangle \mid \mathcal{F}_{j-1}\right] \\ =\;& \mathbb{E}\left[\langle \mathbf{Q}_{j-1}, \lambda_j^{LS} \mathbf{x}_j^* \rangle \mid \mathcal{F}_{j-1}\right] - \mathbb{E}\left[\langle \mathbf{Q}_{j-1}, \zeta_j \mathbf{x}_j^* \rangle \mid \mathcal{F}_{j-1}\right], \end{aligned} \tag{4.36}$$

we establish the bound on ε in Section 4.2.1 and Section 4.2.2 below by bounding the two terms on the right-hand side of (4.36) separately.

Finally, we establish the bound on τ in Section 4.2.3, completing the proof of Lemma 4.2.

4.2.1 Upper bounding $\langle \mathbf{Q}_{j-1}, \lambda_j^{LS} \mathbf{x}_j^* \rangle$

Given $\Omega_j = \{a_1 < a_2 < \cdots < a_k\}$, set $\mathbf{U}_{\Omega_j} \doteq [\mathbf{e}_{a_1} \mid \mathbf{e}_{a_2} \mid \cdots \mid \mathbf{e}_{a_k}] \in \mathbb{R}^{n\times k}$, where each $a_i \in [n]$ and $\mathbf{e}_i$ denotes the ith standard basis in $\mathbb{R}^n$. So we have $\mathbf{P}_{\Omega_j} = \mathbf{U}_{\Omega_j}\mathbf{U}_{\Omega_j}^*$, and we can write

$$\begin{aligned} &\left\langle \mathbf{Q}_{j-1}, \lambda_j^{LS} \mathbf{x}_j^* \right\rangle \\ =\;& \left\langle \mathbf{Q}_{j-1}\,,\, \mathbf{A}_{\Omega_j}(\mathbf{A}_{\Omega_j}^*\mathbf{A}_{\Omega_j})^{-1}\mathbf{U}_{\Omega_j}^*\mathrm{sgn}(\mathbf{v}_j)\mathbf{v}_j^*\mathbf{P}_{\Omega_j} \right\rangle. \end{aligned} \tag{4.37}$$

Since $\mathbb{E}[\text{sgn}(\mathbf{v}_j)\mathbf{v}_j^*] = c_1\sigma\mathbf{I}$, we have

$$\begin{aligned} & \mathbb{E}\left[\left\langle \mathbf{Q}_{j-1}, \lambda_j^{LS}\mathbf{x}_j^*\right\rangle \mid \mathcal{F}_{j-1}\right] \\ = \ & c_1\sigma\mathbb{E}_{\Omega_j}\left[\left\langle \mathbf{Q}_{j-1}\, ,\, \mathbf{A}_{\Omega_j}(\mathbf{A}_{\Omega_j}^*\mathbf{A}_{\Omega_j})^{-1}\mathbf{U}_{\Omega_j}^*\right\rangle\right]. \end{aligned} \tag{4.38}$$

Writing $(\mathbf{A}_{\Omega_j}^*\mathbf{A}_{\Omega_j})^{-1} = \mathbf{I} + \mathbf{\Delta}(\Omega_j)$, then we get

$$\begin{aligned} & \left\langle \mathbf{Q}_{j-1}\, ,\, \mathbf{A}_{\Omega_j}(\mathbf{A}_{\Omega_j}^*\mathbf{A}_{\Omega_j})^{-1}\mathbf{U}_{\Omega_j}^*\right\rangle \\ = \ & \left\langle \mathbf{Q}_{j-1}\mathbf{P}_{\Omega_j}\, ,\, \mathbf{A}_{\Omega_j}(\mathbf{A}_{\Omega_j}^*\mathbf{A}_{\Omega_j})^{-1}\mathbf{U}_{\Omega_j}^*\right\rangle & (4.39) \\ = \ & \left\langle \mathbf{Q}_{j-1}\mathbf{P}_{\Omega_j}\, ,\, \mathbf{A}_{\Omega_j}\mathbf{U}_{\Omega_j}^*\right\rangle + \left\langle \mathbf{Q}_{j-1}\mathbf{P}_{\Omega_j}\, ,\, \mathbf{A}_{\Omega_j}\mathbf{\Delta}(\Omega_j)\mathbf{U}_{\Omega_j}^*\right\rangle. & (4.40) \end{aligned}$$

Since $\mathbf{Q}_{j-1} = \Phi[\sum_{i=0}^{j-1}\lambda_i\mathbf{x}_i^*] \in \text{range}(\Phi)$, each column of $\mathbf{Q}_{j-1}$ is orthogonal to the corresponding column of $\mathbf{A}$. Note that the first inner product in (4.40) is simply the inner product of the restriction of $\mathbf{A}$ to a subset of its columns and the restriction of $\mathbf{Q}_{j-1}$ to a subset of its columns, and thus this term is zero. Applying the Cauchy-Schwarz inequality to the second term of (4.40) gives

$$\left\langle \mathbf{Q}_{j-1}\, ,\, \mathbf{A}_{\Omega_j}(\mathbf{A}_{\Omega_j}^*\mathbf{A}_{\Omega_j})^{-1}\mathbf{U}_{\Omega_j}^*\right\rangle \leq \|\mathbf{Q}_{j-1}\mathbf{P}_{\Omega_j}\|_F\,\|\mathbf{A}_{\Omega_j}\|\|\mathbf{\Delta}(\Omega_j)\|_F. \tag{4.41}$$

Standard calculations in Appendix A show that $\|\mathbf{A}_{\Omega_j}\| \leq (1 + k\mu(\mathbf{A}))^{1/2}$, and $\|\mathbf{\Delta}(\Omega_j)\|_F \leq 2k\mu(\mathbf{A})$. Plugging back into (4.38), we have

$$\mathbb{E}\left[\left\langle \mathbf{Q}_{j-1}, \lambda_j^{LS}\mathbf{x}_j^*\right\rangle \mid \mathcal{F}_{j-1}\right] \leq 2c_1\sigma k\mu(\mathbf{A})(1 + k\mu(\mathbf{A}))^{1/2}\,\mathbb{E}_{\Omega_j}[\|\mathbf{Q}_{j-1}\mathbf{P}_{\Omega_j}\|_F]. \tag{4.42}$$

4.2.2 Lower bounding $\langle \mathbf{Q}_{j-1}, \zeta_j\mathbf{x}_j^*\rangle$

We have

$$\begin{aligned} \langle \mathbf{Q}_{j-1}, \zeta_j\mathbf{x}_j^*\rangle &= \langle \mathbf{Q}_{j-1}\mathbf{x}_j, \zeta_j\rangle & (4.43) \\ &= \tfrac{1}{4}\left\langle \mathbf{Q}_{j-1}\mathbf{x}_j, \frac{\mathbf{\Theta}_j\mathbf{Q}_{j-1}\mathbf{x}_j}{\|\mathbf{\Theta}_j\mathbf{Q}_{j-1}\mathbf{x}_j\|}\right\rangle & (4.44) \\ &= \tfrac{1}{4}\|\mathbf{\Theta}_j\mathbf{Q}_{j-1}\mathbf{x}_j\| & (4.45) \\ &\geq \tfrac{1}{4}\|\mathbf{Q}_{j-1}\mathbf{x}_j\| - \tfrac{1}{4}\left\|\mathbf{P}_{A_{\Omega_j}}\mathbf{Q}_{j-1}\mathbf{x}_j\right\| & (4.46) \end{aligned}$$

where $\mathbf{P}_{A_{\Omega_j}} = \mathbf{A}_{\Omega_j}(\mathbf{A}^*_{\Omega_j}\mathbf{A}_{\Omega_j})^{-1}\mathbf{A}^*_{\Omega_j} \in \mathbb{R}^{m\times m}$. Applying the Kahane-Khintchine inequality in Corollary B.3 to the first term of (4.46) gives

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{Q}_{j-1}\mathbf{x}_j\| \mid \mathcal{F}_{j-1}\right] &= \mathbb{E}_{\Omega_j}\mathbb{E}_{\mathbf{v}_j}\left[\|\mathbf{Q}_{j-1}\mathbf{P}_{\Omega_j}\mathbf{v}_j\|\right] && (4.47)\\
&\geq \frac{\sigma}{\sqrt{\pi}} \times \mathbb{E}_{\Omega_j}\left[\|\mathbf{Q}_{j-1}\mathbf{P}_{\Omega_j}\|_F\right]. && (4.48)
\end{aligned}$$

For the second term of (4.46), write $\mathbf{P}_{\mathbf{A}_{\Omega_j}}$ as

$$\mathbf{P}_{\mathbf{A}_{\Omega_j}} = \mathbf{A}_{\Omega_j}(\mathbf{A}^*_{\Omega_j}\mathbf{A}_{\Omega_j})^{-1/2} \times (\mathbf{A}^*_{\Omega_j}\mathbf{A}_{\Omega_j})^{-1/2}\mathbf{A}^*_{\Omega_j}, \tag{4.49}$$

and we get

$$\begin{aligned}
\left\|\mathbf{P}_{\mathbf{A}_{\Omega_j}}\mathbf{Q}_{j-1}\mathbf{x}_j\right\| &= \left\|(\mathbf{A}^*_{\Omega_j}\mathbf{A}_{\Omega_j})^{-1/2}\mathbf{A}^*_{\Omega_j}\mathbf{Q}_{j-1}\mathbf{P}_{\Omega_j}\mathbf{v}_j\right\| && (4.50)\\
&\leq \left\|(\mathbf{A}^*_{\Omega_j}\mathbf{A}_{\Omega_j})^{-1/2}\right\| \left\|\mathbf{A}^*_{\Omega_j}\mathbf{Q}_{j-1}\mathbf{P}_{\Omega_j}\mathbf{v}_j\right\| && (4.51)\\
&\leq \sqrt{2} \times \left\|\mathbf{A}^*_{\Omega_j}\mathbf{Q}_{j-1}\mathbf{P}_{\Omega_j}\mathbf{v}_j\right\| && (4.52)\\
&= \sqrt{2} \times \left\|\mathbf{P}_{\Omega_j}\mathbf{A}^*\mathbf{Q}_{j-1}\mathbf{P}_{\Omega_j}\mathbf{v}_j\right\|, && (4.53)
\end{aligned}$$

where in (4.52) we have used the fact that $\|(\mathbf{A}^*_{\Omega_j}\mathbf{A}_{\Omega_j})^{-1/2}\| \leq \sqrt{2}$ under the assumption $k\mu(\mathbf{A}) < 1/2$.

Applying the Jensen's inequality to bound the expectation of (4.52), we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{P}_{\mathbf{A}_{\Omega_j}}\mathbf{Q}_{j-1}\mathbf{x}_j\right\| \mid \mathcal{F}_{j-1}\right]\\
\leq\ & \sqrt{2} \times \mathbb{E}\left[\left\|\mathbf{P}_{\Omega_j}\mathbf{A}^*\mathbf{Q}_{j-1}\mathbf{P}_{\Omega_j}\mathbf{v}_j\right\| \mid \mathcal{F}_{j-1}\right] && (4.54)\\
=\ & \sqrt{2} \times \mathbb{E}_{\Omega_j}\mathbb{E}_{\mathbf{v}_j}\left[\left\|\mathbf{P}_{\Omega_j}\mathbf{A}^*\mathbf{Q}_{j-1}\mathbf{P}_{\Omega_j}\mathbf{v}_j\right\|\right] && (4.55)\\
\leq\ & \sigma\sqrt{2} \times \mathbb{E}_{\Omega_j}\left[\left\|\mathbf{P}_{\Omega_j}\mathbf{A}^*\mathbf{Q}_{j-1}\mathbf{P}_{\Omega_j}\right\|_F\right]. && (4.56)
\end{aligned}$$

Since each column of $\mathbf{Q}_{j-1}$ is orthogonal to the corresponding column of $\mathbf{A}$, the diagonal elements of $\mathbf{A}^*\mathbf{Q}_{j-1}$ are zero. By invoking the decoupling lemma in Appendix C, we can remove the first $\mathbf{P}_{\Omega_j}$ in (4.56) and get

$$\begin{aligned}
&\mathbb{E}_{\Omega_j}\left[\left\|\mathbf{P}_{\Omega_j}\mathbf{A}^*\mathbf{Q}_{j-1}\mathbf{P}_{\Omega_j}\right\|_F\right]\\
\leq\ & 16\sqrt{\frac{k}{n}}\,\mathbb{E}_{\Omega_j}\left[\left\|\mathbf{A}^*\mathbf{Q}_{j-1}\mathbf{P}_{\Omega_j}\right\|_F\right] && (4.57)
\end{aligned}$$

$$\leq \quad 16\|\mathbf{A}\|\sqrt{\frac{k}{n}}\,\mathbb{E}_{\Omega_j}\left[\left\|\mathbf{Q}_{j-1}\mathbf{P}_{\Omega_j}\right\|_F\right]. \tag{4.58}$$

Since $\|\mathbf{A}\| \leq \sqrt{1+n\mu(\mathbf{A})}$ (see Appendix A), we have

$$\mathbb{E}_{\Omega_j}\left[\left\|\mathbf{P}_{\Omega_j}\mathbf{A}^*\mathbf{Q}_{j-1}\mathbf{P}_{\Omega_j}\right\|_F\right] \leq c_3\sqrt{k/n+k\mu(\mathbf{A})}\,\mathbb{E}_{\Omega_j}\left[\left\|\mathbf{Q}_{j-1}\mathbf{P}_{\Omega_j}\right\|_F\right], \tag{4.59}$$

for appropriate numerical constant c_3.

Combining (4.36), (4.42), (4.48), (4.59) and (4.58), we get

$$\mathbb{E}\left[\langle \mathbf{Q}_{j-1}, \lambda_j \mathbf{x}_j^* \rangle \mid \mathcal{F}_{j-1}\right] \leq -c_4\sigma\mathbb{E}_{\Omega_j}\left[\left\|\mathbf{Q}_{j-1}\mathbf{P}_{\Omega_j}\right\|_F\right], \tag{4.60}$$

where

$$c_4 = -2c_1(1+k\mu(\mathbf{A}))^{1/2}k\mu(\mathbf{A}) + \frac{c_3}{4}\sqrt{k/n+k\mu(\mathbf{A})} - \frac{1}{4\sqrt{\pi}}.$$

Assuming k/n and $k\mu(\mathbf{A})$ are appropriately small constants, c_4 is strictly positive and we have

$$\begin{aligned}
\mathbb{E}\left[\langle \mathbf{Q}_{j-1}, \lambda_j \mathbf{x}_j^* \rangle \mid \mathcal{F}_{j-1}\right] &\leq -c_4\sigma\mathbb{E}_{\Omega_j}\left[\left\|\mathbf{Q}_{j-1}\mathbf{P}_{\Omega_j}\right\|_F\right] && (4.61)\\
&\leq -c_4\sigma\left\|\mathbb{E}_{\Omega_j}\left[\mathbf{Q}_{j-1}\mathbf{P}_{\Omega_j}\right]\right\|_F && (4.62)\\
&\leq -c_4\sigma(k/n)\|\mathbf{Q}_{j-1}\|_F && (4.63)\\
&= -c_4\sqrt{k/np}\,\|\mathbf{Q}_{j-1}\|_F, && (4.64)
\end{aligned}$$

where we have used Jensen's inequality and the facts that $\mathbb{E}_{\Omega_j}[\mathbf{P}_{\Omega_j}] = (k/n)\mathbf{I}$ and $\sigma = \sqrt{n/kp}$. This establishes the first part of (4.29).

4.2.3 Bounding $\|\lambda_j \mathbf{x}_j^*\|$

We have already shown that under the conditions of Theorem 4.1,

$$\|\lambda_j\|_2 \leq \|\lambda_j^{LS}\|_2 + \|\zeta_j\|_2 \leq c_5\sqrt{k} + 1/4 \leq c_6\sqrt{k}. \tag{4.65}$$

Hence,

$$\left\|\Phi\left[\lambda_j \mathbf{x}_j^*\right]\right\|_F^2 \;\leq\; \left\|\lambda_j \mathbf{x}_j^*\right\|_F^2 \;=\; \|\lambda_j\|^2\|\mathbf{x}_j\|^2 \;\leq\; c_6 k\|\mathbf{x}_j\|^2. \tag{4.66}$$

Since $\mathbb{E}[\|\mathbf{x}_j\|_2^2] = n/p$, we have the following simple bound

$$\mathbb{E}\left[\left\|\Phi[\lambda_j \mathbf{x}_j^*]\right\|_F^2 \mid \mathcal{F}_{j-1}\right] \;\leq\; c_6 kn/p. \tag{4.67}$$

This establishes the second part of (4.29).

CHAPTER 5

BALANCEDNESS PROPERTY

In this chapter, we show that the condition in (3.19) indeed holds with high probability; i.e., we will prove that for any $(\mathbf{\Delta}_A, \mathbf{\Delta}_X)$ in the tangent space to $\mathcal{M}$ at $(\mathbf{A}, \mathbf{X})$,

$$\|\mathcal{P}_{\Omega^c}\mathbf{\Delta}_X\|_F \;\geq\; \alpha\|\mathbf{\Delta}_A\|_F \tag{5.1}$$

for appropriate $\alpha > 0$. This property essentially says that if we locally perturb the basis $\mathbf{A}$, we are guaranteed to pay some penalty by increasing the Frobenius norm of $\mathcal{P}_{\Omega^c}\mathbf{X}$, and thus may not improve the objective function $\|\mathbf{X}\|_1$. However, this balanceness property itself is not sufficient to establish our main result Theorem 2.1, since it does not rule not the possibility that as $\mathbf{A}$ changes, $\|\mathcal{P}_{\Omega}\mathbf{\Delta}_X\|_1$ might decrease faster than $\|\mathcal{P}_{\Omega^c}\mathbf{\Delta}_X\|_1$ increases. For this purpose we need the golfing scheme in Chapter 4 to construct dual variables to show this indeed will not happen.

More precisely, our main result on the balancedness property is the following:

Theorem 5.1. *There exist numerical constants $C_1 \dots C_8 > 0$ such that if*

$$k \;\leq\; C_1 \times \min\left\{n, \frac{1}{\mu(\mathbf{A})}\right\}, \tag{5.2}$$

then whenever $p \geq C_2 n^2$, with probability at least

$$1 \;-\; C_3 p^{-4} - C_4\, n\, \exp\left(-\frac{C_5\, p}{n \log p}\right) \;-\; C_6 n^2 \exp\left(-\frac{C_7 k^2 p}{n^2}\right), \tag{5.3}$$

all pairs $(\mathbf{\Delta}_A, \mathbf{\Delta}_X)$ satisfying (3.1) obey the estimate

$$\|\mathcal{P}_{\Omega^c}\mathbf{\Delta}_X\|_F \;\geq\; C_8\|\mathbf{\Delta}_A\|_F/\|\mathbf{A}\|^2. \tag{5.4}$$

The proof of Theorem 5.1 mainly consists of two parts. First we show the desired property (5.1) holds whenever the random matrix $\mathbf{X}$ satisfies two particular algebraic properties. Then using probabilistic analysis we prove these algebraic properties are indeed satisfied with high probability under the aforementioned probabilistic model. In particular, we will apply two technical tools on matrix norm bound, which are given in Appendix B. The first algebraic property, which is stated in Lemma 5.2, involves a bound on the extreme eigenvalue of $\mathbf{X}\mathbf{X}^*$, and this property is proved in Appendix D using matrix Chernoff bound of [9]. The second algebraic property involves controlling the difference between a matrix operator and its large sample limit, and it is formally stated in Lemma 5.3. Since the proof of Lemma 5.3 is a little technical, the proof is given in Appendix E.

Before proving Theorem 5.1, we introduce one additional definition. Fix $0 < t < 1/2$, and let $\mathcal{E}_{eig}(t)$ denote the event:

$$\mathcal{E}_{eig}(t) \doteq \{ \omega \mid \|\mathbf{X}\mathbf{X}^* - \mathbf{I}\| < t \}. \tag{5.5}$$

In particular, in the event $\mathcal{E}_{eig}(t)$, $\|\mathbf{X}\mathbf{X}^*\| < 1 + t < 2$, and $\|(\mathbf{X}\mathbf{X}^*)^{-1}\| = (\lambda_{min}(\mathbf{X}\mathbf{X}^*))^{-1} < 1/(1-t) < 2$. Lemma 5.2 shows that if $\mathbf{X}$ has sufficiently many i.i.d. columns, then the probability of $\mathcal{E}_{eig}(t)$ is close to one.

Lemma 5.2. *Fix any $0 < t < 1/2$, and let $\mathcal{E}_{eig}(t) \doteq \{ \omega \mid \|\mathbf{X}\mathbf{X}^* - \mathbf{I}\| < t\}$. Then there exists numerical constants C_1, C_2 and C_3 all strictly positive such that for all $p \geq C_1 (n/t)^{1/4}$,*

$$\mathbb{P}\left[\mathcal{E}_{eig}(t)\right] \geq 1 - C_2\, n \exp\left(-\frac{C_3\, t^2\, p}{n \log p} \right) - p^{-7}. \tag{5.6}$$

Lemma 5.2 is essentially a consequence of the matrix Chernoff bound of [9] and the proof is given in Appendix D. For now we take this result as given and apply it to prove Theorem 5.1.

Proof of Theorem 5.1. Due to Lemma 5.2, our analysis can be restricted in the event $\mathcal{E}_{eig}$. In the event $\mathcal{E}_{eig}$, $\mathbf{X}\mathbf{X}^*$ is invertible, and for any pair $(\mathbf{\Delta}_A, \mathbf{\Delta}_X)$ satisfying (3.1) we can write $\mathbf{\Delta}_A$ as

$$\mathbf{\Delta}_A = -\mathbf{A}\mathbf{\Delta}_X \mathbf{X}^* (\mathbf{X}\mathbf{X}^*)^{-1}. \tag{5.7}$$

Using the facts that

$$\|\mathbf{X}^*(\mathbf{X}\mathbf{X}^*)^{-1}\| = 1/\sqrt{\lambda_{min}(\mathbf{X}\mathbf{X}^*)} \tag{5.8}$$

and that for any matrices $\mathbf{P}, \mathbf{Q}, \mathbf{R}$,

$$\|\mathbf{PQR}\|_F \leq \|\mathbf{P}\|\|\mathbf{R}\|\|\mathbf{Q}\|_F, \tag{5.9}$$

on $\mathcal{E}_{eig}(1/2)$ we have

$$\|\boldsymbol{\Delta}_A\|_F \;\leq\; \frac{\|\mathbf{A}\|}{\sqrt{\lambda_{min}(\mathbf{X}\mathbf{X}^*)}}\|\boldsymbol{\Delta}_X\|_F \;\leq\; \sqrt{2}\,\|\mathbf{A}\|\,\|\boldsymbol{\Delta}_X\|_F. \tag{5.10}$$

Next we show that for any pair $(\boldsymbol{\Delta}_A, \boldsymbol{\Delta}_X)$ satisfying (3.1), we further have

$$\|\mathcal{P}_\Omega[\boldsymbol{\Delta}_X]\|_F \;\leq\; \alpha'\|\mathcal{P}_{\Omega^c}[\boldsymbol{\Delta}_X]\|_F, \tag{5.11}$$

for some positive constant α'.

Plug (5.7) into (3.1) and we get

$$\mathbf{0} = \boldsymbol{\Delta}_A\mathbf{X} + \mathbf{A}\boldsymbol{\Delta}_X \;=\; \mathbf{A}\boldsymbol{\Delta}_X\left(\mathbf{I} - \mathbf{X}^*(\mathbf{X}\mathbf{X}^*)^{-1}\mathbf{X}\right). \tag{5.12}$$

Note that above $\mathbf{P}_X \doteq \mathbf{X}^*(\mathbf{X}\mathbf{X}^*)^{-1}\mathbf{X}$ is the projection matrix onto the range of $\mathbf{X}^*$.

We have one further constraint $\mathbf{A}_i^*\boldsymbol{\Delta}_A\mathbf{e}_i = 0\ \forall i$. Let $\mathcal{C}_A : \mathbb{R}^n \to \mathbb{R}^{m\times n}$ via

$$\mathcal{C}_A[\mathbf{z}] = \mathbf{A}\ \mathrm{diag}(\mathbf{z}). \tag{5.13}$$

For $\mathbf{U} = [\mathbf{u}_1 \mid \mathbf{u}_2 \mid \cdots \mid \mathbf{u}_n] \in \mathbb{R}^{m\times n}$, the action of the adjoint of $\mathcal{C}_A$ on $\mathbf{U}$ is given by

$$\mathcal{C}_A^*[\mathbf{U}] = [\langle \mathbf{A}_1, \mathbf{u}_1\rangle, \ldots, \langle \mathbf{A}_n, \mathbf{u}_n\rangle]^* \in \mathbb{R}^n. \tag{5.14}$$

Hence, our second constraint $\mathbf{A}_i^*\boldsymbol{\Delta}_A\mathbf{e}_i = 0\ \forall i$ can be expressed concisely via

$$\mathcal{C}_A^*[\boldsymbol{\Delta}_A] = \mathbf{0} \in \mathbb{R}^n. \tag{5.15}$$

Combining (5.12) and (5.15), we get

$$\mathbf{A}\boldsymbol{\Delta}_X(\mathbf{I} - \mathbf{P}_X) = \mathbf{0} \quad \text{and} \quad \mathcal{C}_A^*[\mathbf{A}\boldsymbol{\Delta}_X\mathbf{X}^*(\mathbf{X}\mathbf{X}^*)^{-1}] = \mathbf{0}. \tag{5.16}$$

It would be convenient to temporarily express the constraint (5.16) in the vector form, as a constraint on $\delta_X \doteq \mathrm{vec}[\mathbf{\Delta}_X] \in \mathbb{R}^{np}$. In vector notation, (5.16) is equivalent to $\mathbf{M}\delta_x = \mathbf{0}$, with

$$\mathbf{M} \doteq \begin{bmatrix} (\mathbf{I} - \mathbf{P}_X) \otimes \mathbf{A} \\ \mathbf{C}_A^*((\mathbf{X}\mathbf{X}^*)^{-1}\mathbf{X} \otimes \mathbf{A}) \end{bmatrix} \in \mathbb{R}^{(mp+n)\times np}, \tag{5.17}$$

where we have used the familiar identity $\mathrm{vec}[\mathbf{QRS}] = (\mathbf{S}^* \otimes \mathbf{Q})\,\mathrm{vec}[\mathbf{R}]$, for matrices $\mathbf{Q}$, $\mathbf{R}$, and $\mathbf{S}$ of compatible size, and we have used $\mathbf{C}_A$ to denote the matrix version of the operator $\mathcal{C}_A$, uniquely defined via[1]

$$\mathrm{vec}[\mathcal{C}_A[\mathbf{z}]] \;=\; \mathbf{C}_A\mathbf{z} \quad \forall\, \mathbf{z} \in \mathbb{R}^{m\times n}. \tag{5.18}$$

Consider a symmetric variant of the equation $\mathbf{M}\delta_x = \mathbf{0}$, by setting

$$\begin{aligned} \mathbf{T} \;&\doteq\; \mathbf{M}^*\mathbf{M} \\ &= (\mathbf{I} - \mathbf{P}_X) \otimes \mathbf{A}^*\mathbf{A} \;+\; \left(\mathbf{X}^*(\mathbf{X}\mathbf{X}^*)^{-1} \otimes \mathbf{A}^*\right)\, \mathbf{C}_A\mathbf{C}_A^* \,\left((\mathbf{X}\mathbf{X}^*)^{-1}\mathbf{X} \otimes \mathbf{A}\right). \end{aligned} \tag{5.19}$$

Then $\mathbf{M}\delta_x = \mathbf{0}$ if and only if

$$\mathbf{T}\delta_x = \mathbf{0}. \tag{5.20}$$

We can split δ_x as $\delta_x = \mathbf{P}_\Omega\delta_x + \mathbf{P}_{\Omega^c}\delta_x$. Multiply (5.20) on the left by $\mathbf{P}_\Omega$ and we get

$$\mathbf{P}_\Omega\mathbf{T}\mathbf{P}_\Omega\delta_x = -\mathbf{P}_\Omega\mathbf{T}\mathbf{P}_{\Omega^c}\delta_x, \tag{5.21}$$

or equivalently,

$$[\mathbf{P}_\Omega\mathbf{T}\mathbf{P}_\Omega]\,(\mathbf{P}_\Omega\delta_x) = -\,[\mathbf{P}_\Omega\mathbf{T}\mathbf{P}_{\Omega^c}]\,(\mathbf{P}_{\Omega^c}\delta_x). \tag{5.22}$$

Let $S_\Omega \subset \mathbb{R}^{np}$ denote the solution space of $\mathbf{P}_\Omega\mathbf{z} = \mathbf{z}$, and define

$$\xi \;\doteq\; \inf_{\mathbf{z}\in S_\Omega\setminus\{\mathbf{0}\}} \frac{\|\mathbf{P}_\Omega\mathbf{T}\mathbf{P}_\Omega\mathbf{z}\|}{\|\mathbf{z}\|}. \tag{5.23}$$

Then if $\xi > 0$, by (5.22) we have

$$\|\mathbf{P}_\Omega\delta_x\| \;\le \xi^{-1}\,\|\mathbf{P}_\Omega\mathbf{T}\mathbf{P}_\Omega\delta_x\| \;=\; \xi^{-1}\,\|[\mathbf{P}_\Omega\mathbf{T}\mathbf{P}_{\Omega^c}]\,\mathbf{P}_{\Omega^c}\delta_x\| \tag{5.24}$$

[1] Indeed, it is easy to see $\mathbf{C}_A \in \mathbb{R}^{mn\times n}$ is a block diagonal matrix whose blocks are the columns of $\mathbf{A}$.

$$\leq \xi^{-1} \left\| \mathbf{P}_\Omega \mathbf{T} \mathbf{P}_{\Omega^c} \right\| \left\| \mathbf{P}_{\Omega^c} \delta_x \right\| . \tag{5.25}$$

We will show that

$$\|\mathbf{P}_\Omega \mathbf{T} \mathbf{P}_{\Omega^c}\| \ \leq \ C\|\mathbf{A}\|, \tag{5.26}$$

for some positive constant C. Indeed, $\|\mathbf{P}_\Omega \mathbf{T} \mathbf{P}_{\Omega^c}\| \leq \|\mathbf{P}_\Omega \mathbf{T}\|\|\mathbf{P}_{\Omega^c}\| = \|\mathbf{P}_\Omega \mathbf{T}\|$, and

$$\|\mathbf{P}_\Omega \mathbf{T}\| \tag{5.27}$$

$$\leq \|\mathbf{P}_\Omega(\mathbf{I} \otimes \mathbf{A}^*)\| \left\|(\mathbf{I} - \mathbf{P}_X) \otimes \mathbf{A} + (\mathbf{X}^*(\mathbf{X}\mathbf{X}^*)^{-1} \otimes \mathbf{I})\mathbf{C}_A \mathbf{C}_A^*((\mathbf{X}\mathbf{X}^*)^{-1}\mathbf{X} \otimes \mathbf{A})\right\| \tag{5.28}$$

$$\leq \|\mathbf{P}_\Omega(\mathbf{I} \otimes \mathbf{A}^*)\|\|(\mathbf{I} - \mathbf{P}_X) \otimes \mathbf{I} + (\mathbf{X}^*(\mathbf{X}\mathbf{X}^*)^{-1} \otimes \mathbf{I})\mathbf{C}_A \mathbf{C}_A^*((\mathbf{X}\mathbf{X}^*)^{-1}\mathbf{X} \otimes \mathbf{I})\|\|\mathbf{A}\| \tag{5.29}$$

$$\leq \|\mathbf{P}_\Omega(\mathbf{I} \otimes \mathbf{A}^*)\|(1 + 1/\lambda_{min}(\mathbf{X}\mathbf{X}^*))\|\mathbf{A}\|. \tag{5.30}$$

Note $\mathbf{P}_\Omega(\mathbf{I}\otimes\mathbf{A}^*)$ is a block-diagonal matrix, with blocks given by $\mathbf{A}^*_{\Omega_1} \dots \mathbf{A}^*_{\Omega_p}$. By the matrix operator norm bound by incoherence (see Appendix A), the operator norm of each of these blocks is upper bounded by a constant. Hence $\|\mathbf{P}_\Omega(\mathbf{I} \otimes \mathbf{A}^*)\|$ is bounded by the same constant. On $\mathcal{E}_{eig}$, $\lambda^{-1}_{min}(\mathbf{X}X^*)$ is also bounded by a constant, and thus (5.26) holds.

From (5.26), (5.25) and (5.26) we have

$$\|\mathbf{\Delta}_A\|_F \leq \sqrt{2}\|\mathbf{A}\|\|\mathbf{\Delta}_X\|_F \ \leq \ \sqrt{2}\|\mathbf{A}\| \left(\|\mathbf{P}_\Omega \mathbf{\Delta}_X\|_F + \|\mathcal{P}_{\Omega^c}\mathbf{\Delta}_X\|_F\right) \tag{5.31}$$

$$\leq \sqrt{2}\|\mathbf{A}\| \left(1 + C\xi^{-1}\|\mathbf{A}\|\right) \|\mathcal{P}_{\Omega^c}\mathbf{\Delta}_X\|_F. \tag{5.32}$$

So our only remaining task is to lower bound ξ in (5.32) to complete the bound on α. Specifically, in the following we will show that ξ is lower bounded by a positive constant with high probability.

Notice that as $p \to \infty$, $\mathbf{X}\mathbf{X}^* \to \mathbf{I}$ almost surely. We can replace $(\mathbf{X}\mathbf{X}^*)^{-1}$ with $\mathbf{I}$ in (5.19) to get a simplified approximation of $\mathbf{T}$ given by

$$\hat{\mathbf{T}} \ \doteq \ (\mathbf{I} - \mathbf{X}^*\mathbf{X}) \otimes \mathbf{A}^*\mathbf{A} + (\mathbf{X}^* \otimes \mathbf{A}^*)\, \mathbf{C}_A \mathbf{C}_A^* \,(\mathbf{X} \otimes \mathbf{A}) \tag{5.33}$$

$$= \ \mathbf{I} \otimes \mathbf{A}^*\mathbf{A} - (\mathbf{X}^* \otimes \mathbf{A}^*)(\mathbf{I} - \mathbf{C}_A \mathbf{C}_A^*)(\mathbf{X} \otimes \mathbf{A}). \tag{5.34}$$

Let $\mathbf{R}$ denote the second term in (5.34)

$$\mathbf{R} \ \doteq \ (\mathbf{X}^* \otimes \mathbf{A}^*)(\mathbf{I} - \mathbf{C}_A \mathbf{C}_A^*)(\mathbf{X} \otimes \mathbf{A}). \tag{5.35}$$

Then $\hat{\mathbf{T}} = \mathbf{I} \otimes \mathbf{A}^*\mathbf{A} - \mathbf{R}$, and

$$\mathbf{T} = \mathbf{I} \otimes \mathbf{A}^*\mathbf{A} - \mathbf{R} + (\mathbf{T} - \hat{\mathbf{T}}). \tag{5.36}$$

In terms of $\mathbf{T}, \hat{\mathbf{T}}$ and $\mathbf{R}$,

$$\begin{aligned} \xi &= \inf_{\mathbf{z} \in S_\Omega \setminus \{\mathbf{0}\}} \left\{ \frac{\|\mathbf{P}_\Omega(\mathbf{I} \otimes \mathbf{A}^*\mathbf{A} - \mathbf{R} + \mathbf{T} - \hat{\mathbf{T}})\mathbf{P}_\Omega \mathbf{z}\|}{\|\mathbf{z}\|} \right\} \\ &\geq \inf_{\mathbf{z} \in S_\Omega \setminus \{\mathbf{0}\}} \left\{ \frac{\|\mathbf{P}_\Omega(\mathbf{I} \otimes \mathbf{A}^*\mathbf{A})\mathbf{P}_\Omega \mathbf{z}\|}{\|\mathbf{z}\|} \right\} - \|\mathbf{P}_\Omega \mathbf{R} \mathbf{P}_\Omega\| - \|\mathbf{P}_\Omega(\mathbf{T} - \hat{\mathbf{T}})\mathbf{P}_\Omega\|. \end{aligned} \tag{5.37}$$

In the following we will bound the three terms in (5.37) separately. In particular, we will show

$$\inf_{\mathbf{z} \in S_\Omega \setminus \{\mathbf{0}\}} \left\{ \frac{\|\mathbf{P}_\Omega(\mathbf{I} \otimes \mathbf{A}^*\mathbf{A})\mathbf{P}_\Omega \mathbf{z}\|}{\|\mathbf{z}\|} \right\} \ \geq\ 1 - k\mu(\mathbf{A}), \tag{5.38}$$

and there is a constant $t_\star > 0$ such that on $\mathcal{E}_{eig}(t_\star)$,

$$\|\mathbf{P}_\Omega(\mathbf{T} - \hat{\mathbf{T}})\mathbf{P}_\Omega\| \ \leq\ 1/8. \tag{5.39}$$

The analysis of the middle term $\mathbf{P}_\Omega \mathbf{R} \mathbf{P}_\Omega$ is a little technical, requiring both additional algebraic manipulations and additional probability estimates. Now define an event $\mathcal{E}_R$, on which the norm of $\mathbf{P}_\Omega \mathbf{R} \mathbf{P}_\Omega$ is small:

$$\mathcal{E}_R \ \doteq\ \{\omega \mid \|\mathbf{P}_\Omega \mathbf{R} \mathbf{P}_\Omega\| \leq 1/8\}. \tag{5.40}$$

In Appendix E, we prove the following Lemma 5.3, which shows that the probability of $\mathcal{E}_R$ is indeed close to one. More precisely,

Lemma 5.3. *Let $\mathcal{E}_R$ be the event that $\|\mathbf{P}_\Omega \mathbf{R} \mathbf{P}_\Omega\| \leq 1/8$. Then there exist positive numerical constants C_1, C_2, C_3, C_4, C_5 and C_6 such that whenever*

$$k \ \leq\ \min\left\{ C_1 n, \frac{C_2}{\mu(\mathbf{A})} \right\} \tag{5.41}$$

and $p > C_3\, n^2$ we have

$$\mathbb{P}\left[\mathcal{E}_R\right] \ \geq\ 1 - C_4 p^{-4} - C_5 n^2 \exp\left(-C_6 k^2 p / n^2\right). \tag{5.42}$$

Combining (5.38), (5.39) and (5.40), and further assuming $k\mu(\mathbf{A}) < 1/2$ (by choosing an appropriately small constant C_1 in the statement of Theorem 5.1), on $\mathcal{E}_{eig}(t_\star) \cap \mathcal{E}_R$, we have

$$\xi \;\geq\; \frac{3}{4} - k\mu(\mathbf{A}) \;>\; \frac{1}{4}. \tag{5.43}$$

Therefore, on $\mathcal{E}_{eig}(t_\star) \cap \mathcal{E}_R$,

$$\begin{aligned} \|\mathbf{\Delta}_A\|_F \;&\leq\; \sqrt{2}\|\mathbf{A}\| \left(1 + C\xi^{-1}\|\mathbf{A}\|\right) \|\mathcal{P}_{\Omega^c}\mathbf{\Delta}_X\|_F & (5.44)\\ &\leq\; \sqrt{2}\|\mathbf{A}\|(1 + C'\|\mathbf{A}\|)\|\mathcal{P}_{\Omega^c}[\mathbf{\Delta}_X]\|_F & (5.45)\\ &\leq\; C''\,\|\mathbf{A}\|^2\|\mathcal{P}_{\Omega^c}[\mathbf{\Delta}_X]\|_F, & (5.46)\end{aligned}$$

for some positive constants C' and C''.

Therefore, the bound in Theorem 5.1 holds with probability at least $1 - \mathbb{P}[\mathcal{E}_{eig}(t_\star)^c] - \mathbb{P}[\mathcal{E}_R^c]$. In Lemma 5.2 we have shown

$$\mathbb{P}[\mathcal{E}_{eig}(t_\star)^c] \;<\; c_1 n \exp(-c_2 p/n \log(p)) + p^{-7}, \tag{5.47}$$

and in Lemma 5.3 we prove

$$\mathbb{P}[\mathcal{E}_R^c] \;<\; c_3 p^{-4} + c_4 n^2 \exp(-c_5 k^2 p/n^2). \tag{5.48}$$

Combining these two probability bounds yields the desired result in Theorem 5.1.

All that remains to show is that (5.38) and (5.39) indeed hold, and we prove them in the following.

(i) Establishing (5.38) Fix any $\mathbf{Z} \in \mathbb{R}^{n\times p}$ with support set contained in Ω and write $\mathbf{z} \doteq \mathrm{vec}[\mathbf{Z}]$. Then

$$\begin{aligned} \|\mathbf{P}_\Omega(\mathbf{I}\otimes\mathbf{A}^*\mathbf{A})\mathbf{P}_\Omega\mathbf{z}\|^2 \;&=\; \|\mathcal{P}_\Omega\,[\;\mathbf{A}^*\mathbf{A}\;\mathcal{P}_\Omega[\mathbf{Z}]\;]\|_F^2 & (5.49)\\ &=\; \sum_{j=1}^{p} \left\|\mathbf{A}_{\Omega_j}^*\mathbf{A}_{\Omega_j}\mathbf{Z}(\Omega_j, j)\right\|_2^2 & (5.50)\\ &\geq\; \sigma_{min}^2(\mathbf{A}_{\Omega_j}^*\mathbf{A}_{\Omega_j})\sum_j \|\mathbf{Z}(\Omega_j, j)\|_2^2 & (5.51)\\ &\geq\; \|\mathbf{Z}\|_F^2(1 - k\mu(\mathbf{A}))^2, & (5.52)\end{aligned}$$

where in (5.51) we used that $\mathbf{Z}$ is supported on Ω and in (5.52) we used the bound $\sigma_{min}(\mathbf{A}_{\Omega_j}^*\mathbf{A}_{\Omega_j}) > 1-k\mu(\mathbf{A})$ (see Appendix A). Since this bound holds for all $z \in S_\Omega \setminus \{\mathbf{0}\}$, (5.38) holds.

(ii) Establishing (5.39) Write $\mathbf{\Xi} \doteq (\mathbf{X}\mathbf{X}^*)^{-1} - \mathbf{I}$, and then $\mathbf{T} - \hat{\mathbf{T}}$ can be written as

$$
\begin{aligned}
\mathbf{T} - \hat{\mathbf{T}} &= \mathbf{X}^*\mathbf{X} \otimes \mathbf{A}^*\mathbf{A} - \big(\mathbf{X}^*(\mathbf{X}\mathbf{X}^*)^{-1}\mathbf{X}\big) \otimes \mathbf{A}^*\mathbf{A} \\
&\quad + \big(\mathbf{X}^*(\mathbf{X}\mathbf{X}^*)^{-1} \otimes \mathbf{A}^*\big)\, \mathbf{C}_A\mathbf{C}_A^* \, \big((\mathbf{X}\mathbf{X}^*)^{-1}\mathbf{X} \otimes \mathbf{A}\big) \\
&\quad - (\mathbf{X}^* \otimes \mathbf{A}^*)\, \mathbf{C}_A\mathbf{C}_A^* \, (\mathbf{X} \otimes \mathbf{A}) \\
&= (\mathbf{X}^* \otimes \mathbf{A}^*)(-\mathbf{\Xi} \otimes \mathbf{I})(\mathbf{X} \otimes \mathbf{A}) \\
&\quad + \big(\mathbf{X}^*(\mathbf{X}\mathbf{X}^*)^{-1} \otimes \mathbf{A}^*\big)\, \mathbf{C}_A\mathbf{C}_A^* \, \big((\mathbf{X}\mathbf{X}^*)^{-1}\mathbf{X} \otimes \mathbf{A}\big) \\
&\quad - \ (\mathbf{X}^* \otimes \mathbf{A}^*)\, \mathbf{C}_A\mathbf{C}_A^* \, \big((\mathbf{X}\mathbf{X}^*)^{-1}\mathbf{X} \otimes \mathbf{A}\big) \\
&\quad + \ (\mathbf{X}^* \otimes \mathbf{A}^*)\, \mathbf{C}_A\mathbf{C}_A^* \, \big((\mathbf{X}\mathbf{X}^*)^{-1}\mathbf{X} \otimes \mathbf{A}\big) \\
&\quad - (\mathbf{X}^* \otimes \mathbf{A}^*)\, \mathbf{C}_A\mathbf{C}_A^* \, (\mathbf{X} \otimes \mathbf{A}) \\
&= (\mathbf{X}^* \otimes \mathbf{A}^*)(-\mathbf{\Xi} \otimes \mathbf{I})(\mathbf{X} \otimes \mathbf{A}) \\
&\quad + (\mathbf{X}^* \otimes \mathbf{A}^*)(\mathbf{\Xi} \otimes \mathbf{I})\mathbf{C}_A\mathbf{C}_A^* \big((\mathbf{X}\mathbf{X}^*)^{-1}\mathbf{X} \otimes \mathbf{A}\big) \\
&\quad + (\mathbf{X}^* \otimes \mathbf{A}^*)\mathbf{C}_A\mathbf{C}_A^*(\mathbf{\Xi} \otimes \mathbf{I})(\mathbf{X} \otimes \mathbf{A}) \\
&= (\mathbf{X}^* \otimes \mathbf{A}^*)\Big((\mathbf{C}_A\mathbf{C}_A^* - \mathbf{I})\mathbf{\Xi} \otimes \mathbf{I} + (\mathbf{\Xi} \otimes \mathbf{I})\mathbf{C}_A\mathbf{C}_A^*(\mathbf{X}\mathbf{X}^*)^{-1}\Big)(\mathbf{X} \otimes \mathbf{A}).
\end{aligned}
$$

Using the facts that $\|\mathbf{C}_A\mathbf{C}_A^* - \mathbf{I}\| = 1$[2] and $\|\mathbf{C}_A\| = 1$, we have the estimate

$$
\begin{aligned}
\|\mathbf{P}_\Omega(\mathbf{T} - \hat{\mathbf{T}})\mathbf{P}_\Omega\| &\leq \|\mathbf{P}_\Omega(\mathbf{X}^* \otimes \mathbf{A})\|^2 \times \big(\|\mathbf{\Xi}\| + \|\mathbf{X}\mathbf{X}^*\|^{-1}\|\mathbf{\Xi}\|\big) \\
&\leq \|\mathbf{P}_\Omega(\mathbf{I} \otimes \mathbf{A})\|^2\|\mathbf{X}\mathbf{X}^*\| \left(1 + \|(\mathbf{X}\mathbf{X}^*)^{-1}\|\right) \|\mathbf{\Xi}\| \\
&\leq 6 \times \|\mathbf{P}_\Omega(\mathbf{I} \otimes \mathbf{A})\|^2\|\mathbf{\Xi}\|, \qquad (5.53)
\end{aligned}
$$

where the last bound holds on $\mathcal{E}_{eig}(t)$ for small enough t (say, $t < 1/2$ is sufficient). From the matrix norm bound by incoherence of $\mathbf{A}$ (see Appendix A),

$$
\|\mathbf{P}_\Omega(\mathbf{I} \otimes \mathbf{A})\|^2 \ = \ \max_j \|\mathbf{A}_{\Omega_j}\|^2 \ \leq \ 1 + k\mu(\mathbf{A}) < 2. \qquad (5.54)
$$

Hence, on $\mathcal{E}_{eig}(t)$, $\|\mathbf{P}_\Omega(\mathbf{T} - \hat{\mathbf{T}})\mathbf{P}_\Omega\| \leq 12\|\mathbf{\Xi}\|$. In addition, on $\mathcal{E}_{eig}(t)$, $\|\mathbf{\Xi}\| \leq t/(1-t)$. By choosing t small enough (say, $t < 1/97$) we can guarantees that

[2]Since $\mathbf{C}_A^*\mathbf{C}_A = \mathbf{I}$, $\|\mathbf{C}_A\mathbf{C}_A^* - \mathbf{I}\|^2 = \|(\mathbf{C}_A\mathbf{C}_A^* - \mathbf{I})(\mathbf{C}_A\mathbf{C}_A^* - \mathbf{I})^*\| = \|(\mathbf{C}_A\mathbf{C}_A^* - \mathbf{I})\|$, and thus $\|\mathbf{C}_A\mathbf{C}_A^* - \mathbf{I}\|^2 = 1$.

on $\mathcal{E}_{eig}(t)$, $\|\mathbf{P}_\Omega(\mathbf{T} - \hat{\mathbf{T}})\mathbf{P}_\Omega\| \leq 1/8$, which establishes the bound in (5.39).

Thus, we have proved (5.38) and (5.39) hold, and Theorem 5.1 is established.

$\square$

CHAPTER 6

CONCLUSION

The idea that many classes of signals can be represented by linear combination of a small set of atoms of a dictionary has had a great impact on various signal processing applications, e.g., image compression, super resolution imaging and robust face recognition [10]. However, for practical problems such a sparsifying dictionary is usually unknown ahead of time. In this thesis we study an ℓ_1-minimization approach to the dictionary learning problem. We prove that that under mild conditions, the dictionary learning problem is locally well-posed, i.e., the desired solution is indeed a local optimum, and thus a local algorithm can hope to recover the sparsifying dictionary. Intriguingly, the simulation results even suggest global optimality: When the problem is well-structured, from any random initial point, a local algorithm always converges to the desired solution up to sign and permutation ambiguity.

To fully understand the dictionary learning problem is a long-term goal and there are many interesting open problems. While we have proved the local correctness result, simulation results even suggest global optimality. We conjecture that when the problem is well-structured ($\mathbf{X}$ is sufficiently sparse), the desired solution $(\mathbf{A}, \mathbf{X})$ is indeed the unique global optimal point to (2.8) up to sign and permutation ambiguity. To establish this global correctness result will require new ideas and tools. Another interesting open problem involves with the uniqueness of sparse matrix factorization. Given $\mathbf{Y} = \mathbf{AX}$ as a product of a overcomplete dictionary $\mathbf{A} \in \mathbb{R}^{m \times n}(m < n)$ and a column-wise sparse matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, we want to determine under what conditions in terms of number of samples p, $\mathbf{Y}$ has a unique sparse matrix factorization, again up to sign and permutation ambiguity. This problem was proposed and studied in [11], which establishes a partial result on the uniqueness of overcomplete dictionary when we have $(k+1)\binom{n}{k}$ samples, in which for each k-dimension subspace of $\mathbf{A}$ there are exactly $(k+1)$ samples. However, under the random sparsity model in Section 2.1, it is still an open problem to determine how many samples are needed to guarantee the uniqueness of the overcomplete dictionary. Another interesting direction is "robust dictionary learning," where samples are corrupted by certain noises. We believe the proposed ℓ_1-minimization approach (2.8) (or slightly modified one) is also robust against noise. The techniques we used in this work may be still useful to attack these problems.

APPENDIX A

MATRIX NORM BOUND BY MUTUAL COHERENCE

In this section, we give some bounds on the operator norm of matrix $\mathbf{A}$ based on the mutual coherence of $\mathbf{A}$ [12]. Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ be a matrix with unit norm columns. Recall that mutual coherence $\mu(\mathbf{A})$ is defined as

$$\mu(\mathbf{A}) = \max_{i \neq j} |\langle \mathbf{A}_i, \mathbf{A}_j \rangle|. \tag{A.1}$$

Set $\mathbf{\Delta} = \mathbf{A}^*\mathbf{A} - \mathbf{I}$. Then we can use the mutual coherence to bound the operator norm of $\mathbf{A}$ by

$$\|\mathbf{A}\|^2 = \|\mathbf{A}^*\mathbf{A}\| = \|\mathbf{I}+\mathbf{\Delta}\| \leq 1+\|\mathbf{\Delta}\| \leq 1+\|\mathbf{\Delta}\|_F \leq 1+n\|\mathbf{\Delta}\|_\infty = 1+n\mu(\mathbf{A}). \tag{A.2}$$

Further, we can get a tighter bound for the submatrices of $\mathbf{A}$. Let $L \in \binom{[n]}{k}$. Then the same argument works for $\mathbf{A}_L$:

$$\|\mathbf{A}_L\|^2 = \|\mathbf{A}_L^*\mathbf{A}_L\| \leq 1 + k\mu(\mathbf{A}). \tag{A.3}$$

Similarly, we can bound the smallest eigenvalue of $\mathbf{A}_L^*\mathbf{A}_L$ by

$$\lambda_{min}(\mathbf{A}_L^*\mathbf{A}_L) = \lambda_{min}(I+\mathbf{A}_L^*\mathbf{A}_L-I) \geq 1-\|\mathbf{A}_L^*\mathbf{A}_L-I\|_F \geq 1-k\mu(\mathbf{A}). \tag{A.4}$$

If we assume that $k\mu(\mathbf{A}) < 1/2$, then

$$\|(\mathbf{A}_L^*\mathbf{A}_L)^{-1}\| \leq 2. \tag{A.5}$$

We can get a tighter bound for the operator norm of $(\mathbf{A}_L^*\mathbf{A}_L)^{-1}$ using the Neumann series representation of the inverse. Write $\mathbf{A}_L^*\mathbf{A}_L = \mathbf{I} + \mathbf{H}$. Then $\|\mathbf{H}\|_F < k\mu(\mathbf{A})$. Using the fact that

$$(\mathbf{A}_L^*\mathbf{A}_L)^{-1} = \sum_{t=0}^{\infty} (-1)^t \mathbf{H}^t, \tag{A.6}$$

we have

$$\begin{aligned}
\|(\mathbf{A}_L^*\mathbf{A}_L)^{-1} - \mathbf{I}\|_F &= \|\sum_{t=1}^{\infty}(-\mathbf{H})^t\|_F && \text{(A.7)}\\
&\le \sum_{t=1}^{\infty}\|-\mathbf{H}\|_F^t \;\le\; k\mu(\mathbf{A})/(1-k\mu(\mathbf{A})) && \text{(A.8)}\\
&< 2k\mu(\mathbf{A}). && \text{(A.9)}
\end{aligned}$$

APPENDIX B

MATRIX CHERNOFF BOUND AND KAHANE-KHINTCHINE INEQUALITY

In this section, we quote two technical tools used in the proof of Theorem 5.1. The first one is the matrix Chernoff bound of Tropp [9], which builds on ideas introduced by Ahlswede and Winter [13] to bound the eigenvalues of the sum of independent random positive semidefinite matrices.

Theorem B.1 (Matrix Chernoff Bound, [9] Theorem 2.5)**.** *Let $\mathbf{M}_1, \dots, \mathbf{M}_n$ be a finite sequence of independent random positive-semidefinite matrices of dimension d. Suppose that for each $\mathbf{M}_i$, $\lambda_{max}(\mathbf{M}_i) \le B$ almost surely. Set $\mu_{min} = \lambda_{min}(\sum_i \mathbb{E}[\mathbf{M}_i])$ and $\mu_{max} = \lambda_{max}(\sum_i \mathbb{E}[\mathbf{M}_i])$. Then the following two bounds hold:*

$$\mathbb{P}\Big[\lambda_{min}\Big(\sum_i \mathbf{M}_i\Big) \le t\mu_{min}\Big] \le d\, \exp\big(-(1-t)^2\mu_{min}/2B\big), \quad \forall\, t \in [0,1), \tag{B.1}$$

$$\mathbb{P}\Big[\lambda_{max}\Big(\sum_i \mathbf{M}_i\Big) \ge (1+t)\mu_{max}\Big] \le d\, \left(\frac{e^t}{(1+t)^{1+t}}\right)^{\mu_{max}/B}, \quad \forall\, t \ge 0. \tag{B.2}$$

Two simplifications of (B.2) are useful:

$$\mathbb{P}\Big[\ \Big\|\sum_i \mathbf{M}_i\Big\|\ \ge\ (1+t)\,\mu_{max}\ \Big] \le d \exp\big(-t^2\mu_{max}/4B\big), \quad \forall\, t \in [0,1], \tag{B.3}$$

and
$$\mathbb{P}\Big[\ \Big\|\sum_i \mathbf{M}_i\Big\|\ \ge\ t\,\mu_{max}\ \Big] \le d\, \left(\frac{e}{t}\right)^{t\mu_{max}/B}, \quad \forall\, t > e. \tag{B.4}$$

The second is given in [9], while the first follows from (B.2) and the inequality $t - (1+t)\log(1+t) \le -t^2/4$, which holds on $[0,1]$.

The second tool we quote here is the classical Kahane-Khintchine inequality, with constant $1/\sqrt{2}$ found by Latala and Oleszkiewicz [14]:

Theorem B.2 (Kahane-Khintchine Inequality [15], [14] Theorem 1)**.** *Let $\sigma_1, \ldots, \sigma_n$ be an i.i.d. sequence of Rademacher random variables (i.e., variables that take on ± 1 with equal probability), and let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a fixed sequence of vectors in a normed space V. Then*

$$\frac{1}{\sqrt{2}} \left(\mathbb{E}\left[\left\|\sum_i \sigma_i \mathbf{x}_i\right\|_V^2\right]\right)^{1/2} \;\leq\; \mathbb{E}\left[\left\|\sum_i \sigma_i \mathbf{x}_i\right\|_V\right] \;\leq\; \left(\mathbb{E}\left[\left\|\sum_i \sigma_i \mathbf{x}_i\right\|_V^2\right]\right)^{1/2}. \tag{B.5}$$

We use the Kahane-Khintchine inequality to prove Corollary B.3, which is used in Section 4.2.2 to prove Theorem 4.1.

Corollary B.3. *Let $\mathbf{M} \in \mathbb{R}^{m\times n}$ be any fixed matrix, and $\mathbf{v} \in \mathbb{R}^n$ with each component being an i.i.d. $\mathcal{N}(0, \sigma^2)$ random variable. Then*

$$\frac{\sigma}{\sqrt{\pi}} \|\mathbf{M}\|_F \;\leq\; \mathbb{E}\left[\|\mathbf{M}\mathbf{v}\|_2\right] \;\leq\; \sigma \|\mathbf{M}\|_F. \tag{B.6}$$

Proof. For the right side, we simply use Jensen's inequality:

$$\begin{aligned}
\mathbb{E}\left[\|Mv\|_2\right] &\leq \left(\mathbb{E}\left[\|Mv\|_2^2\right]\right)^{1/2} && \text{(B.7)}\\
&= \left(\mathbb{E}\left[\sum_i \Big(\sum_j M_{ij} v_j\Big)^2\right]\right)^{1/2} && \text{(B.8)}\\
&= \left(\mathbb{E}\left[\sum_{i,j,j'} M_{ij} M_{ij'} v_j v_{j'}\right]\right)^{1/2} && \text{(B.9)}\\
&= \left(\sum_{ij} M_{ij}^2 \mathbb{E}[v_j^2]\right)^{1/2} \quad = \quad \sigma \|M\|_F. && \text{(B.10)}
\end{aligned}$$

For the left side, we write $v_j = \varepsilon_j \nu_j$, where ν_j is a nonnegative random variable with the same distribution as $|v_j|$, and ε_j is an independent Rademacher random variable. Then

$$\begin{aligned}
\mathbb{E}\left[\|Mv\|_2\right] &= \mathbb{E}\left[\left\|\sum_i M_i v_i\right\|_2\right] && \text{(B.11)}\\
&= \mathbb{E}_\varepsilon \mathbb{E}_\nu \left[\left\|\sum_i M_i \varepsilon_i \nu_i\right\|_2\right] && \text{(B.12)}\\
&\geq \mathbb{E}_\varepsilon \left[\left\|\mathbb{E}_\nu \left[\sum_i M_i \varepsilon_i \nu_i\right]\right\|_2\right] && \text{(B.13)}
\end{aligned}$$

$$= \sigma\sqrt{\frac{2}{\pi}}\mathbb{E}_\varepsilon\left[\left\|\sum_i M_i\varepsilon_i\right\|_2\right] \tag{B.14}$$

$$\geq \sigma\sqrt{\frac{1}{\pi}}\left(\mathbb{E}_\varepsilon\left[\left\|\sum_i M_i\varepsilon_i\right\|_2^2\right]\right)^{1/2}, \tag{B.15}$$

where in (B.13) we have used Jensen's inequality, and we have applied the Kahane-Khintchine inequality in (B.15) . Since $\mathbb{E}_\varepsilon\left[\|\sum_i M_i\varepsilon_i\|^2\right] = \|M\|_F^2$, we have

$$\frac{\sigma}{\sqrt{\pi}}\|\mathbf{M}\|_F \;\leq\; \mathbb{E}\left[\|\mathbf{M}\mathbf{v}\|_2\right]. \tag{B.16}$$

$\square$

APPENDIX C

A DECOUPLING LEMMA

In Lemma C.1, we establish an upper bound of the expected norm of $\mathbf{P}_\Omega \mathbf{M} \mathbf{P}_\Omega$ in which $\mathbf{M}$ is a matrix with no diagonal elements. The proof is an application of a well-known decoupling technique [16] and several steps are quite similar to the argument in the proof of Proposition 2.1 of [17].

Lemma C.1. *Fix $\mathbf{M} \in \mathbb{R}^{n \times n}$ with all diagonal elements equal to zero. Let $\Omega \sim \mathrm{uni}\binom{[n]}{k}$ be a uniform random subset of size k. Then the following estimate holds:*

$$\mathbb{E}\left[\|\mathbf{P}_\Omega \mathbf{M} \mathbf{P}_\Omega\|_F\right] \;\le\; 16\sqrt{\frac{k}{n}}\,\mathbb{E}\left[\|\mathbf{M}\mathbf{P}_\Omega\|_F\right]. \tag{C.1}$$

Proof. Let $\mathbf{\Lambda}$ be a diagonal matrix whose entries are i.i.d. Bernoulli random variables taking on value 1 with probability k/n . Let k' be the trace of $\mathbf{\Lambda}$, i.e., the number of nonzeros in $\mathbf{\Lambda}$, and thus k' is a binomial random variable. Then

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{\Lambda M \Lambda}\|_F\right] &= \sum_{s=0}^{n} \mathbb{P}[k'=s]\,\mathbb{E}\left[\|\mathbf{\Lambda M \Lambda}\|_F \mid k'=s\right], &\text{(C.2)}\\
&\ge \sum_{s=k}^{n} \mathbb{P}[k'=s]\,\mathbb{E}\left[\|\mathbf{\Lambda M \Lambda}\|_F \mid k'=s\right]. &\text{(C.3)}
\end{aligned}$$

Conditioned on $k'=s$, the nonzero entries on the diagonal of $\mathbf{\Lambda}$ are uniformly distributed on $\binom{[n]}{s}$. Furthermore, note that if $\Omega \subset \mathrm{support}(\mathrm{diag}(\mathbf{\Lambda}))$, $\|\mathbf{P}_\Omega \mathbf{M} \mathbf{P}_\Omega\|_F \le \|\mathbf{\Lambda M \Lambda}\|_F$. Hence,

$$\forall\, s \ge k, \quad \mathbb{E}\left[\|\mathbf{\Lambda M \Lambda}\|_F \mid k'=s\right] \;\ge\; \mathbb{E}\left[\|\mathbf{P}_\Omega \mathbf{M} \mathbf{P}_\Omega\|_F\right]. \tag{C.4}$$

Plugging into (C.3), and using the fact that k is a median of the binomial random variable k', we have

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{\Lambda M \Lambda}\|_F\right] &\geq \sum_{s=k}^{n} \mathbb{P}[k'=s]\,\mathbb{E}\left[\|\mathbf{P}_\Omega \mathbf{M} \mathbf{P}_\Omega\|_F\right], && \text{(C.5)}\\
&= \mathbb{P}[k' \geq k]\,\mathbb{E}\left[\|\mathbf{P}_\Omega \mathbf{M} \mathbf{P}_\Omega\|_F\right], && \text{(C.6)}\\
&\geq \frac{1}{2}\,\mathbb{E}\left[\|\mathbf{P}_\Omega \mathbf{M} \mathbf{P}_\Omega\|_F\right]. && \text{(C.7)}
\end{aligned}$$

Therefore,

$$\mathbb{E}\left[\|\mathbf{P}_\Omega \mathbf{M} \mathbf{P}_\Omega\|_F\right] \leq 2\,\mathbb{E}\left[\|\mathbf{\Lambda M \Lambda}\|_F\right]. \tag{C.8}$$

Similar to [17], for each i, j, let $\mathbf{M}_{ij} \in \mathbb{R}^{n\times n}$ be a matrix of which the (i,j) entry is equal to the (i,j) entry of $\mathbf{M}$ and all other entries are zero. So we can write $\mathbb{E}\left[\|\mathbf{\Lambda M \Lambda}\|_F\right]$ as

$$\mathbb{E}\left[\|\mathbf{\Lambda M \Lambda}\|_F\right] = \mathbb{E}\Big[\Big\|\sum_{i>j} \lambda_i \lambda_j (\mathbf{M}_{ij} + \mathbf{M}_{ji})\Big\|_F\Big]. \tag{C.9}$$

Let $\eta_1, \eta_2, \ldots, \eta_n$ be a sequence of independent Bernoulli random variables, each taking on value 1 with probability 1/2. Then we have

$$\begin{aligned}
&\mathbb{E}\Big[\Big\|\sum_{i>j} \lambda_i \lambda_j (\mathbf{M}_{ij} + \mathbf{M}_{ji})\Big\|_F\Big] && \text{(C.10)}\\
&= 2\,\mathbb{E}_{\mathbf{\Lambda}}\Big[\Big\|\mathbb{E}_\eta\Big[\sum_{i>j}\Big(\eta_i(1-\eta_j) + \eta_j(1-\eta_i)\Big)\lambda_i\lambda_j(\mathbf{M}_{ij} + \mathbf{M}_{ji})\Big]\Big\|_F\Big] && \text{(C.11)}\\
&\leq 2\,\mathbb{E}_{\mathbf{\Lambda}}\mathbb{E}_\eta\Big[\Big\|\sum_{i>j}\Big(\eta_i(1-\eta_j) + \eta_j(1-\eta_i)\Big)\lambda_i\lambda_j(\mathbf{M}_{ij} + \mathbf{M}_{ji})\Big\|_F\Big] && \text{(C.12)}\\
&= 2\,\mathbb{E}_\eta\mathbb{E}_{\mathbf{\Lambda}}\Big[\Big\|\sum_{i>j}\Big(\eta_i(1-\eta_j) + \eta_j(1-\eta_i)\Big)\lambda_i\lambda_j(\mathbf{M}_{ij} + \mathbf{M}_{ji})\Big\|_F\Big], && \text{(C.13)}
\end{aligned}$$

where in (C.12) we used Jensen's inequality to pull the expectation out of the norm. So there must exist at least one sequence $\eta^\star$ such that the right hand side of (C.13) is larger than or equal to its expectation over η. Let $T \subset [n]$ be the support of $\eta^\star$, and let T^c be its complement. Then combining (C.8) and (C.13), we have

$$\begin{aligned}
&\mathbb{E}\Big[\|\mathbf{P}_\Omega \mathbf{M} \mathbf{P}_\Omega\|_F\Big]\\
&\leq 4\,\mathbb{E}_{\mathbf{\Lambda}}\Big[\Big\|\sum_{i>j}\Big(\eta_i^\star(1-\eta_j^\star) + \eta_j^\star(1-\eta_i^\star)\Big)\lambda_i\lambda_j(\mathbf{M}_{ij} + \mathbf{M}_{ji})\Big\|_F\Big] && \text{(C.14)}
\end{aligned}$$

$$
\begin{aligned}
&= 4\,\mathbb{E}_{\Lambda}\Big[\Big\| \sum_{i\in T, j\in T^c} \lambda_i\lambda_j(\mathbf{M}_{ij}+\mathbf{M}_{ji})\Big\|_F\Big] && \text{(C.15)}\\
&\le 4\,\mathbb{E}_{\mathbf{\Lambda}}\Big[\Big\| \sum_{i\in T, j\in T^c} \lambda_i\lambda_j\mathbf{M}_{ij}\Big\|_F\Big] + 4\,\mathbb{E}_{\mathbf{\Lambda}}\Big[\Big\| \sum_{i\in T, j\in T^c} \lambda_i\lambda_j\mathbf{M}_{ji}\Big\|_F\Big]. && \text{(C.16)}
\end{aligned}
$$

Let $\mathbf{\Lambda}'$ be an independent copy of $\mathbf{\Lambda}$, and then (C.16) is equal to

$$
\begin{aligned}
&4\,\mathbb{E}_{\mathbf{\Lambda},\mathbf{\Lambda}'}\Big[\Big\| \sum_{i\in T, j\in T^c} \lambda_i'\lambda_j\mathbf{M}_{ij}\Big\|_F\Big] + 4\,\mathbb{E}_{\mathbf{\Lambda},\mathbf{\Lambda}'}\Big[\Big\| \sum_{i\in T, j\in T^c} \lambda_i\lambda_j'\mathbf{M}_{ji}\Big\|_F\Big]\\
&\le 8\,\mathbb{E}_{\mathbf{\Lambda},\mathbf{\Lambda}'}\Big[\Big\| \sum_{i,j=1}^{n} \lambda_i'\lambda_j\mathbf{M}_{ij}\Big\|_F\Big] && \text{(C.17)}\\
&= 8\,\mathbb{E}_{\mathbf{\Lambda},\mathbf{\Lambda}'}\Big[\|\mathbf{\Lambda}'\mathbf{M}\mathbf{\Lambda}\|_F\Big] && \text{(C.18)}\\
&\le 8\,\mathbb{E}_{\mathbf{\Lambda}}\Big(\mathbb{E}_{\mathbf{\Lambda}'}\Big[\|\mathbf{\Lambda}'\mathbf{M}\mathbf{\Lambda}\|_F^2\Big]\Big)^{1/2} && \text{(C.19)}\\
&= 8\sqrt{k/n}\,\mathbb{E}_{\mathbf{\Lambda}}\Big[\|\mathbf{M}\mathbf{\Lambda}\|_F\Big]. && \text{(C.20)}
\end{aligned}
$$

Now we move from the Bernoulli model back to the uniform model. Conditioned on $k'=s$, we can divide support($\mathbf{\Lambda}$) into $a=\lceil k'/k\rceil$ random subsets $S_1,\ldots,S_a$ of size at most k, and the marginal distribution of each S_i is uniform on $\binom{[n]}{|S_i|}$. Hence

$$
\mathbb{E}_{\mathbf{\Lambda}}\Big[\|\mathbf{M}\mathbf{P}_{S_i}\|_F \mid k'=s\Big] \;\le\; \begin{cases} \mathbb{E}_{\Omega}\left[\|\mathbf{M}\mathbf{P}_{\Omega}\|_F\right] & (i-1)k<s \\ 0 & \text{else} \end{cases}. \tag{C.21}
$$

Therefore,

$$
\begin{aligned}
\mathbb{E}_{\mathbf{\Lambda}}\Big[\,\|\mathbf{M}\mathbf{\Lambda}\|_F\,\Big] &\le \mathbb{E}_{\mathbf{\Lambda}}\Big[\sum_i \|\mathbf{M}\mathbf{P}_{S_i}\|_F\Big] && \text{(C.22)}\\
&= \sum_{s=0}^{n}\sum_i \mathbb{E}_{\Lambda}\left[\|\mathbf{M}\mathbf{P}_{S_i}\|_F \mid k'=s\right]\,\mathbb{P}[k'=s]. && \text{(C.23)}\\
&\le \sum_{s=0}^{n}\Big\lfloor \frac{s}{k}+1\Big\rfloor\;\mathbb{E}_{\Omega}\left[\|\mathbf{M}\mathbf{P}_{\Omega}\|_F\right]\,\mathbb{P}[k'=s] && \text{(C.24)}\\
&= \mathbb{E}_{\Omega}\left[\,\|\mathbf{M}\mathbf{P}_{\Omega}\|_F\,\right]\mathbb{E}\left[k'/k+1\right] && \text{(C.25)}\\
&= 2\,\mathbb{E}_{\Omega}\left[\,\|\mathbf{M}\mathbf{P}_{\Omega}\|_F\,\right]. && \text{(C.26)}
\end{aligned}
$$

Combining (C.20) and (C.26) gives (C.1). $\square$

APPENDIX D

PROOF OF LEMMA 5.2

In this section we prove Lemma 5.2 by applying the matrix Chernoff bound in Appendix B to the extreme eigenvalues of

$$\mathbf{X}\mathbf{X}^* = \sum_{j=1}^{p} \mathbf{x}_j \mathbf{x}_j^* \tag{D.1}$$

as a sum of independent positive semidefinite matrices.

Since the summands $\mathbf{x}_j \mathbf{x}_j^*$ may have unbounded norm, we will use truncation technique to replace $\mathbf{x}_j \mathbf{x}_j^*$ by truncated terms $\bar{\mathbf{x}}_j \bar{\mathbf{x}}_j^*$ which are equivalent to $\mathbf{x}_j \mathbf{x}_j^*$ with very high probability.

Proof of Lemma 5.2. Let

$$\bar{\mathbf{x}}_j = \begin{cases} \mathbf{x}_j & \|\mathbf{x}_j\| \ \leq \ (1+\beta)\sqrt{n \log p / p} \\ \mathbf{0} & \text{else} \end{cases} \tag{D.2}$$

were $\beta > 0$ is a constant which we will choose later. From Gaussian measure concentration, for each j

$$\mathbb{P}\left[\|\mathbf{x}_j\| > (1+\beta)\sqrt{\frac{n \log p}{p}} \right] \ < \ p^{-\beta^2/2}. \tag{D.3}$$

Therefore, by union bound $\max_j \|\mathbf{x}_j\|$ is bounded by $(1+\beta)\sqrt{n \log p/p}$ with probability at least $1 - p^{1-\beta^2/2}$. Hence, with probability at least $1 - p^{1-\beta^2/2}$, $\bar{\mathbf{x}}_j = \mathbf{x}_j, \ \forall \, j$ and thus $\sum_j \mathbf{x}_j \mathbf{x}_j^* = \sum_j \bar{\mathbf{x}}_j \bar{\mathbf{x}}_j^*$. Due to truncation, we have:

$$\|\bar{\mathbf{x}}_j \bar{\mathbf{x}}_j^*\| \ = \ \|\bar{\mathbf{x}}_j\|^2 \ \leq \ B \ \doteq \ (1+\beta)^2 \frac{n \log p}{p}. \tag{D.4}$$

Since by definition $\mathbf{x}_j\mathbf{x}_j^* \succeq \bar{\mathbf{x}}_j\bar{\mathbf{x}}_j^*$ for all j, $\mathbb{E}[\mathbf{x}_j\mathbf{x}_j^*] \succeq \mathbb{E}[\bar{\mathbf{x}}_j\bar{\mathbf{x}}_j^*]$, and thus

$$\mu_{max} \;\doteq\; \left\|\mathbb{E}\Big[\sum_j \bar{\mathbf{x}}_j\bar{\mathbf{x}}_j^*\Big]\right\| \;\le\; \left\|\mathbb{E}\Big[\sum_j \mathbf{x}_j\mathbf{x}_j^*\Big]\right\| \;=\; \|\mathbf{I}\| \;=\; 1. \tag{D.5}$$

Plugging (D.5) into (B.3), we have

$$\mathbb{P}\Big[\;\lambda_{max}\Big(\sum_j \bar{\mathbf{x}}_j\bar{\mathbf{x}}_j^*\Big) \;\ge\; 1+t\;\Big] \le n\exp\left(-\frac{t^2\,\mu_{max}\,p}{4\,(1+\beta)^2\,n\,\log p}\right). \tag{D.6}$$

Notice that the right-hand side of (D.6) still depends on $\mu_{max} \le 1$. We will resolve this by developing a lower bound on μ_{min} and thus on μ_{max}.

For the smallest eigenvalue, we have

$$\begin{aligned}
\mu_{min} &\doteq \lambda_{min}\Big(\mathbb{E}\Big[\sum_j \bar{\mathbf{x}}_j\bar{\mathbf{x}}_j^*\Big]\Big) && \text{(D.7)}\\
&\ge \lambda_{min}\Big(\mathbb{E}\Big[\sum_j \mathbf{x}_j\mathbf{x}_j^*\Big]\Big) - \left\|\mathbb{E}\Big[\sum_j \bar{\mathbf{x}}_j\bar{\mathbf{x}}_j^* - \mathbf{x}_j\mathbf{x}_j^*\Big]\right\| && \text{(D.8)}\\
&\ge \lambda_{min}\Big(\mathbb{E}\Big[\sum_j \mathbf{x}_j\mathbf{x}_j^*\Big]\Big) - \sum_j \mathbb{E}\Big[\left\|\bar{\mathbf{x}}_j\bar{\mathbf{x}}_j^* - \mathbf{x}_j\mathbf{x}_j^*\right\|\Big] && \text{(D.9)}\\
&= \lambda_{min}(\mathbf{I}) - \sum_j \mathbb{E}\left[\left\|\mathbf{x}_j\mathbf{x}_j^*\right\|\,\mathbf{1}_{\|\mathbf{x}_j\|>\sqrt{B}}\right] && \text{(D.10)}\\
&= 1 - \sum_j \mathbb{E}\left[\|\mathbf{x}_j\|_2^2\mathbf{1}_{\|\mathbf{x}_j\|>\sqrt{B}}\right] && \text{(D.11)}\\
&\ge 1 - \sum_j \sqrt{\mathbb{E}[\|\mathbf{x}_j\|_2^4]}\sqrt{\mathbb{E}[(\mathbf{1}_{\|\mathbf{x}_j\|>\sqrt{B}})^2]} && \text{(D.12)}\\
&= 1 - p\sqrt{\mathbb{E}[\|\mathbf{x}_1\|_2^4]}\sqrt{\mathbb{P}[\|\mathbf{x}_1\| > \sqrt{B}]} && \text{(D.13)}\\
&\ge 1 - p \times \sqrt{3}\,n/p \times p^{-\beta^2/4} && \text{(D.14)}\\
&= 1 - \sqrt{3}\,np^{-\beta^2/4}, && \text{(D.15)}
\end{aligned}$$

where in (D.9) we have used Jensen's inequality, and in (D.12) we have used the Cauchy-Schwarz inequality. Finally, in (D.15) we use the following bound on $\mathbb{E}\|\mathbf{x}_1\|^4$:

$$\mathbb{E}\left[\|\mathbf{x}_1\|_2^4\right] = (k^2+2k)\sigma^4 \;\le\; 3k^2\sigma^4 \;=\; 3\,n^2/p^2. \tag{D.16}$$

So we can write $\mu_{min} \ge 1 - g(p)$, where $g(p) = \sqrt{3}\,np^{-\beta^2/4}$. Apply Tropp's bound (B.1) and we get

$$\mathbb{P}\Big[\lambda_{min}\Big(\sum_j \bar{\mathbf{x}}_j\bar{\mathbf{x}}_j^*\Big) < 1-t\Big] \;\leq\; n\exp\left(-\frac{(t-g(p))^2}{2(\beta+1)^2}\frac{p}{n\log p}\right). \qquad \text{(D.17)}$$

Now let $\beta = 4$. Then provided $p > (Cn/t)^{1/4}$, we have $g(p) < t/2 < 1/2$. Combining (D.6) and (D.17) establishes Lemma 5.2. □

APPENDIX E

PROOF OF LEMMA 5.3

In this section, we prove Lemma 5.3, which estimates the norm of the residual $\mathbf{P}_\Omega \mathbf{R} \mathbf{P}_\Omega$. To establish this result in Lemma 5.3, we first write $\mathbf{P}_\Omega \mathbf{R} \mathbf{P}_\Omega$ as a sum of random semidefinite matrices that are independent conditioned on Ω and then apply the matrix Chernoff bound in Appendix B to show with high probability that $\|\mathbf{P}_\Omega \mathbf{R} \mathbf{P}_\Omega\|$ is bounded by a small constant.

E.1 Proof of Lemma 5.3

Proof. To simplify notations, let $\mathbf{x}^i$ denote the i-th row of $\mathbf{X}$, and $\mathbf{x}_j$ be j-th column of $\mathbf{X}$, where matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$. Similarly, we define

$$\Omega^i \doteq \{j \mid (i,j) \in \Omega\} \subseteq [p], \tag{E.1}$$

and

$$\Omega_j \doteq \{i \mid (i,j) \in \Omega\} \subset [n]. \tag{E.2}$$

By using the familiar identity $(\mathbf{P} \otimes \mathbf{Q}) = (\mathbf{P} \otimes \mathbf{I})(\mathbf{I} \otimes \mathbf{Q})$, we can write $\mathbf{R}$ as

$$\mathbf{R} \;=\; (\mathbf{X}^* \otimes \mathbf{I})(\mathbf{I} \otimes \mathbf{A}^*)(\mathbf{I} - \mathbf{C}_A \mathbf{C}_A^*)(\mathbf{I} \otimes \mathbf{A})(\mathbf{X} \otimes \mathbf{I}), \tag{E.3}$$

where the product of the middle three terms is a block diagonal matrix

$$(\mathbf{I} \otimes \mathbf{A}^*)(\mathbf{I} - \mathbf{C}_A \mathbf{C}_A^*)(\mathbf{I} \otimes \mathbf{A}) \;=\; \begin{bmatrix} \mathbf{A}^*(\mathbf{I} - \mathbf{A}_1 \mathbf{A}_1^*)\mathbf{A} & & \\ & \ddots & \\ & & \mathbf{A}^*(\mathbf{I} - \mathbf{A}_n \mathbf{A}_n^*)\mathbf{A} \end{bmatrix}. \tag{E.4}$$

Let $\mathbf{P}_i = \mathbf{I} - \mathbf{A}_i \mathbf{A}_i^*$. Then we can expand the product in (E.3) more explicitly by

$$\mathbf{R} \;=\; \begin{bmatrix} \sum_{b=1}^{n} X_{b,1} X_{b,1} \mathbf{A}^* \mathbf{P}_b \mathbf{A} & \dots & \sum_{b=1}^{n} X_{b,1} X_{b,p} \mathbf{A}^* \mathbf{P}_b \mathbf{A} \\ \vdots & \ddots & \vdots \\ \sum_{b=1}^{n} X_{b,p} X_{b,1} \mathbf{A}^* \mathbf{P}_b \mathbf{A} & \dots & \sum_{b=1}^{n} X_{b,p} X_{b,p} \mathbf{A}^* \mathbf{P}_b \mathbf{A} \end{bmatrix}. \tag{E.5}$$

Thus

$$\mathbf{P}_\Omega \mathbf{R} \mathbf{P}_\Omega = \sum_{b=1}^{n} \mathbf{P}_\Omega \left(\mathbf{x}^{b*} \mathbf{x}^b \otimes \mathbf{A}^* \mathbf{P}_b \mathbf{A} \right) \mathbf{P}_\Omega. \tag{E.6}$$

Let

$$\begin{aligned} \mathbf{\Psi}_i \;&\doteq\; \mathbf{P}_\Omega \left(\mathbf{x}^{i*} \mathbf{x}^i \otimes \mathbf{A}^* \mathbf{P}_i \mathbf{A} \right) \mathbf{P}_\Omega \\ &=\; \mathbf{P}_\Omega \left(\mathbf{P}_{\Omega^i} \mathbf{v}^{i*} \mathbf{v}^i \mathbf{P}_{\Omega^i} \otimes \mathbf{A}^* \mathbf{P}_i \mathbf{A} \right) \mathbf{P}_\Omega; \end{aligned} \tag{E.7}$$

then we can write $\mathbf{P}_\Omega \mathbf{R} \mathbf{P}_\Omega$ as a sum of random positive semidefinite matrices

$$\mathbf{P}_\Omega \mathbf{R} \mathbf{P}_\Omega \;=\; \sum_{i=1}^{n} \mathbf{\Psi}_i. \tag{E.8}$$

Note that by definition (E.7), conditioned on Ω, $\mathbf{\Psi}_i$ only depends on independent random vectors $\mathbf{v}^i$. Hence, conditioned Ω, $\{\mathbf{\Psi}_i\}_{1 \le i \le n}$ are independent. So we would like to apply the matrix Chernoff bound in Appendix B to bound the size of the sum of $\{\mathbf{\Psi}_i\}_{1 \le i \le n}$ conditioned on Ω. Before doing this, we need to understand how Ω affects the size of $\mathbf{\Psi}_i$.

Since the support of each $\mathbf{x}_i$ is independent, with high probability Ω is quite regular. Indeed, the expected size of Ω^i is simply pk/n for any $i \in [n]$. Furthermore, for distinct i and i', $|\Omega^i \cap \Omega^{i'}|$ concentrates about the expectation, which is simply bounded by $k^2 p/n^2$. We define a set of supports, in which these quantities do not greatly exceed their expectations:

$$\mathcal{O} \doteq \left\{ \Omega \subset [n] \times [p] \;\middle|\; \begin{array}{l} \max_{i=1,\dots,n} |\Omega^i| \le 3pk/2n \\ \max_{i \ne i'} |\Omega^i \cap \Omega^{i'}| \le 3pk^2/2n^2 \end{array} \right\}. \tag{E.9}$$

By measure concentration, it should be expected that the event $\Omega \in \mathcal{O}$ is highly likely. More precisely,

Lemma E.1. *With overwhelming probability, $\Omega \in \mathcal{O}$:*

$$\mathbb{P}\left[\Omega \in \mathcal{O}\right] \;\geq\; 1 - n^2 \exp\left(-\frac{pk^2}{10n^2}\right). \tag{E.10}$$

The proof of Lemma E.1 is a standard application of Bernstein's inequality, and thus is omitted.

Conditioned on $\Omega \in \mathcal{O}$, for any $i \in [n]$, the norm of the i-th row $\mathbf{x}^i$ also concentrates about their conditional expectation, and further it will not concentrate too strongly on the intersection $\Omega^{i'} \cap \Omega^i$ for any $i' \neq i$. For each $i \in [n]$, we define

$$\mathcal{E}_i \doteq \left\{ \omega \;\middle|\; \max_{a \neq i} \|\mathbf{x}^i \mathbf{P}_{\Omega^a}\| \;\leq\; 2\sqrt{k/n}, \;\text{ and }\; \|\mathbf{x}^i\| \;\leq\; 2 \right\}, \tag{E.11}$$

and set

$$\mathcal{E}_X \doteq \cap_{i=1}^n \mathcal{E}_i. \tag{E.12}$$

Similarly, by measure concentration we expect $\mathcal{E}_X$ is overwhelmingly likely:

Lemma E.2. *For any $\Omega \in \mathcal{O}$,*

$$\mathbb{P}[\mathcal{E}_X \mid \Omega] \;\geq\; 1 - n^2 \exp\left(-\frac{k^2 p}{4n^2}\right). \tag{E.13}$$

We prove Lemma E.2 in Section E.2. In Lemma E.3, we show that conditioned on $\mathcal{E}_i$, $\mathbf{\Psi}_i$ indeed has a small norm:

Lemma E.3. *Let $\mathcal{E}_i$ be the event defined in* (E.11)*, and let $\mathbf{\Psi}_i$ denote the i-th residual term:*

$$\mathbf{\Psi}_i \;=\; \mathbf{P}_\Omega \left(\mathbf{x}^{i*}\mathbf{x}^i \otimes \mathbf{A}^*\mathbf{P}_i\mathbf{A}\right) \mathbf{P}_\Omega \tag{E.14}$$

Then on event $\mathcal{E}_i$, we have

$$\|\mathbf{\Psi}_i\| \;\leq\; 4k/n + 24\, k\mu(\mathbf{A}). \tag{E.15}$$

We prove Lemma E.3 in Section E.3.

Next we show how to apply Lemma E.1, E.2 and E.3 to establish the desired result on bounding $\|\mathbf{P}_\Omega \mathbf{R} \mathbf{P}_\Omega\|$. Write $\mathbf{\Psi} \doteq \sum_i \mathbf{\Psi}_i$, and set

$$\bar{\mathbf{\Psi}}_i = \mathbf{\Psi}_i \times \mathbf{1}_{\mathcal{E}_i}, \tag{E.16}$$

where $\mathbf{1}_{\mathcal{E}_i}$ denotes the indicator function of the event $\mathcal{E}_i$. By Lemma E.3, $\bar{\mathbf{\Psi}}_i$ satisfies

$$\|\bar{\mathbf{\Psi}}_i\| \ \leq \ 4k/n + 24\, k\mu(\mathbf{A}) \ \doteq \ B. \tag{E.17}$$

Then we can bound the probability that $\|\mathbf{\Psi}\|$ exceeds 1/8 by analyzing the probability that $\|\bar{\mathbf{\Psi}}\|$ exceeds 1/8:

$$\begin{aligned}
&\mathbb{P}\left[\|\mathbf{\Psi}\| \geq 1/8\right] \\
&= \mathbb{P}\left[\|\mathbf{\Psi}\| \geq 1/8 \mid \Omega \in \mathcal{O}\right] \mathbb{P}\left[\Omega \in \mathcal{O}\right] + \mathbb{P}\left[\|\mathbf{\Psi}\| \geq 1/8 \mid \Omega \in \mathcal{O}^c\right] \mathbb{P}\left[\Omega \in \mathcal{O}^c\right] \\
&\leq \mathbb{P}\left[\|\mathbf{\Psi}\| \geq 1/8 \mid \Omega \in \mathcal{O}\right] \ + \ \mathbb{P}\left[\Omega \in \mathcal{O}^c\right] \\
&\leq \ \max_{\Omega_0 \in \mathcal{O}} \mathbb{P}\left[\|\mathbf{\Psi}\| \geq 1/8 \mid \Omega_0\right] \ + \ \mathbb{P}\left[\Omega \in \mathcal{O}^c\right] \\
&\leq \ \max_{\Omega_0 \in \mathcal{O}} \left\{ \mathbb{P}\left[\|\bar{\mathbf{\Psi}}\| \geq 1/8 \mid \Omega_0\right] + \mathbb{P}\left[\mathbf{\Psi} \neq \bar{\mathbf{\Psi}} \mid \Omega_0\right] \right\} \ + \ \mathbb{P}\left[\Omega \in \mathcal{O}^c\right] \\
&\leq \ \max_{\Omega_0 \in \mathcal{O}} \left\{ \mathbb{P}\left[\|\bar{\mathbf{\Psi}}\| \geq 1/8 \mid \Omega_0\right] + \mathbb{P}\left[\cup_i \mathcal{E}_i^c \mid \Omega_0\right] \right\} \ + \ \mathbb{P}\left[\Omega \in \mathcal{O}^c\right] \\
&= \ \max_{\Omega_0 \in \mathcal{O}} \left\{ \mathbb{P}\left[\|\bar{\mathbf{\Psi}}\| \geq 1/8 \mid \Omega_0\right] + \mathbb{P}\left[\mathcal{E}_X^c \mid \Omega_0\right] \right\} \ + \ \mathbb{P}\left[\Omega \in \mathcal{O}^c\right] \\
&\leq \ \max_{\Omega_0 \in \mathcal{O}} \mathbb{P}\left[\|\bar{\mathbf{\Psi}}\| \geq 1/8 \mid \Omega_0\right] + n^2 \exp\left(-\frac{k^2 p}{4n^2}\right) + n^2 \exp\left(-\frac{k^2 p}{10 n^2}\right) && \text{(E.18)} \\
&\leq \ \max_{\Omega_0 \in \mathcal{O}} \mathbb{P}\left[\|\bar{\mathbf{\Psi}}\| \geq 1/8 \mid \Omega_0\right] + 2n^2 \exp\left(-\frac{k^2 p}{10 n^2}\right), && \text{(E.19)}
\end{aligned}$$

where in (E.18) we have used Lemma E.1 and E.2.

To complete the proof, we only need to bound the first term in (E.19). Since $\bar{\mathbf{\Psi}}_i$ is the sum of a sequence of independent random positive semidefinite matrices conditioned on Ω_0, we can apply the matrix Chernoff bound in Appendix B. First, we need to estimate $\mu_{max} = \|\mathbb{E}[\bar{\mathbf{\Psi}} \mid \Omega_0]\|$, which can be bounded by

$$\begin{aligned}
\mu_{max} \ &= \|\mathbb{E}[\bar{\mathbf{\Psi}} \mid \Omega_0]\| \\
&\leq \|\mathbb{E}[\mathbf{\Psi} \mid \Omega_0]\| \\
&= \|\mathbb{E}_V \left[\sum_{i=1}^n \mathbf{P}_{\Omega_0} \left(\mathbf{P}_{\Omega_0^i} \mathbf{v}^{i*} \mathbf{v}^i \mathbf{P}_{\Omega_0^i} \otimes \mathbf{A}^* \mathbf{P}_i \mathbf{A} \right) \mathbf{P}_{\Omega_0} \right] \|
\end{aligned}$$

$$= \|\frac{n}{kp}\sum_{i=1}^{n}\mathbf{P}_{\Omega_0}\left(\mathbf{P}_{\Omega_0^i}\otimes\mathbf{A}^*\mathbf{P}_i\mathbf{A}\right)\mathbf{P}_{\Omega_0}\|$$
$$\leq \|\frac{n}{kp}\sum_{i=1}^{n}\mathbf{P}_{\Omega_0}\left(\mathbf{P}_{\Omega_0^i}\otimes\mathbf{A}^*\mathbf{A}\right)P_{\Omega_0}\|$$
$$= \|\frac{n}{kp}\mathbf{P}_{\Omega_0}\left(\sum_{i=1}^{n}\mathbf{P}_{\Omega_0^i}\otimes\mathbf{A}^*\mathbf{A}\right)\mathbf{P}_{\Omega_0}\|$$
$$= \frac{n}{p}\|\mathbf{P}_{\Omega_0}\left(\mathbf{I}\otimes\mathbf{A}^*\mathbf{A}\right)\mathbf{P}_{\Omega_0}\|, \tag{E.20}$$

where in (E.20) we have used the fact that $\sum_i \mathbf{P}_{\Omega_0^i} = k\mathbf{I}$.

In (E.20), $\mathbf{P}_{\Omega_0}\left(\mathbf{I}\otimes\mathbf{A}^*\mathbf{A}\right)\mathbf{P}_{\Omega_0}$ is a block diagonal matrix and each block has a norm bounded by $\|A_{\Omega_j}\|^2$, which is again upper bounded by $1 + k\mu(\mathbf{A})$ due to (A.3). Therefore, provided $k\mu(\mathbf{A}) < 1/2$, we have

$$\mu_{max} \ \leq \ 3n/2p. \tag{E.21}$$

Let $t\mu_{max} = 1/8$. Then $t \geq p/12n \geq e$, so we can apply the matrix Chernoff bound (B.4) and get

$$\mathbb{P}\left[\left\|\bar{\mathbf{\Psi}}\right\| \geq 1/8 \mid \Omega\right] \ \leq \ np\left(\frac{12en}{p}\right)^{\frac{1}{8B}}, \tag{E.22}$$

where B is the bound on the norm of the summands $\bar{\mathbf{\Psi}}_i$. By choosing sufficiently small constants C_1 and C_2 in the statement of Lemma 5.3 , we can make the exponent $\nu = \frac{1}{8B}$ as large as desired. Assuming $p \geq Cn^2$, and by appropriate choice of C_1 and C_2, we can bound the probability that $\|\bar{\mathbf{\Psi}}\|$ exceeds $1/8$ by

$$\mathbb{P}\left[\left\|\bar{\mathbf{\Psi}}\right\| \geq 1/8 \mid \Omega\right] \ \leq \ C(\nu)p^{-3}. \tag{E.23}$$

Plugging into (E.19) completes the proof. □

E.2 Proof of Lemma E.2

Proof. This proof is an application of Gaussian measure concentration [18]. If $\mathbf{v}$ is an i.i.d. sequence of $\mathcal{N}(0,\sigma^2)$ random variables, and f is a positively homogeneous, 1-Lipschitz function, then

$$\mathbb{P}\left[f(\mathbf{v}) \geq \mathbb{E}[f(\mathbf{v})] + t\right] \leq \exp\left(-\frac{t^2}{2\sigma^2}\right). \tag{E.24}$$

Given $\Omega \in \mathcal{O}$, we define $f(\mathbf{v}^i) \doteq \|\mathbf{v}^i \mathbf{P}_{\Omega^i}\| = \|\mathbf{x}^i\|$. It is easy to see $f(\mathbf{v}^i)$ is a 1-Lipschitz function of $\mathbf{v}^i$. Since $\Omega \in \mathcal{O}$, $|\Omega^i| \leq 3pk/2n$ and thus

$$\mathbb{E}\left[\|\mathbf{x}^i\| \mid \Omega\right] \leq \sqrt{\mathbb{E}\left[\|\mathbf{x}^i\|^2 \mid \Omega\right]} = \sqrt{|\Omega^i| n/kp} \leq \sqrt{3/2}. \tag{E.25}$$

Applying Gaussian measure concentration, we get

$$\begin{aligned}\mathbb{P}[\|\mathbf{x}^i\| \geq 2 \mid \Omega] \;&\leq\; \mathbb{P}\left[f(\mathbf{v}^i) \geq \mathbb{E}[f(\mathbf{v}^i) \mid \Omega] + (2 - \sqrt{3/2}) \mid \Omega\right] && \text{(E.26)}\\ &\leq\; \exp\left(-\frac{kp}{4n}\right). && \text{(E.27)}\end{aligned}$$

Similarly, given $i' \neq i$, we define $g(\mathbf{v}^i) \doteq \|\mathbf{v}^i \mathbf{P}_{\Omega^i \cap \Omega^{i'}}\| = \|\mathbf{x}^i \mathbf{P}_{\Omega^{i'}}\|$. It is easy to check that $g(\cdot)$ is also a 1-Lipschitz function of $\mathbf{v}^i$. Since $\Omega \in \mathcal{O}$, $|\Omega^i \cap \Omega^{i'}| \leq 3pk^2/2n^2$, and by Jensen's inequality,

$$\mathbb{E}\left[g(\mathbf{v}^i) \mid \Omega\right] \leq \sqrt{\mathbb{E}\left[g(\mathbf{v}^i)^2 \mid \Omega\right]} \;=\; \leq \; \sqrt{3k/2n}. \tag{E.28}$$

By Gaussian measure concentration,

$$\begin{aligned}\mathbb{P}\left[g(\mathbf{v}^i) \geq 2\sqrt{k/n} \mid \Omega\right] \;&\leq\; \mathbb{P}\left[g(\mathbf{v}^i) \geq \mathbb{E}[g(\mathbf{v}^i) \mid \Omega] + (2 - \sqrt{3/2})\sqrt{k/n} \mid \Omega\right]\\ &\leq\; \exp\left(-\frac{k^2 p}{4n^2}\right). && \text{(E.29)}\end{aligned}$$

Taking the union bound over all n choices of i in (E.26) and all $n(n-1)$ ordered pairs (i, i') in (E.29) gives (E.13). $\square$

E.3 Proof of Lemma E.3

Proof. We will show the calculations for $i = 1$ only, and the same argument works for $i = 2, \ldots, n$ as well. Recall that $\mathbf{\Psi}_1 \;=\; \mathbf{P}_\Omega \left(\mathbf{x}^{1*}\mathbf{x}^1 \otimes \mathbf{A}^*\mathbf{P}_1\mathbf{A}\right) \mathbf{P}_\Omega$. The term $\mathbf{A}^*\mathbf{P}_1\mathbf{A} \;=\; \mathbf{A}^*\mathbf{A} - \mathbf{A}^*\mathbf{A}_1\mathbf{A}_1^*\mathbf{A} \;\approx\; \mathbf{I} - \mathbf{e}_1\mathbf{e}_1^*$, since $\mathbf{A}$ is incoherent and thus an approximately orthogonal matrix. Let

$$\mathbf{\Delta} \;\doteq\; \mathbf{A}^*\mathbf{P}_1\mathbf{A} - (\mathbf{I} - \mathbf{e}_1\mathbf{e}_1^*) \in \mathbb{R}^{n\times n}; \tag{E.30}$$

then

$$\|\mathbf{\Delta}\|_\infty \;\le\; \|\mathbf{A}^*\mathbf{A} - \mathbf{I}\|_\infty + \|\mathbf{A}^*\mathbf{A}_1\mathbf{A}_1^*\mathbf{A} - \mathbf{e}_1\mathbf{e}_1^*\|_\infty \;\le\; 2\mu(\mathbf{A}). \tag{E.31}$$

Write

$$\begin{aligned} \|\mathbf{\Psi}_1\| &= \left\|\mathbf{P}_\Omega\left(\mathbf{x}^{1^*}\mathbf{x}^1 \otimes (\mathbf{I} - \mathbf{e}_1\mathbf{e}_1^* + \mathbf{\Delta})\right)\mathbf{P}_\Omega\right\| \\ &\le \left\|\mathbf{P}_\Omega\left(\mathbf{x}^{1^*}\mathbf{x}^1 \otimes (\mathbf{I} - \mathbf{e}_1\mathbf{e}_1^*)\right)\mathbf{P}_\Omega\right\| + \left\|\mathbf{P}_\Omega(\mathbf{x}^{1^*}\mathbf{x}^1 \otimes \mathbf{\Delta})\mathbf{P}_\Omega\right\|. \end{aligned} \tag{E.32}$$

We handle these two terms in (E.32) separately. For the first term

$$\mathbf{L} \;\doteq\; \mathbf{P}_\Omega\left(\mathbf{x}^{1^*}\mathbf{x}^1 \otimes (\mathbf{I} - \mathbf{e}_1\mathbf{e}_1^*)\right)\mathbf{P}_\Omega \;\in\; \mathbb{R}^{np\times np}, \tag{E.33}$$

we let $\mathcal{L} : \mathbb{R}^{n\times p} \to \mathbb{R}^{n\times p}$ be the equivalent linear operator via

$$\mathrm{vec}\,[\,\mathcal{L}[\mathbf{Q}]\,] \;=\; \mathbf{L}\,\mathrm{vec}[\mathbf{Q}] \tag{E.34}$$

for all $\mathbf{Q} \in \mathbb{R}^{n\times p}$. Therefore,

$$\|\mathbf{L}\| \;=\; \|\mathcal{L}\| \;\doteq\; \sup_{\mathbf{Q}\neq\mathbf{0}} \frac{\|\mathcal{L}[\mathbf{Q}]\|_F}{\|\mathbf{Q}\|_F}. \tag{E.35}$$

By the familiar identity $\mathrm{vec}[\mathbf{PQR}] = (\mathbf{R}^* \otimes \mathbf{P})\,\mathrm{vec}[\mathbf{Q}]$, the operator $\mathcal{L}[\mathbf{Q}]$ is given by

$$\mathcal{L}[\mathbf{Q}] \;=\; \mathcal{P}_\Omega\left[(\mathbf{I} - \mathbf{e}_1\mathbf{e}_1^*)\;\mathcal{P}_\Omega[\mathbf{Q}]\;\mathbf{x}^{1^*}\mathbf{x}^1\right]. \tag{E.36}$$

We can expand (E.36) by expressing $\mathcal{P}_\Omega[\mathbf{H}]$ as $\sum_{a=1}^n \mathbf{e}_a\mathbf{e}_a^*\mathbf{H}\mathbf{P}_{\Omega^a}$, for any $\mathbf{H} \in \mathbb{R}^{n\times p}$. Then (E.36) becomes

$$\begin{aligned} \mathcal{L}[\mathbf{Q}] \;&=\; \sum_{a=1}^n \mathbf{e}_a\mathbf{e}_a^*\left[(\mathbf{I} - \mathbf{e}_1\mathbf{e}_1^*)\;\mathcal{P}_\Omega[\mathbf{Q}]\;\mathbf{x}^{1^*}\mathbf{x}^1\right]\mathbf{P}_{\Omega^a} \\ &=\; \sum_{a=2}^n \mathbf{e}_a\mathbf{e}_a^*\;\mathcal{P}_\Omega[\mathbf{Q}]\;\mathbf{x}^{1^*}\mathbf{x}^1\mathbf{P}_{\Omega^a} \\ &=\; \sum_{a=2}^n \mathbf{e}_a\mathbf{e}_a^*\left(\sum_{b=1}^n \mathbf{e}_b\mathbf{e}_b^*\;\mathbf{Q}\;\mathbf{P}_{\Omega^b}\right)\mathbf{x}^{1^*}\mathbf{x}^1\mathbf{P}_{\Omega^a} \end{aligned}$$

$$= \sum_{a=2}^{n} \mathbf{e}_a \mathbf{e}_a^* \; \mathbf{Q} \; \mathbf{P}_{\Omega^a} \mathbf{x}^{1^*} \mathbf{x}^1 \mathbf{P}_{\Omega^a}. \tag{E.37}$$

Since the summands in (E.37) do not overlap with each other,

$$\begin{aligned} \|\mathcal{L}[\mathbf{Q}]\|_F^2 \;&=\; \sum_{a=2}^{n} \|\mathbf{e}_a^* \; \mathbf{Q} \; \mathbf{P}_{\Omega^a} \mathbf{x}^{1^*} \mathbf{x}^1 \mathbf{P}_{\Omega^a}\|^2 && \text{(E.38)} \\ &\le\; \sum_{a=2}^{n} \|\mathbf{e}_a^* \mathbf{Q}\|^2 \|\mathbf{x}^1 \mathbf{P}_{\Omega^a}\|^4 && \text{(E.39)} \\ &\le\; 16 \frac{k^2}{n^2} \|\mathbf{Q}\|_F^2. && \text{(E.40)} \end{aligned}$$

Therefore, we conclude

$$\|\mathbf{L}\| \;\le\; 4k/n. \tag{E.41}$$

Next we address the second term $\mathbf{W} \doteq \mathbf{P}_\Omega \left(\mathbf{x}^{1^*} \mathbf{x}^1 \otimes \mathbf{\Delta} \right)$ in (E.32). Similarly, we consider the associated linear map operator $\mathcal{W} : \mathbb{R}^{n \times p} \to \mathbb{R}^{n \times p}$, given by

$$\begin{aligned} \mathcal{W}[\mathbf{Q}] \;&= \mathcal{P}_\Omega \left[\mathbf{\Delta} \; \mathcal{P}_\Omega[\mathbf{Q}] \; \mathbf{x}^{1^*} \mathbf{x}^1 \right] \\ &= \sum_{a=1}^{n} \sum_{b=1}^{n} \; \mathbf{e}_a \mathbf{e}_a^* \; \mathbf{\Delta} \; \mathbf{e}_b \mathbf{e}_b^* \; \mathbf{Q} \; \mathbf{P}_{\Omega^b} \; \mathbf{x}^{1^*} \mathbf{x}^1 \mathbf{P}_{\Omega^a}. && \text{(E.42)} \end{aligned}$$

We break the summation in (E.42) into four terms:

$$\begin{aligned} \mathbf{T}_1 \;&\doteq \mathbf{e}_1 \mathbf{e}_1^* \; \mathbf{\Delta} \; \mathbf{e}_1 \mathbf{e}_1^* \; \mathbf{Q} \; \mathbf{P}_{\Omega^1} \mathbf{x}^{1^*} \mathbf{x}^1 \mathbf{P}_{\Omega^1}, && \text{(E.43)} \\ \mathbf{T}_2 \;&\doteq\; \sum_{b=2}^{n} \; \mathbf{e}_1 \mathbf{e}_1^* \; \mathbf{\Delta} \; \mathbf{e}_b \mathbf{e}_b^* \; \mathbf{Q} \; \mathbf{P}_{\Omega^b} \mathbf{x}^{1^*} \mathbf{x}^1 \mathbf{P}_{\Omega^1}, && \text{(E.44)} \\ \mathbf{T}_3 \;&\doteq\; \sum_{a=2}^{n} \; \mathbf{e}_a \mathbf{e}_a^* \; \mathbf{\Delta} \; \mathbf{e}_1 \mathbf{e}_1^* \; \mathbf{Q} \; \mathbf{P}_{\Omega^1} \mathbf{x}^{1^*} \mathbf{x}^1 \mathbf{P}_{\Omega^a}, && \text{(E.45)} \\ \mathbf{T}_4 \;&\doteq\; \sum_{a=2}^{n} \sum_{b=2}^{n} \; \mathbf{e}_a \mathbf{e}_a^* \; \mathbf{\Delta} \; \mathbf{e}_b \mathbf{e}_b^* \; \mathbf{Q} \; \mathbf{P}_{\Omega^b} \mathbf{x}^{1^*} \mathbf{x}^1 \mathbf{P}_{\Omega^a}, && \text{(E.46)} \end{aligned}$$

and thus

$$\mathcal{W}[\mathbf{Q}] \;=\; \mathbf{T}_1 + \mathbf{T}_2 + \mathbf{T}_3 + \mathbf{T}_4. \tag{E.47}$$

Since $\mathbf{e}_1^* \mathbf{\Delta} \mathbf{e}_1 \;=\; \mathbf{A}_1^* \mathbf{A}_1 - (\mathbf{A}_1^* \mathbf{A}_1)^2 \;=\; 0$, the first term $\mathbf{T}_1 = \mathbf{0}$. We will show that on $\mathcal{E}_1$, we can bound $\mathbf{T}_2, \mathbf{T}_3$ and $\mathbf{T}_4$ by:

$$\|\mathbf{T}_2\|_F \leq 8\mu(\mathbf{A})\sqrt{k}\|\mathbf{Q}\|_F, \tag{E.48}$$

$$\|\mathbf{T}_3\|_F \leq 8\mu(\mathbf{A})\sqrt{k}\|\mathbf{Q}\|_F, \tag{E.49}$$

$$\|\mathbf{T}_4\|_F \leq 8\mu(\mathbf{A})k\|\mathbf{Q}\|_F. \tag{E.50}$$

Hence, on $\mathcal{E}_1$,

$$\|\mathcal{W}[\mathbf{Q}]\|_F \leq \left(16\mu(\mathbf{A})\sqrt{k} + 8\mu(\mathbf{A})k\right) \|\mathbf{Q}\|_F. \tag{E.51}$$

Therefore, $\|\mathbf{W}\| \leq 16\mu(\mathbf{A})\sqrt{k} + 8\mu(\mathbf{A})k \leq 24k\mu(\mathbf{A})$. Combining this with (E.41), we get the desired result (E.15).

All that remains to do is prove the three inequalities (E.48), (E.48) and (E.50). We establish these bounds in the following.

(i) Establishing (E.48)**.** For the term $\mathbf{T}_2$ defined in (E.44),

$$\begin{aligned}
\|\mathbf{T}_2\|_F &= \Big\|\mathbf{e}_1^* \boldsymbol{\Delta} \big(\sum_b \mathbf{e}_b\mathbf{e}_b^* \,\mathbf{Q}\,\mathbf{P}_{\Omega^b}\mathbf{x}^{1^*}\big)\mathbf{x}^1\mathbf{P}_{\Omega^1}\Big\|_2 \\
&\leq \Big\|\mathbf{e}_1^*\boldsymbol{\Delta}\Big\|_2 \Big\|\sum_b \mathbf{e}_b\mathbf{e}_b^* \,\mathbf{Q}\,\mathbf{P}_{\Omega^b}\mathbf{x}^{1^*}\Big\|_2 \Big\|\mathbf{x}^1\mathbf{P}_{\Omega^1}\Big\|_2 && \text{(E.52)} \\
&\leq \sqrt{n}\|\mathbf{e}_1^*\boldsymbol{\Delta}\|_\infty \times \Big\|\sum_b \mathbf{e}_b\mathbf{e}_b^* \,\mathbf{Q}\,\mathbf{P}_{\Omega^b}\mathbf{x}^{1^*}\Big\|_2 \times 2 && \text{(E.53)} \\
&\leq 4\sqrt{n}\,\mu(\mathbf{A}) \Big\|\sum_b \mathbf{e}_b\mathbf{e}_b^* \,\mathbf{Q}\,\mathbf{P}_{\Omega^b}\mathbf{x}^{1^*}\Big\|_2. && \text{(E.54)}
\end{aligned}$$

For the last term in (E.54), we have

$$\begin{aligned}
\big\|\sum_b \mathbf{e}_b\mathbf{e}_b^* \,\mathbf{Q}\,\mathbf{P}_{\Omega^b}\mathbf{x}^{1^*}\big\|_2^2 &= \sum_b (\mathbf{e}_b^* \,\mathbf{Q}\,\mathbf{P}_{\Omega^b}\mathbf{x}^{1^*})^2 \\
&\leq \sum_b \|\mathbf{e}_b^*\mathbf{Q}\|_2^2 \|\mathbf{x}^1\mathbf{P}_{\Omega^b}\|_2^2 \\
&\leq 4k\,\|\mathbf{Q}\|_F^2\,/n. && \text{(E.55)}
\end{aligned}$$

Combining (E.54) and (E.55) establishes (E.48).

(ii) Establishing (E.49)**.** For $\mathbf{T}_3$ defined in (E.45), we have

$$\begin{aligned}
\|\mathbf{T}_3\|_F^2 &= \sum_{a=2}^{n} (\mathbf{e}_a^* \boldsymbol{\Delta} \mathbf{e}_1)^2 \left\| \mathbf{e}_1^* \mathbf{Q} \;\; \mathbf{P}_{\Omega^1} \mathbf{x}^{1^*} \;\; \mathbf{x}^1 \mathbf{P}_{\Omega^a} \right\|^2 \\
&\leq 4\mu^2(\mathbf{A}) \sum_{a=2}^{n} \|\mathbf{e}_1^* \mathbf{Q}\|^2 \|\mathbf{x}^1 \mathbf{P}_{\Omega^1}\|^2 \|\mathbf{x}^1 \mathbf{P}_{\Omega^a}\|^2 && \text{(E.56)} \\
&\leq 4\mu^2(\mathbf{A}) \times 4 \times 4(k/n) \times (n-1) \|\mathbf{e}_1^* \mathbf{Q}\|^2 && \text{(E.57)} \\
&\leq 64\mu^2(\mathbf{A}) \frac{k(n-1)}{n} \|\mathbf{Q}\|^2, && \text{(E.58)}
\end{aligned}$$

where in (E.56) we have used the bound $\|\boldsymbol{\Delta}\|_\infty \leq 2\mu(\mathbf{A})$ and the Cauchy-Schwarz inequality. Thus we have established (E.49).

(iii) Establishing (E.50)**.** Express $\|\mathbf{T}_4\|_F^2$ as a sum of squared ℓ_2-norms of the rows of $\mathbf{T}_4$:

$$\begin{aligned}
\|\mathbf{T}_4\|_F^2 &= \sum_{a=2}^{n} \left\| \sum_{b=2}^{n} \mathbf{e}_a^* \boldsymbol{\Delta} \mathbf{e}_b \, \mathbf{e}_b^* \mathbf{Q} \mathbf{P}_{\Omega^b} \mathbf{x}^{1^*} \times \mathbf{x}^1 \mathbf{P}_{\Omega^a} \right\|^2 \\
&= \sum_{a=2}^{n} \|\mathbf{x}^1 \mathbf{P}_{\Omega^a}\|^2 \left(\sum_{b=2}^{n} \mathbf{e}_a^* \boldsymbol{\Delta} \mathbf{e}_b \times \mathbf{e}_b^* \mathbf{Q} \mathbf{P}_{\Omega^b} \mathbf{x}^{1^*} \right)^2 && \text{(E.59)} \\
&\leq 4(k/n) \times \sum_{a=2}^{n} \left(\mathbf{e}_a^* \boldsymbol{\Delta} \sum_{b=2}^{n} \mathbf{e}_b \mathbf{e}_b^* \mathbf{Q} \mathbf{P}_{\Omega^b} \mathbf{x}^{1^*} \right)^2 && \text{(E.60)} \\
&\leq 4(k/n) \times \sum_{a=2}^{n} \|\mathbf{e}_a^* \boldsymbol{\Delta}\|_2^2 \left\| \sum_{b=2}^{n} \mathbf{e}_b \mathbf{e}_b^* \mathbf{Q} \mathbf{P}_{\Omega^b} \mathbf{x}^{1^*} \right\|_2^2 && \text{(E.61)} \\
&\leq 4(k/n) \times \|\boldsymbol{\Delta}\|_F^2 \times \left\| \sum_{b=2}^{n} \mathbf{e}_b \mathbf{e}_b^* \mathbf{Q} \mathbf{P}_{\Omega^b} \mathbf{x}^{1^*} \right\|_2^2 && \text{(E.62)} \\
&\leq 4(k/n) \times n^2 \|\boldsymbol{\Delta}\|_\infty^2 \times \sum_{b=2}^{n} \left(\mathbf{e}_b^* \mathbf{Q} \mathbf{P}_{\Omega^b} \mathbf{x}^{1^*} \right)^2 && \text{(E.63)} \\
&\leq 16\, k\, n\, \mu^2(\mathbf{A}) \times \sum_{b=2}^{n} \|\mathbf{e}_b^* \mathbf{Q}\|_2^2 \|\mathbf{P}_{\Omega^b} \mathbf{x}^{1^*}\|_2^2 && \text{(E.64)} \\
&\leq 64\, k^2\, \mu^2(\mathbf{A}) \times \sum_{b=2}^{n} \|\mathbf{e}_b^* \mathbf{Q}\|^2. && \text{(E.65)}
\end{aligned}$$

Bound the summation $\sum_{b=2}^{n} \|\mathbf{e}_b^* \mathbf{Q}\|^2$ by $\|\mathbf{Q}\|_F^2$, and then we get (E.50). This completes the proof of Lemma E.3. □

REFERENCES

[1] G. Wallace, "The JPEG still picture compression standard," *Communications of the ACM*, vol. 34, no. 4, pp. 30–44, 1991.

[2] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Review*, vol. 51, no. 1, pp. 34–81, 2009.

[3] E. Candès and T. Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[4] O. Bryt and M. Elad, "Compression of facial images using the K-SVD algorithm," *Journal of Visual Communication and Image Representation*, vol. 19, no. 4, pp. 270–283, 2008.

[5] R. Gribonval and K. Schnass, "Dictionary identification - sparse matrix factorization via ℓ_1-minimization," *IEEE Transactions on Information Theory*, vol. 56, no. 7, pp. 3523–3539, 2010.

[6] Q. Geng, H. Wang, and J. Wright, "Algorithms for exact dictionary learning by ℓ_1-minimization," *Technical Report*, 2011.

[7] E. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, pp. 717–772, 2008.

[8] D. Gross, "Recovering low-rank matrices from a few coefficients in any basis," 2009, available at http://arxiv.org/abs/0910.1879.

[9] J. Tropp, "User-friendly tail bounds for matrix martingales," 2010, available at http://arxiv.org/abs/1004.4389v4.

[10] J. Wright, A. Y. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, pp. 210–227, February 2009.

[11] M. Aharon, M. Elad, and A. Bruckstein, "On the uniqueness of overcomplete dictionaries and a practical way to retrieve them," *Journal of Linear Algebra and Applications*, vol. 416, pp. 48–67, 2006.

[12] J. Fuchs, "On sparse representations in arbitrary redundant bases," *IEEE Transactions on Information Theory*, vol. 50, no. 6, pp. 1341–1344, 2004.

[13] R. Ahlswede and A. Winter, "Strong converse for identification via quantum channels," *IEEE Transactions on Information Theory*, vol. 48, no. 3, pp. 569–579, 2002.

[14] R. Latala and K. Oleszkiewicz, "On the best constant in the Khintchine-Kahane inequality," *Studia Mathematica*, vol. 109, no. 1, pp. 101–104, 1994.

[15] J. Kahane, "Sur les sommes vectorielles $\sum \pm u_n$," *Comptes Rendus Mathematique*, vol. 259, pp. 2577–2580, 1964.

[16] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: Isoperimetry and Processes*. New York: Springer-Verlag, 1991.

[17] J. Tropp, "Norms of random submatrices and sparse approximation," *Comptes Rendus Mathematique*, vol. 346, pp. 1271–1274, 2008.

[18] M. Ledoux, *The Concentration of Measure Phenomenon*, ser. Mathematical Surveys and Monographs. Providence, RI: American Mathematical Society, 2001, vol. 89.