Foundations of Computational Mathematics (2020) 20:451–632
https://doi.org/10.1007/s10208-019-09429-9

FOUNDATIONS OF COMPUTATIONAL MATHEMATICS
The Journal of the Society for the Foundations of Computational Mathematics

Implicit Regularization in Nonconvex Statistical Estimation: Gradient Descent Converges Linearly for Phase Retrieval, Matrix Completion, and Blind Deconvolution

Cong Ma[1] · Kaizheng Wang[1] · Yuejie Chi[2] · Yuxin Chen[3]

Received: 14 December 2017 / Revised: 8 May 2019 / Accepted: 18 June 2019 / Published online: 5 August 2019

Abstract
Recent years have seen a flurry of activities in designing provably efficient nonconvex procedures for solving statistical estimation problems. Due to the highly nonconvex nature of the empirical loss, state-of-the-art procedures often require proper regularization (e.g., trimming, regularized cost, projection) in order to guarantee fast convergence. For vanilla procedures such as gradient descent, however, prior theory either recommends highly conservative learning rates to avoid overshooting, or completely lacks performance guarantees. This paper uncovers a striking phenomenon in nonconvex optimization: even in the absence of explicit regularization, gradient descent enforces proper regularization implicitly under various statistical models. In fact, gradient descent follows a trajectory staying within a basin that enjoys nice geometry, consisting of points incoherent with the sampling mechanism. This "implicit regularization" feature allows gradient descent to proceed in a far more aggressive fashion without overshooting, which in turn results in substantial computational savings. Focusing on three fundamental statistical estimation problems, i.e., phase retrieval, low-rank matrix completion, and blind deconvolution, we establish that gradient descent achieves near-optimal statistical and computational guarantees without explicit regularization. In particular, by marrying statistical modeling with generic optimization theory, we develop a general recipe for analyzing the trajectories of iterative algorithms via a leave-one-out perturbation argument. As a by-product, for noisy matrix completion, we demonstrate that gradient descent achieves near-optimal error control—measured entrywise and by the spectral norm—which might be of independent interest.

Keywords Nonconvex optimization · Gradient descent · Leave-one-out analysis · Phase retrieval · Matrix completion · Blind deconvolution

Communicated by Emmanuel J. Candès.

Extended author information available on the last page of the article

Mathematics Subject Classification 90C26

1 Introduction

1.1 Nonlinear Systems and Empirical Loss Minimization

A wide spectrum of science and engineering applications calls for solutions to a nonlinear system of equations. Imagine we have collected a set of data points $\boldsymbol{y} = \{y_j\}_{1\le j\le m}$, generated by a nonlinear sensing system,

$$y_j \approx \mathcal{A}_j(\boldsymbol{x}^\star), \quad 1 \le j \le m,$$

where $\boldsymbol{x}^\star$ is the unknown object of interest and the $\mathcal{A}_j$'s are certain nonlinear maps known *a priori*. Can we reconstruct the underlying object $\boldsymbol{x}^\star$ in a faithful yet efficient manner? Problems of this kind abound in information and statistical science, prominent examples including low-rank matrix recovery [19,64], robust principal component analysis [17,21], phase retrieval [20,59], neural networks [103,132], to name just a few.

In principle, it is possible to attempt reconstruction by searching for a solution that minimizes the empirical loss, namely

$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}) = \sum_{j=1}^{m} \left|y_j - \mathcal{A}_j(\boldsymbol{x})\right|^2. \tag{1}$$

Unfortunately, this empirical loss minimization problem is, in many cases, nonconvex, making it NP-hard in general. This issue of nonconvexity comes up in, for example, several representative problems that epitomize the structures of nonlinear systems encountered in practice.[1]

- *Phase retrieval/solving quadratic systems of equations* Imagine we are asked to recover an unknown object $\boldsymbol{x}^\star \in \mathbb{R}^n$, but are only given the square modulus of certain linear measurements about the object, with all sign/phase information of the measurements missing. This arises, for example, in X-ray crystallography [15], and in latent-variable models where the hidden variables are captured by the missing signs [33]. To fix ideas, assume we would like to solve for $\boldsymbol{x}^\star \in \mathbb{R}^n$ in the following quadratic system of m equations

$$y_j = \left(\boldsymbol{a}_j^\top \boldsymbol{x}^\star\right)^2, \qquad 1 \le j \le m,$$

[1] Here, we choose different pre-constants in front of the empirical loss in order to be consistent with the literature of the respective problems. In addition, we only introduce the problem in the noiseless case for simplicity of presentation.

where $\{\boldsymbol{a}_j\}_{1\le j\le m}$ are the known design vectors. One strategy is thus to solve the following problem

$$\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} \quad f(\boldsymbol{x}) = \frac{1}{4m}\sum_{j=1}^{m}\left[y_j - \left(\boldsymbol{a}_j^\top \boldsymbol{x}\right)^2\right]^2. \tag{2}$$

- *Low-rank matrix completion* In many scenarios such as collaborative filtering, we wish to make predictions about all entries of an (approximately) low-rank matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n\times n}$ (e.g., a matrix consisting of users' ratings about many movies), yet only a highly incomplete subset of the entries are revealed to us [19]. For clarity of presentation, assume $\boldsymbol{M}^\star$ to be rank-r ($r \ll n$) and positive semidefinite (PSD), i.e., $\boldsymbol{M}^\star = \boldsymbol{X}^\star\boldsymbol{X}^{\star\top}$ with $\boldsymbol{X}^\star \in \mathbb{R}^{n\times r}$, and suppose we have only seen the entries

$$Y_{j,k} = M^\star_{j,k} = (\boldsymbol{X}^\star\boldsymbol{X}^{\star\top})_{j,k}, \qquad (j,k)\in\Omega$$

within some index subset Ω of cardinality m. These entries can be viewed as nonlinear measurements about the low-rank factor $\boldsymbol{X}^\star$. The task of completing the true matrix $\boldsymbol{M}^\star$ can then be cast as solving

$$\text{minimize}_{\boldsymbol{X}\in\mathbb{R}^{n\times r}} \quad f(\boldsymbol{X}) = \frac{n^2}{4m}\sum_{(j,k)\in\Omega}\left(Y_{j,k} - \boldsymbol{e}_j^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{e}_k\right)^2, \tag{3}$$

where the $\boldsymbol{e}_j$'s stand for the canonical basis vectors in $\mathbb{R}^n$.
- *Blind deconvolution/solving bilinear systems of equations* Imagine we are interested in estimating two signals of interest $\boldsymbol{h}^\star, \boldsymbol{x}^\star \in \mathbb{C}^K$, but only get to collect a few bilinear measurements about them. This problem arises from mathematical modeling of blind deconvolution [3,76], which frequently arises in astronomy, imaging, communications, etc. The goal is to recover two signals from their convolution. Put more formally, suppose we have acquired m bilinear measurements taking the following form

$$y_j = \boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j, \qquad 1\le j\le m,$$

where $\boldsymbol{a}_j, \boldsymbol{b}_j \in \mathbb{C}^K$ are distinct design vectors (e.g., Fourier and/or random design vectors) known *a priori* and $\boldsymbol{b}_j^{\mathsf{H}}$ denotes the conjugate transpose of $\boldsymbol{b}_j$. In order to reconstruct the underlying signals, one asks for solutions to the following problem

$$\text{minimize}_{\boldsymbol{h},\boldsymbol{x}\in\mathbb{C}^K} \quad f(\boldsymbol{h},\boldsymbol{x}) = \sum_{j=1}^{m}\left|y_j - \boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}\boldsymbol{a}_j\right|^2.$$

1.2 Nonconvex Optimization via Regularized Gradient Descent

First-order methods have been a popular heuristic in practice for solving nonconvex problems including (1). For instance, a widely adopted procedure is gradient descent, which follows the update rule

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_t \nabla f\left(\boldsymbol{x}^t\right), \qquad t \geq 0, \tag{4}$$

where η_t is the learning rate (or step size) and $\boldsymbol{x}^0$ is some proper initial guess. Given that it only performs a single gradient calculation $\nabla f(\cdot)$ per iteration (which typically can be completed within near-linear time), this paradigm emerges as a candidate for solving large-scale problems. The concern is: whether $\boldsymbol{x}^t$ converges to the global solution and, if so, how long it takes for convergence, especially since (1) is highly nonconvex.

Fortunately, despite the worst-case hardness, appealing convergence properties have been discovered in various statistical estimation problems, the blessing being that the statistical models help rule out ill-behaved instances. For the average case, the empirical loss often enjoys benign geometry, in a *local* region (or at least along certain directions) surrounding the global optimum. In light of this, an effective nonconvex iterative method typically consists of two stages:

1. a carefully designed initialization scheme (e.g., spectral method);
2. an iterative refinement procedure (e.g., gradient descent).

This strategy has recently spurred a great deal of interest, owing to its promise of achieving computational efficiency and statistical accuracy at once for a growing list of problems (e.g., [18,25,32,61,64,76,78,107]). However, rather than directly applying gradient descent (4), existing theory often suggests enforcing proper regularization. Such explicit regularization enables improved computational convergence by properly "stabilizing" the search directions. The following regularization schemes, among others, have been suggested to obtain or improve computational guarantees. We refer to these algorithms collectively as *Regularized Gradient Descent*.

- *Trimming/truncation*, which discards/truncates a subset of the gradient components when forming the descent direction. For instance, when solving quadratic systems of equations, one can modify the gradient descent update rule as

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_t \mathcal{T}\left(\nabla f\left(\boldsymbol{x}^t\right)\right), \tag{5}$$

  where $\mathcal{T}$ is an operator that effectively drops samples bearing too much influence on the search direction. This strategy [25,118,126] has been shown to enable exact recovery with linear-time computational complexity and optimal sample complexity.
- *Regularized loss*, which attempts to optimize a regularized empirical risk

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_t \left(\nabla f\left(\boldsymbol{x}^t\right) + \nabla R\left(\boldsymbol{x}^t\right)\right), \tag{6}$$

  where $R(\boldsymbol{x})$ stands for an additional penalty term in the empirical loss. For example, in low-rank matrix completion $R(\cdot)$ imposes penalty based on the ℓ_2 row norm [64,107] as well as the Frobenius norm [107] of the decision matrix, while in blind deconvolution, it penalizes the ℓ_2 norm as well as certain component-wise incoherence measure of the decision vectors [58,76,82].

Table 1 Prior theory for gradient descent (with spectral initialization)

	Vanilla gradient descent			Regularized gradient descent		
	Sample complexity	Iteration complexity	Step size	Sample complexity	Iteration complexity	Type of regularization
Phase retrieval	$n \log n$	$n \log \frac{1}{\epsilon}$	$\frac{1}{n}$	n	$\log \frac{1}{\epsilon}$	Trimming [25,126]
Matrix completion	n/a	n/a	n/a	nr^7	$\frac{n}{r} \log \frac{1}{\epsilon}$	Regularized loss [107]
				nr^2	$r^2 \log \frac{1}{\epsilon}$	Projection [32,131]
Blind deconvolution	n/a	n/a	n/a	$K \operatorname{poly} \log m$	$m \log \frac{1}{\epsilon}$	Regularized loss and projection [76]

- *Projection*, which projects the iterates onto certain sets based on prior knowledge, that is,

$$\boldsymbol{x}^{t+1} = \mathcal{P}\left(\boldsymbol{x}^t - \eta_t \nabla f\left(\boldsymbol{x}^t\right)\right), \tag{7}$$

  where $\mathcal{P}$ is a certain projection operator used to enforce, for example, incoherence properties. This strategy has been employed in both low-rank matrix completion [32,131] and blind deconvolution [76].

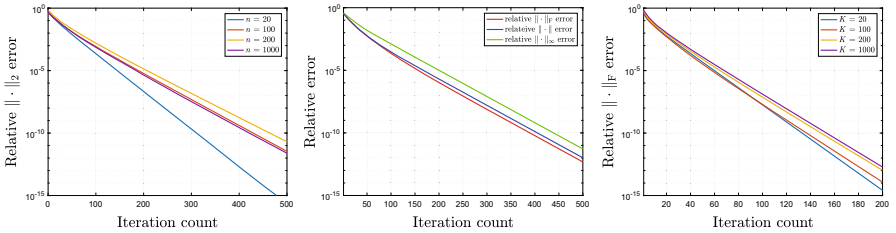
Equipped with such regularization procedures, existing works uncover appealing computational and statistical properties under various statistical models. Table 1 summarizes the performance guarantees derived in the prior literature; for simplicity, only orderwise results are provided.

Remark 1 There is another role of regularization commonly studied in the literature, which exploits prior knowledge about the structure of the unknown object, such as sparsity to prevent over-fitting and improve statistical generalization ability. This is, however, not the focal point of this paper, since we are primarily pursuing solutions to (1) without imposing additional structures.

1.3 Regularization-Free Procedures?

The regularized gradient descent algorithms, while exhibiting appealing performance, usually introduce more algorithmic parameters that need to be carefully tuned based on the assumed statistical models. In contrast, vanilla gradient descent (cf. (4))—which is perhaps the very first method that comes into mind and requires minimal tuning parameters—is far less understood (cf. Table 1). Take matrix completion and blind deconvolution as examples: to the best of our knowledge, there is currently no theoretical guarantee derived for vanilla gradient descent.

The situation is better for phase retrieval: the local convergence of vanilla gradient descent, also known as Wirtinger flow (WF), has been investigated in [18,96]. Under i.i.d. Gaussian design and with near-optimal sample complexity, WF (combined with spectral initialization) provably achieves ϵ-accuracy (in a relative sense)



(a) phase retrieval **(b)** matrix completion **(c)** blind deconvolution

Fig. 1 **a** Relative ℓ_2 error of $\boldsymbol{x}^t$ (modulo the global phase) versus iteration count for phase retrieval under i.i.d. Gaussian design, where $m = 10n$ and $\eta_t = 0.1$. **b** Relative error of $\boldsymbol{X}^t\boldsymbol{X}^{t\top}$ (measured by $\|\cdot\|_{\mathrm{F}}, \|\cdot\|, \|\cdot\|_\infty$) versus iteration count for matrix completion, where $n = 1000$, $r = 10$, $p = 0.1$, and $\eta_t = 0.2$. **c** Relative error of $\boldsymbol{h}^t\boldsymbol{x}^{t\mathsf{H}}$ (measured by $\|\cdot\|_{\mathrm{F}}$) versus iteration count for blind deconvolution, where $m = 10K$ and $\eta_t = 0.5$

within $O\big(n\log(1/\varepsilon)\big)$ iterations. Nevertheless, the computational guarantee is significantly outperformed by the regularized version (called truncated Wirtinger flow [25]), which only requires $O\big(\log(1/\varepsilon)\big)$ iterations to converge with similar per-iteration cost. On closer inspection, the high computational cost of WF is largely due to the vanishingly small step size $\eta_t = O\big(1/(n\|\boldsymbol{x}^\star\|_2^2)\big)$—and hence slow movement—suggested by the theory [18]. While this is already the largest possible step size allowed in the theory published in [18], it is considerably more conservative than the choice $\eta_t = O\big(1/\|\boldsymbol{x}^\star\|_2^2\big)$ theoretically justified for the regularized version [25,126].

The lack of understanding and suboptimal results about vanilla gradient descent raise a very natural question: *Are regularization-free iterative algorithms inherently suboptimal when solving nonconvex statistical estimation problems of this kind?*

1.4 Numerical Surprise of Unregularized Gradient Descent

To answer the preceding question, it is perhaps best to first collect some numerical evidence. In what follows, we test the performance of vanilla gradient descent for phase retrieval, matrix completion, and blind deconvolution, using a *constant* step size. For all of these experiments, the initial guess is obtained by means of the standard spectral method. Our numerical findings are as follows:

- *Phase retrieval* For each n, set $m = 10n$, take $\boldsymbol{x}^\star \in \mathbb{R}^n$ to be a random vector with unit norm, and generate the design vectors $\boldsymbol{a}_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, $1 \le j \le m$. Figure 1a illustrates the relative ℓ_2 error $\min\{\|\boldsymbol{x}^t - \boldsymbol{x}^\star\|_2, \|\boldsymbol{x}^t + \boldsymbol{x}^\star\|_2\}/\|\boldsymbol{x}^\star\|_2$ (modulo the unrecoverable global phase) versus the iteration count. The results are shown for $n = 20, 100, 200, 1000$, with the step size taken to be $\eta_t = 0.1$ in all settings.
- *Matrix completion* Generate a random PSD matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n\times n}$ with dimension $n = 1000$, rank $r = 10$, and all nonzero eigenvalues equal to one. Each entry of $\boldsymbol{M}^\star$ is observed independently with probability $p = 0.1$. Figure 1b plots the relative error $\big|\big|\big|\boldsymbol{X}^t\boldsymbol{X}^{t\top} - \boldsymbol{M}^\star\big|\big|\big| / \big|\big|\big|\boldsymbol{M}^\star\big|\big|\big|$ versus the iteration count, where $|||\cdot|||$ can either be the Frobenius norm $\|\cdot\|_{\mathrm{F}}$, the spectral norm $\|\cdot\|$, or the entrywise ℓ_∞ norm $\|\cdot\|_\infty$. Here, we pick the step size as $\eta_t = 0.2$.

- *Blind deconvolution* For each $K \in \{20, 100, 200, 1000\}$ and $m = 10K$, generate the design vectors $\boldsymbol{a}_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \frac{1}{2}\boldsymbol{I}_K) + i\mathcal{N}(\boldsymbol{0}, \frac{1}{2}\boldsymbol{I}_K)$ for $1 \le j \le m$ independently,[2] and the $\boldsymbol{b}_j$'s are drawn from a partial discrete Fourier transform (DFT) matrix (to be described in Sect. 3.3). The underlying signals $\boldsymbol{h}^\star, \boldsymbol{x}^\star \in \mathbb{C}^K$ are produced as random vectors with unit norm. Figure 1c plots the relative error $\|\boldsymbol{h}^t\boldsymbol{x}^{t\mathsf{H}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\|_{\mathrm{F}}/\|\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\|_{\mathrm{F}}$ versus the iteration count, with the step size taken to be $\eta_t = 0.5$ in all settings.

In all of these numerical experiments, vanilla gradient descent enjoys remarkable linear convergence, always yielding an accuracy of 10^{-5} (in a relative sense) within around 200 iterations. In particular, for the phase retrieval problem, the step size is taken to be $\eta_t = 0.1$ although we vary the problem size from $n = 20$ to $n = 1000$. The consequence is that the convergence rates experience little changes when the problem sizes vary. In comparison, the theory published in [18] seems overly pessimistic, as it suggests a diminishing step size inversely proportional to n and, as a result, an iteration complexity that worsens as the problem size grows.

In addition, it has been empirically observed in prior literature [25,76,127] that vanilla gradient descent performs comparably with the regularized counterpart for phase retrieval and blind deconvolution. To complete the picture, we further conduct experiments on matrix completion. In particular, we follow the experimental setup for matrix completion used above. We vary p from 0.01 to 0.1 with 51 logarithmically spaced points. For each p, we apply vanilla gradient descent, projected gradient descent [32] and gradient descent with additional regularization terms [107] with step size $\eta = 0.2$ to 50 randomly generated instances. Successful recovery is declared if $\|\boldsymbol{X}^t\boldsymbol{X}^{t\top} - \boldsymbol{M}^\star\|_{\mathrm{F}}/\|\boldsymbol{M}^\star\|_{\mathrm{F}} \le 10^{-5}$ in 10^4 iterations. Figure 2 reports the success rate versus the sampling rate. As can be seen, the phase transition of vanilla GD and that of GD with regularized cost are almost identical, whereas projected GD performs slightly better than the other two.

In short, the above empirical results are surprisingly positive yet puzzling. Why was the computational efficiency of vanilla gradient descent unexplained or substantially underestimated in prior theory?

1.5 This Paper

The main contribution of this paper is toward demystifying the "unreasonable" effectiveness of regularization-free nonconvex iterative methods. As asserted in previous work, regularized gradient descent succeeds by properly enforcing/promoting certain incoherence conditions throughout the execution of the algorithm. In contrast, we discover that

> *Vanilla gradient descent automatically forces the iterates to stay incoherent with the measurement mechanism, thus implicitly regularizing the search directions.*

This "implicit regularization" phenomenon is of fundamental importance, suggesting that vanilla gradient descent proceeds as if it were properly regularized. This

[2] Here and throughout, i represents the imaginary unit.

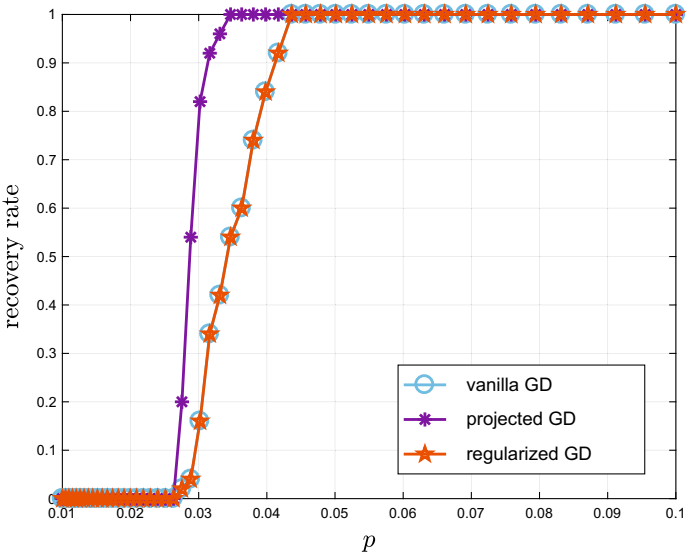


Fig. 2 Success rate versus sampling rate p over 50 Monte Carlo trials for matrix completion with $n = 1000$ and $r = 10$

explains the remarkably favorable performance of unregularized gradient descent in practice. Focusing on the three representative problems mentioned in Sect. 1.1, our theory guarantees both statistical and computational efficiency of vanilla gradient descent under random designs and spectral initialization. With near-optimal sample complexity, to attain ϵ-accuracy,

- *Phase retrieval (informal)* vanilla gradient descent converges in $O\big(\log n \log \frac{1}{\epsilon}\big)$ iterations;
- *Matrix completion (informal)* vanilla gradient descent converges in $O\big(\log \frac{1}{\epsilon}\big)$ iterations;
- *Blind deconvolution (informal)* vanilla gradient descent converges in $O\big(\log \frac{1}{\epsilon}\big)$ iterations.

In other words, gradient descent provably achieves (nearly) linear convergence in all of these examples. Throughout this paper, an algorithm is said to *converge (nearly) linearly* to $\boldsymbol{x}^\star$ in the noiseless case if the iterates $\{\boldsymbol{x}^t\}$ obey

$$\operatorname{dist}(\boldsymbol{x}^{t+1}, \boldsymbol{x}^\star) \le (1 - c)\operatorname{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\star), \qquad \forall t \ge 0$$

for some $0 < c \le 1$ that is (almost) independent of the problem size. Here, $\operatorname{dist}(\cdot, \cdot)$ can be any appropriate discrepancy measure.

As a by-product of our theory, gradient descent also provably controls the *entrywise* empirical risk uniformly across all iterations; for instance, this implies that vanilla gradient descent controls entrywise estimation error for the matrix completion task. Precise statements of these results are deferred to Sect. 3 and are briefly summarized in Table 2.

Table 2 Prior theory versus our theory for vanilla gradient descent (with spectral initialization)

	Prior theory			Our theory		
	Sample complexity	Iteration complexity	Step size	Sample complexity	Iteration complexity	Step size
Phase retrieval	$n \log n$	$n \log(1/\varepsilon)$	$1/n$	$n \log n$	$\log n \log(1/\varepsilon)$	$1/\log n$
Matrix completion	n/a	n/a	n/a	$nr^3 \operatorname{poly} \log n$	$\log(1/\varepsilon)$	1
Blind deconvolution	n/a	n/a	n/a	$K \operatorname{poly} \log m$	$\log(1/\varepsilon)$	1

Notably, our study of implicit regularization suggests that the behavior of *nonconvex optimization* algorithms for statistical estimation needs to be examined in the context of *statistical models*, which induces an objective function as a finite sum. Our proof is accomplished via a leave-one-out perturbation argument, which is inherently tied to statistical models and leverages homogeneity across samples. Altogether, this allows us to localize benign landscapes for optimization and characterize finer dynamics not accounted for in generic gradient descent theory.

1.6 Notations

Before continuing, we introduce several notations used throughout the paper. First of all, boldfaced symbols are reserved for vectors and matrices. For any vector $\boldsymbol{v}$, we use $\|\boldsymbol{v}\|_2$ to denote its Euclidean norm. For any matrix $\boldsymbol{A}$, we use $\sigma_j(\boldsymbol{A})$ and $\lambda_j(\boldsymbol{A})$ to denote its jth largest singular value and eigenvalue, respectively, and let $\boldsymbol{A}_{j,\cdot}$ and $\boldsymbol{A}_{\cdot,j}$ denote its jth row and jth column, respectively. In addition, $\|\boldsymbol{A}\|$, $\|\boldsymbol{A}\|_{\mathrm{F}}$, $\|\boldsymbol{A}\|_{2,\infty}$, and $\|\boldsymbol{A}\|_{\infty}$ stand for the spectral norm (i.e., the largest singular value), the Frobenius norm, the ℓ_2/ℓ_∞ norm (i.e., the largest ℓ_2 norm of the rows), and the entrywise ℓ_∞ norm (the largest magnitude of all entries) of a matrix $\boldsymbol{A}$. Also, $\boldsymbol{A}^\top$, $\boldsymbol{A}^{\mathsf{H}}$, and $\overline{\boldsymbol{A}}$ denote the transpose, the conjugate transpose, and the entrywise conjugate of $\boldsymbol{A}$, respectively. $\boldsymbol{I}_n$ denotes the identity matrix with dimension $n \times n$. The notation $\mathcal{O}^{n\times r}$ represents the set of all $n \times r$ orthonormal matrices. The notation $[n]$ refers to the set $\{1, \cdots, n\}$. Also, we use $\mathrm{Re}(x)$ to denote the real part of a complex number x. Throughout the paper, we use the terms "samples" and "measurements" interchangeably.

Additionally, the standard notation $f(n) = O(g(n))$ or $f(n) \lesssim g(n)$ means that there exists a constant $c > 0$ such that $|f(n)| \le c|g(n)|$, $f(n) \gtrsim g(n)$ means that there exists a constant $c > 0$ such that $|f(n)| \ge c\,|g(n)|$, and $f(n) \asymp g(n)$ means that there exist constants $c_1, c_2 > 0$ such that $c_1|g(n)| \le |f(n)| \le c_2|g(n)|$. Also, $f(n) \gg g(n)$ means that there exists some large enough constant $c > 0$ such that $|f(n)| \ge c\,|g(n)|$. Similarly, $f(n) \ll g(n)$ means that there exists some sufficiently small constant $c > 0$ such that $|f(n)| \le c\,|g(n)|$.

2 Implicit Regularization: A Case Study

To reveal reasons behind the effectiveness of vanilla gradient descent, we first examine existing theory of gradient descent and identify the geometric properties that enable

linear convergence. We then develop an understanding as to why prior theory is conservative, and describe the phenomenon of implicit regularization that helps explain the effectiveness of vanilla gradient descent. To facilitate discussion, we will use the problem of solving random quadratic systems (phase retrieval) and Wirtinger flow as a case study, but our diagnosis applies more generally, as will be seen in later sections.

2.1 Gradient Descent Theory Revisited

In the convex optimization literature, there are two standard conditions about the objective function—strong convexity and smoothness—that allow for linear convergence of gradient descent.

Definition 1 *(Strong convexity)* A twice continuously differentiable function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is said to be α-strongly convex for $\alpha > 0$ if

$$\nabla^2 f(\boldsymbol{x}) \succeq \alpha \boldsymbol{I}_n, \qquad \forall \boldsymbol{x} \in \mathbb{R}^n.$$

Definition 2 *(Smoothness)* A twice continuously differentiable function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is said to be β-smooth for $\beta > 0$ if

$$\left\| \nabla^2 f(\boldsymbol{x}) \right\| \leq \beta, \qquad \forall \boldsymbol{x} \in \mathbb{R}^n.$$

It is well known that for an unconstrained optimization problem, if the objective function f is both α-strongly convex and β-smooth, then vanilla gradient descent (4) enjoys ℓ_2 error contraction [9, Theorem 3.12], namely

$$\begin{aligned} \left\| \boldsymbol{x}^{t+1} - \boldsymbol{x}^\star \right\|_2 &\leq \left(1 - \frac{2}{\beta/\alpha + 1} \right) \left\| \boldsymbol{x}^t - \boldsymbol{x}^\star \right\|_2, \quad \text{and} \quad \left\| \boldsymbol{x}^t - \boldsymbol{x}^\star \right\|_2 \\ &\leq \left(1 - \frac{2}{\beta/\alpha + 1} \right)^t \left\| \boldsymbol{x}^0 - \boldsymbol{x}^\star \right\|_2, \quad t \geq 0, \end{aligned} \tag{8}$$

as long as the step size is chosen as $\eta_t = 2/(\alpha + \beta)$. Here, $\boldsymbol{x}^\star$ denotes the global minimum. This immediately reveals the iteration complexity for gradient descent: the number of iterations taken to attain ϵ-accuracy (in a relative sense) is bounded by

$$O\left(\frac{\beta}{\alpha} \log \frac{1}{\epsilon} \right).$$

In other words, the iteration complexity is dictated by and scales linearly with the condition number—the ratio β/α of smoothness to strong convexity parameters.

Moving beyond convex optimization, one can easily extend the above theory to *nonconvex* problems with *local* strong convexity and smoothness. More precisely, suppose the objective function f satisfies

$$\nabla^2 f(\boldsymbol{x}) \succeq \alpha \boldsymbol{I} \qquad \text{and} \qquad \left\| \nabla^2 f(\boldsymbol{x}) \right\| \leq \beta$$

over a local ℓ_2 ball surrounding the global minimum $\boldsymbol{x}^\star$:

$$\mathcal{B}_\delta(\boldsymbol{x}) := \left\{\boldsymbol{x} \mid \|\boldsymbol{x}-\boldsymbol{x}^\star\|_2 \le \delta\|\boldsymbol{x}^\star\|_2\right\}. \tag{9}$$

Then the contraction result (8) continues to hold, as long as the algorithm is seeded with an initial point that falls inside $\mathcal{B}_\delta(\boldsymbol{x})$.

2.2 Local Geometry for Solving Random Quadratic Systems

To invoke generic gradient descent theory, it is critical to characterize the local strong convexity and smoothness properties of the loss function. Take the problem of solving random quadratic systems (phase retrieval) as an example. Consider the i.i.d. Gaussian design in which $\boldsymbol{a}_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, $1 \le j \le m$, and suppose without loss of generality that the underlying signal obeys $\|\boldsymbol{x}^\star\|_2 = 1$. It is well known that $\boldsymbol{x}^\star$ is the unique minimizer—up to global phase—of (2) under this statistical model, provided that the ratio m/n of equations to unknowns is sufficiently large. The Hessian of the loss function $f(\boldsymbol{x})$ is given by

$$\nabla^2 f(\boldsymbol{x}) = \frac{1}{m}\sum_{j=1}^{m}\left[3\left(\boldsymbol{a}_j^\top\boldsymbol{x}\right)^2 - y_j\right]\boldsymbol{a}_j\boldsymbol{a}_j^\top. \tag{10}$$

- *Population-level analysis* Consider the case with an infinite number of equations or samples, i.e., $m \to \infty$, where $\nabla^2 f(\boldsymbol{x})$ converges to its expectation. Simple calculation yields that

$$\mathbb{E}\left[\nabla^2 f(\boldsymbol{x})\right] = 3\left(\|\boldsymbol{x}\|_2^2\boldsymbol{I}_n + 2\boldsymbol{x}\boldsymbol{x}^\top\right) - \left(\boldsymbol{I}_n + 2\boldsymbol{x}^\star\boldsymbol{x}^{\star\top}\right).$$

  It is straightforward to verify that for any sufficiently small constant $\delta > 0$, one has the crude bound

$$\boldsymbol{I}_n \preceq \mathbb{E}\left[\nabla^2 f(\boldsymbol{x})\right] \preceq 10\boldsymbol{I}_n, \qquad \forall \boldsymbol{x} \in \mathcal{B}_\delta(\boldsymbol{x}) : \left\|\boldsymbol{x}-\boldsymbol{x}^\star\right\|_2 \le \delta\left\|\boldsymbol{x}^\star\right\|_2,$$

  meaning that f is 1-strongly convex and 10-smooth within a local ball around $\boldsymbol{x}^\star$. As a consequence, when we have infinite samples and an initial guess $\boldsymbol{x}^0$ such that $\|\boldsymbol{x}^0-\boldsymbol{x}^\star\|_2 \le \delta\left\|\boldsymbol{x}^\star\right\|_2$, vanilla gradient descent with a constant step size converges to the global minimum within logarithmic iterations.
- *Finite-sample regime with $m \asymp n\log n$* Now that f exhibits favorable landscape in the population level, one thus hopes that the fluctuation can be well controlled so that the nice geometry carries over to the finite-sample regime. In the regime where $m \asymp n\log n$ (which is the regime considered in [18]), the local strong convexity is still preserved, in the sense that

$$\nabla^2 f(\boldsymbol{x}) \succeq (1/2)\cdot\boldsymbol{I}_n, \qquad \forall \boldsymbol{x} : \left\|\boldsymbol{x}-\boldsymbol{x}^\star\right\|_2 \le \delta\left\|\boldsymbol{x}^\star\right\|_2$$

occurs with high probability, provided that $\delta > 0$ is sufficiently small (see [96,101] and Lemma 1). The smoothness parameter, however, is not well controlled. In fact, it can be as large as (up to logarithmic factors)[3]

$$\left\|\nabla^2 f(\boldsymbol{x})\right\| \lesssim n$$

even when we restrict attention to the local ℓ_2 ball (9) with $\delta > 0$ being a fixed small constant. This means that the condition number β/α (defined in Sect. 2.1) may scale as $O(n)$, leading to the step size recommendation

$$\eta_t \asymp 1/n,$$

and, as a consequence, a high iteration complexity $O\big(n\log(1/\epsilon)\big)$. This underpins the analysis in [18].

In summary, the geometric properties of the loss function—even in the local ℓ_2 ball centering around the global minimum—are not as favorable as one anticipates, in particular in view of its population counterpart. A direct application of generic gradient descent theory leads to an overly conservative step size and a pessimistic convergence rate, unless the number of samples is enormously larger than the number of unknowns.

Remark 2 Notably, due to Gaussian designs, the phase retrieval problem enjoys more favorable geometry compared to other nonconvex problems. In matrix completion and blind deconvolution, the Hessian matrices are rank-deficient even at the population level. In such cases, the above discussions need to be adjusted, e.g., strong convexity is only possible when we restrict attention to certain directions.

2.3 Which Region Enjoys Nicer Geometry?

Interestingly, our theory identifies a local region surrounding $\boldsymbol{x}^\star$ with a large diameter that enjoys much nicer geometry. This region does not mimic an ℓ_2 ball, but rather, the intersection of an ℓ_2 ball and a polytope. We term it the *region of incoherence and contraction* (RIC). For phase retrieval, the RIC includes all points $\boldsymbol{x} \in \mathbb{R}^n$ obeying

$$\left\|\boldsymbol{x}-\boldsymbol{x}^\star\right\|_2 \le \delta\left\|\boldsymbol{x}^\star\right\|_2 \quad \text{and} \tag{11a}$$

$$\max_{1\le j\le m}\left|\boldsymbol{a}_j^\top\left(\boldsymbol{x}-\boldsymbol{x}^\star\right)\right| \lesssim \sqrt{\log n}\left\|\boldsymbol{x}^\star\right\|_2, \tag{11b}$$

where $\delta > 0$ is some small numerical constant. As will be formalized in Lemma 1, with high probability the Hessian matrix satisfies

$$(1/2)\cdot\boldsymbol{I}_n \preceq \nabla^2 f(\boldsymbol{x}) \preceq O(\log n)\cdot\boldsymbol{I}_n$$

[3] To demonstrate this, taking $\boldsymbol{x} = \boldsymbol{x}^\star + (\delta/\|\boldsymbol{a}_1\|_2)\cdot\boldsymbol{a}_1$ in (10), one can easily verify that, with high probability, $\left\|\nabla^2 f(\boldsymbol{x})\right\| \ge \left|3(\boldsymbol{a}_1^\top\boldsymbol{x})^2 - y_1\right|\left\|\boldsymbol{a}_1\boldsymbol{a}_1^\top\right\|/m - O(1) \gtrsim \delta^2 n^2/m \asymp \delta^2 n/\log n$.

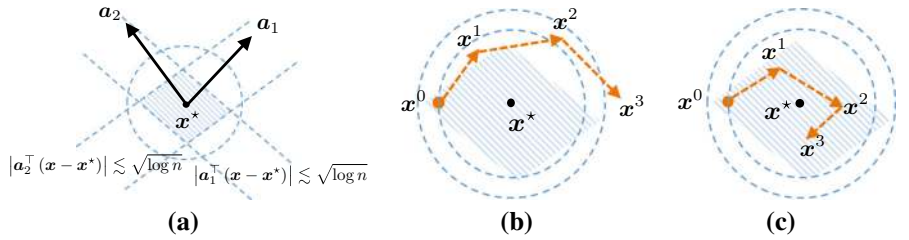


Fig. 3 **a** The shaded region is an illustration of the incoherence region, which satisfies $|a_j^\top (x - x^\star)| \lesssim \sqrt{\log n}$ for all points x in the region. **b** When x^0 resides in the desired region, we know that x^1 remains within the ℓ_2 ball but might fall out of the incoherence region (the shaded region). Once x^1 leaves the incoherence region, we lose control and may overshoot. **c** Our theory reveals that with high probability, all iterates will stay within the incoherence region, enabling fast convergence

simultaneously for x in the RIC. In words, the Hessian matrix is nearly well conditioned (with the condition number bounded by $O(\log n)$), as long as (i) the iterate is not very far from the global minimizer (cf. (11a)) and (ii) the iterate remains incoherent[4] with respect to the sensing vectors (cf. (11b)). Another way to interpret the incoherence condition (11b) is that the empirical risk needs to be well controlled uniformly across all samples. See Fig. 3a for an illustration of the above region.

The following observation is thus immediate: one can safely adopt a far more aggressive step size (as large as $\eta_t = O(1/\log n)$) to achieve acceleration, as long as the iterates stay within the RIC. This, however, fails to be guaranteed by generic gradient descent theory. To be more precise, if the current iterate x^t falls within the desired region, then in view of (8), we can ensure ℓ_2 error contraction after one iteration, namely

$$\|x^{t+1} - x^\star\|_2 \le \|x^t - x^\star\|_2,$$

and hence x^{t+1} stays within the local ℓ_2 ball and hence satisfies (11a). However, it is not immediately obvious that x^{t+1} would still stay incoherent with the sensing vectors and satisfy (11b). If x^{t+1} leaves the RIC, it no longer enjoys the benign local geometry of the loss function, and the algorithm has to slow down in order to avoid overshooting. See Fig. 3b for a visual illustration. In fact, in almost all regularized gradient descent algorithms mentioned in Sect. 1.2, one of the main purposes of the proposed regularization procedures is to enforce such incoherence constraints.

2.4 Implicit Regularization

However, is regularization really necessary for the iterates to stay within the RIC? To answer this question, we plot in Fig. 4a (resp. Fig. 4b) the incoherence measure $\frac{\max_j |a_j^\top x^t|}{\sqrt{\log n}\|x^\star\|_2}$ (resp. $\frac{\max_j |a_j^\top (x^t - x^\star)|}{\sqrt{\log n}\|x^\star\|_2}$) versus the iteration count in a typical Monte Carlo

[4] If x is aligned with (and hence very coherent with) one vector a_j, then with high probability one has $|a_j^\top (x - x^\star)| \gtrsim |a_j^\top x| \asymp \sqrt{n}\|x\|_2$, which is significantly larger than $\sqrt{\log n}\|x\|_2$.

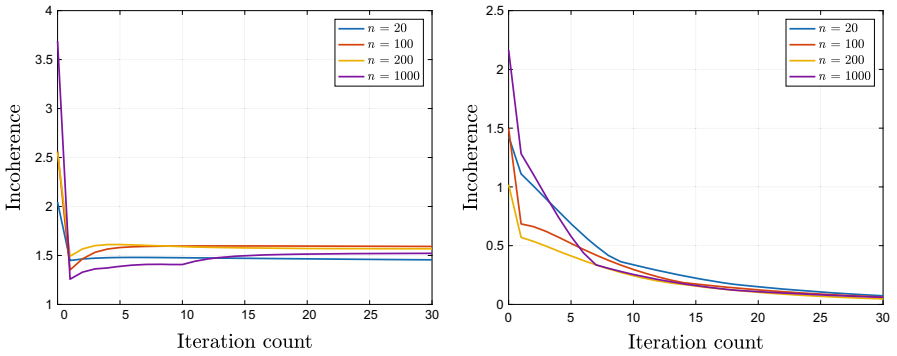


Fig. 4 The incoherence measure $\frac{\max_{1\le j\le m}\left|\boldsymbol{a}_j^\top \boldsymbol{x}^t\right|}{\sqrt{\log n}\|\boldsymbol{x}^\star\|_2}$ (in **a**) and $\frac{\max_{1\le j\le m}\left|\boldsymbol{a}_j^\top (\boldsymbol{x}^t-\boldsymbol{x}^\star)\right|}{\sqrt{\log n}\|\boldsymbol{x}^\star\|_2}$ (in **b**) of the gradient iterates versus iteration count for the phase retrieval problem. The results are shown for $n \in \{20, 100, 200, 1000\}$ and $m = 10n$, with the step size taken to be $\eta_t = 0.1$. The problem instances are generated in the same way as in Fig. 1a

trial, generated in the same way as for Fig. 1a. Interestingly, the incoherence measure remains bounded by 2 for all iterations $t > 1$. This important observation suggests that one may adopt a substantially more aggressive step size throughout the whole algorithm.

The main objective of this paper is thus to provide a theoretical validation of the above empirical observation. As we will demonstrate shortly, with high probability all iterates along the execution of the algorithm (as well as the spectral initialization) are provably constrained within the RIC, implying fast convergence of vanilla gradient descent (cf. Fig. 3c). The fact that the iterates stay incoherent with the measurement mechanism automatically, without explicit enforcement, is termed "implicit regularization."

2.5 A Glimpse of the Analysis: A Leave-One-Out Trick

In order to rigorously establish (11b) for all iterates, the current paper develops a powerful mechanism based on the leave-one-out perturbation argument, a trick rooted and widely used in probability and random matrix theory. Note that the iterate $\boldsymbol{x}^t$ is statistically dependent with the design vectors $\{\boldsymbol{a}_j\}$. Under such circumstances, one often resorts to generic bounds like the Cauchy–Schwarz inequality, which would not yield a desirable estimate. To address this issue, we introduce a sequence of auxiliary iterates $\{\boldsymbol{x}^{t,(l)}\}$ for each $1 \le l \le m$ (for analytical purposes only), obtained by running vanilla gradient descent using all but the lth sample. As one can expect, such auxiliary trajectories serve as extremely good surrogates of $\{\boldsymbol{x}^t\}$ in the sense that

$$\boldsymbol{x}^t \approx \boldsymbol{x}^{t,(l)}, \qquad 1 \le l \le m, \quad t \ge 0, \tag{12}$$

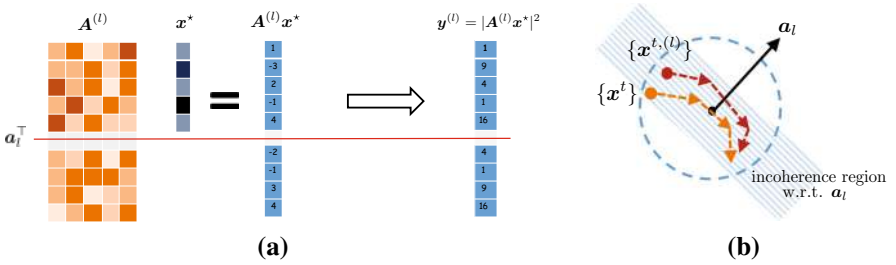


Fig. 5 Illustration of the leave-one-out sequence w.r.t. $\boldsymbol{a}_l$. **a** The sequence $\{\boldsymbol{x}^{t,(l)}\}_{t\geq 0}$ is constructed without using the lth sample. **b** Since the auxiliary sequence $\{\boldsymbol{x}^{t,(l)}\}$ is constructed without using $\boldsymbol{a}_l$, the leave-one-out iterates stay within the incoherence region w.r.t. $\boldsymbol{a}_l$ with high probability. Meanwhile, $\{\boldsymbol{x}^t\}$ and $\{\boldsymbol{x}^{t,(l)}\}$ are expected to remain close as their construction differ only in a single sample

since their constructions only differ by a single sample. Most importantly, since $\boldsymbol{x}^{t,(l)}$ is independent with the lth design vector, it is much easier to control its incoherence w.r.t. $\boldsymbol{a}_l$ to the desired level:

$$\left|\boldsymbol{a}_l^\top\left(\boldsymbol{x}^{t,(l)}-\boldsymbol{x}^\star\right)\right| \lesssim \sqrt{\log n}\left\|\boldsymbol{x}^\star\right\|_2. \tag{13}$$

Combining (12) and (13) then leads to (11b). See Fig. 5 for a graphical illustration of this argument. Notably, this technique is very general and applicable to many other problems. We invite the readers to Sect. 5 for more details.

3 Main Results

This section formalizes the implicit regularization phenomenon underlying unregularized gradient descent and presents its consequences, namely near-optimal statistical and computational guarantees for phase retrieval, matrix completion, and blind deconvolution. Note that the discrepancy measure $\mathrm{dist}(\cdot,\cdot)$ may vary from problem to problem.

3.1 Phase Retrieval

Suppose the m quadratic equations

$$y_j = \left(\boldsymbol{a}_j^\top \boldsymbol{x}^\star\right)^2, \qquad j = 1, 2, \ldots, m \tag{14}$$

are collected using random design vectors, namely $\boldsymbol{a}_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, and the nonconvex problem to solve is

$$\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} \quad f(\boldsymbol{x}) := \frac{1}{4m}\sum_{j=1}^m\left[\left(\boldsymbol{a}_j^\top\boldsymbol{x}\right)^2 - y_j\right]^2. \tag{15}$$

The Wirtinger flow (WF) algorithm, first introduced in [18], is a combination of spectral initialization and vanilla gradient descent; see Algorithm 1.

Algorithm 1 Wirtinger flow/gradient descent for phase retrieval

Input: $\{\boldsymbol{a}_j\}_{1\le j\le m}$ and $\{y_j\}_{1\le j\le m}$.
Spectral initialization: Let $\lambda_1(\boldsymbol{Y})$ and $\widetilde{\boldsymbol{x}}^0$ be the leading eigenvalue and eigenvector of

$$\boldsymbol{Y} = \frac{1}{m}\sum_{j=1}^{m} y_j \boldsymbol{a}_j \boldsymbol{a}_j^\top, \tag{16}$$

respectively, and set $\boldsymbol{x}^0 = \sqrt{\lambda_1(\boldsymbol{Y})/3}\,\widetilde{\boldsymbol{x}}^0$.
Gradient updates: **for** $t = 0, 1, 2, \ldots, T-1$ **do**

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_t \nabla f(\boldsymbol{x}^t). \tag{17}$$

Recognizing that the global phase/sign is unrecoverable from quadratic measurements, we introduce the ℓ_2 distance modulo the global phase as follows

$$\operatorname{dist}(\boldsymbol{x}, \boldsymbol{x}^\star) := \min\left\{\|\boldsymbol{x} - \boldsymbol{x}^\star\|_2, \|\boldsymbol{x} + \boldsymbol{x}^\star\|_2\right\}. \tag{18}$$

Our finding is summarized in the following theorem.

Theorem 1 *Let $\boldsymbol{x}^\star \in \mathbb{R}^n$ be a fixed vector. Suppose $\boldsymbol{a}_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$ for each $1 \le j \le m$ and $m \ge c_0 n \log n$ for some sufficiently large constant $c_0 > 0$. Assume the step size obeys $\eta_t \equiv \eta = c_1/\left(\log n \cdot \|\boldsymbol{x}_0\|_2^2\right)$ for any sufficiently small constant $c_1 > 0$. Then there exist some absolute constants $0 < \varepsilon < 1$ and $c_2 > 0$ such that with probability at least $1 - O\left(mn^{-5}\right)$, Algorithm 1 satisfies that for all $t \ge 0$,*

$$\operatorname{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\star) \le \varepsilon(1 - \eta\|\boldsymbol{x}^\star\|_2^2/2)^t \|\boldsymbol{x}^\star\|_2, \tag{19a}$$

$$\max_{1\le j\le m} \left|\boldsymbol{a}_j^\top\left(\boldsymbol{x}^t - \boldsymbol{x}^\star\right)\right| \le c_2\sqrt{\log n}\|\boldsymbol{x}^\star\|_2. \tag{19b}$$

Theorem 1 reveals a few intriguing properties of Algorithm 1.

- *Implicit regularization* Theorem 1 asserts that the incoherence properties are satisfied throughout the execution of the algorithm (see (19b)), which formally justifies the implicit regularization feature we hypothesized.
- *Near-constant step size* Consider the case where $\|\boldsymbol{x}^\star\|_2 = 1$. Theorem 1 establishes near-linear convergence of WF with a substantially more aggressive step size $\eta \asymp 1/\log n$. Compared with the choice $\eta \lesssim 1/n$ admissible in [18, Theorem 3.3], Theorem 1 allows WF/GD to attain ϵ-accuracy within $O(\log n \log(1/\epsilon))$ iterations. The resulting computational complexity of the algorithm is

$$O\left(mn \log n \log\frac{1}{\epsilon}\right),$$

which significantly improves upon the result $O\big(mn^2 \log(1/\epsilon)\big)$ derived in [18]. As a side note, if the sample size further increases to $m \asymp n \log^2 n$, then a constant step size $\eta \asymp 1$ is also feasible, resulting in an iteration complexity $\log(1/\epsilon)$. This follows since with high probability, the entire trajectory resides within a more refined incoherence region $\max_j \left|\boldsymbol{a}_j^\top \left(\boldsymbol{x}^t - \boldsymbol{x}^\star\right)\right| \lesssim \|\boldsymbol{x}^\star\|_2$. We omit the details here.

- *Incoherence of spectral initialization* We have also demonstrated in Theorem 1 that the initial guess $\boldsymbol{x}^0$ falls within the RIC and is hence nearly orthogonal to all design vectors. This provides a finer characterization of spectral initialization, in comparison with prior theory that focuses primarily on the ℓ_2 accuracy [18,90]. We expect our leave-one-out analysis to accommodate other variants of spectral initialization studied in the literature [12,25,83,88,118].

Remark 3 As it turns out, a carefully designed initialization is not pivotal in enabling fast convergence. In fact, randomly initialized gradient descent provably attains ε-accuracy in $O(\log n + \log \frac{1}{\varepsilon})$ iterations; see [27] for details.

3.2 Low-Rank Matrix Completion

Let $\boldsymbol{M}^\star \in \mathbb{R}^{n\times n}$ be a positive semidefinite matrix[5] with rank r, and suppose its eigendecomposition is

$$\boldsymbol{M}^\star = \boldsymbol{U}^\star \boldsymbol{\Sigma}^\star \boldsymbol{U}^{\star\top}, \tag{20}$$

where $\boldsymbol{U}^\star \in \mathbb{R}^{n\times r}$ consists of orthonormal columns and $\boldsymbol{\Sigma}^\star$ is an $r \times r$ diagonal matrix with eigenvalues in a descending order, i.e., $\sigma_{\max} = \sigma_1 \geq \cdots \geq \sigma_r = \sigma_{\min} > 0$. Throughout this paper, we assume the condition number $\kappa := \sigma_{\max}/\sigma_{\min}$ is bounded by a fixed constant, independent of the problem size (i.e., n and r). Denoting $\boldsymbol{X}^\star = \boldsymbol{U}^\star(\boldsymbol{\Sigma}^\star)^{1/2}$ allows us to factorize $\boldsymbol{M}^\star$ as

$$\boldsymbol{M}^\star = \boldsymbol{X}^\star \boldsymbol{X}^{\star\top}. \tag{21}$$

Consider a random sampling model such that each entry of $\boldsymbol{M}^\star$ is observed independently with probability $0 < p \leq 1$, i.e., for $1 \leq j \leq k \leq n$,

$$Y_{j,k} = \begin{cases} M^\star_{j,k} + E_{j,k}, & \text{with probability } p, \\ 0, & \text{else}, \end{cases} \tag{22}$$

where the entries of $\boldsymbol{E} = [E_{j,k}]_{1\leq j\leq k\leq n}$ are independent sub-Gaussian noise with sub-Gaussian norm σ (see [116, Definition 5.7]). We denote by Ω the set of locations being sampled, and $\mathcal{P}_\Omega(\boldsymbol{Y})$ represents the projection of $\boldsymbol{Y}$ onto the set of matrices supported in Ω. We note here that the sampling rate p, if not known, can be faithfully estimated by the sample proportion $|\Omega|/n^2$.

[5] Here, we assume $\boldsymbol{M}^\star$ to be positive semidefinite to simplify the presentation, but note that our analysis easily extends to asymmetric low-rank matrices.

To fix ideas, we consider the following nonconvex optimization problem

$$\text{minimize}_{\boldsymbol{X}\in\mathbb{R}^{n\times r}} \quad f(\boldsymbol{X}) := \frac{1}{4p}\sum_{(j,k)\in\Omega}\left(\boldsymbol{e}_j^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{e}_k - Y_{j,k}\right)^2. \tag{23}$$

The vanilla gradient descent algorithm (with spectral initialization) is summarized in Algorithm 2.

Algorithm 2 Vanilla gradient descent for matrix completion (with spectral initialization)

Input: $\boldsymbol{Y} = \left[Y_{j,k}\right]_{1\le j,k\le n}$, r, p.

Spectral initialization: Let $\boldsymbol{U}^0\boldsymbol{\Sigma}^0\boldsymbol{U}^{0\top}$ be the rank-r eigendecomposition of

$$\boldsymbol{M}^0 := \frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{Y}) = \frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{M}^\star + \boldsymbol{E}\right),$$

and set $\boldsymbol{X}^0 = \boldsymbol{U}^0\left(\boldsymbol{\Sigma}^0\right)^{1/2}$.

Gradient updates: **for** $t = 0, 1, 2, \ldots, T-1$ **do**

$$\boldsymbol{X}^{t+1} = \boldsymbol{X}^t - \eta_t \nabla f\left(\boldsymbol{X}^t\right). \tag{24}$$

Before proceeding to the main theorem, we first introduce a standard incoherence parameter required for matrix completion [19].

Definition 3 *(Incoherence for matrix completion)* A rank-r matrix $\boldsymbol{M}^\star$ with eigendecomposition $\boldsymbol{M}^\star = \boldsymbol{U}^\star\boldsymbol{\Sigma}^\star\boldsymbol{U}^{\star\top}$ is said to be μ-incoherent if

$$\left\|\boldsymbol{U}^\star\right\|_{2,\infty} \le \sqrt{\frac{\mu}{n}}\left\|\boldsymbol{U}^\star\right\|_{\mathrm{F}} = \sqrt{\frac{\mu r}{n}}. \tag{25}$$

In addition, recognizing that $\boldsymbol{X}^\star$ is identifiable only up to orthogonal transformation, we define the optimal transform from the tth iterate $\boldsymbol{X}^t$ to $\boldsymbol{X}^\star$ as

$$\widehat{\boldsymbol{H}}^t := \operatorname*{argmin}_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \left\|\boldsymbol{X}^t\boldsymbol{R} - \boldsymbol{X}^\star\right\|_{\mathrm{F}}, \tag{26}$$

where $\mathcal{O}^{r\times r}$ is the set of $r\times r$ orthonormal matrices. With these definitions in place, we have the following theorem.

Theorem 2 *Let $\boldsymbol{M}^\star$ be a rank-r, μ-incoherent PSD matrix, and its condition number κ is a fixed constant. Suppose the sample size satisfies $n^2 p \ge C\mu^3 r^3 n\log^3 n$ for some sufficiently large constant $C > 0$, and the noise satisfies*

$$\sigma\sqrt{\frac{n}{p}} \ll \frac{\sigma_{\min}}{\sqrt{\kappa^3\mu r\log^3 n}}. \tag{27}$$

With probability at least $1 - O\left(n^{-3}\right)$*, the iterates of Algorithm 2 satisfy*

$$\left\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t - \boldsymbol{X}^\star\right\|_{\mathrm{F}} \le \left(C_4\rho^t\mu r\frac{1}{\sqrt{np}} + C_1\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\right)\left\|\boldsymbol{X}^\star\right\|_{\mathrm{F}}, \tag{28a}$$

$$\left\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t - \boldsymbol{X}^\star\right\|_{2,\infty} \le \left(C_5\rho^t\mu r\sqrt{\frac{\log n}{np}} + C_8\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\right)\left\|\boldsymbol{X}^\star\right\|_{2,\infty}, \tag{28b}$$

$$\left\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t - \boldsymbol{X}^\star\right\| \le \left(C_9\rho^t\mu r\frac{1}{\sqrt{np}} + C_{10}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\right)\left\|\boldsymbol{X}^\star\right\| \tag{28c}$$

for all $0 \le t \le T = O(n^5)$*, where* C_1*,* C_4*,* C_5*,* C_8*,* C_9*, and* C_{10} *are some absolute positive constants and* $1 - (\sigma_{\min}/5)\cdot\eta \le \rho < 1$*, provided that* $0 < \eta_t \equiv \eta \le 2/(25\kappa\sigma_{\max})$.

Theorem 2 provides the first theoretical guarantee of unregularized gradient descent for matrix completion, demonstrating near-optimal statistical accuracy and computational complexity.

- *Implicit regularization* In Theorem 2, we bound the ℓ_2/ℓ_∞ error of the iterates in a uniform manner via (28b). Note that $\left\|\boldsymbol{X} - \boldsymbol{X}^\star\right\|_{2,\infty} = \max_j \left\|\boldsymbol{e}_j^\top\left(\boldsymbol{X} - \boldsymbol{X}^\star\right)\right\|_2$, which implies the iterates remain incoherent with the sensing vectors throughout and have small incoherence parameters (cf. (25)). In comparison, prior works either include a penalty term on $\{\|\boldsymbol{e}_j^\top\boldsymbol{X}\|_2\}_{1\le j\le n}$ [64,107] and/or $\|\boldsymbol{X}\|_{\mathrm{F}}$ [107] to encourage an incoherent and/or low-norm solution, or add an extra projection operation to enforce incoherence [32,131]. Our results demonstrate that such explicit regularization is unnecessary.
- *Constant step size* Without loss of generality, we may assume that $\sigma_{\max} = \|\boldsymbol{M}^\star\| = O(1)$, which can be done by choosing proper scaling of $\boldsymbol{M}^\star$. Hence, we have a constant step size $\eta_t \asymp 1$. Actually, it is more convenient to consider the scale-invariant parameter ρ: Theorem 2 guarantees linear convergence of the vanilla gradient descent at a constant rate ρ. Remarkably, the convergence occurs with respect to three different unitarily invariant norms: the Frobenius norm $\|\cdot\|_{\mathrm{F}}$, the ℓ_2/ℓ_∞ norm $\|\cdot\|_{2,\infty}$, and the spectral norm $\|\cdot\|$. As far as we know, the latter two are established for the first time. Note that our result even improves upon that for regularized gradient descent; see Table 1.
- *Near-optimal sample complexity* When the rank $r = O(1)$, vanilla gradient descent succeeds under a near-optimal sample complexity $n^2 p \gtrsim n\mathrm{poly}\log n$, which is statistically optimal up to some logarithmic factor.
- *Near-minimal Euclidean error* In view of (28a), as t increases, the Euclidean error of vanilla GD converges to

$$\left\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t - \boldsymbol{X}^\star\right\|_{\mathrm{F}} \lesssim \frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\left\|\boldsymbol{X}^\star\right\|_{\mathrm{F}}, \tag{29}$$

  which coincides with the theoretical guarantee in [32, Corollary 1] and matches the minimax lower bound established in [67,89].

- *Near-optimal entrywise error* The ℓ_2/ℓ_∞ error bound (28b) immediately yields entrywise control of the empirical risk. Specifically, as soon as t is sufficiently large (so that the first term in (28b) is negligible), we have

$$\begin{aligned}\left\|\boldsymbol{X}^t\boldsymbol{X}^{t\top}-\boldsymbol{M}^\star\right\|_\infty &\le \left\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t\left(\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t-\boldsymbol{X}^\star\right)^\top\right\|_\infty+\left\|\left(\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t-\boldsymbol{X}^\star\right)\boldsymbol{X}^{\star\top}\right\|_\infty\\ &\le \left\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t\right\|_{2,\infty}\left\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t-\boldsymbol{X}^\star\right\|_{2,\infty}+\left\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t-\boldsymbol{X}^\star\right\|_{2,\infty}\left\|\boldsymbol{X}^\star\right\|_{2,\infty}\\ &\lesssim \frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{M}^\star\right\|_\infty,\end{aligned}$$

  where the last line follows from (28b) as well as the facts that $\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t-\boldsymbol{X}^\star\|_{2,\infty}\le\|\boldsymbol{X}^\star\|_{2,\infty}$ and $\|\boldsymbol{M}^\star\|_\infty=\|\boldsymbol{X}^\star\|_{2,\infty}^2$. Compared with the Euclidean loss (29), this implies that when $r=O(1)$, the entrywise error of $\boldsymbol{X}^t\boldsymbol{X}^{t\top}$ is uniformly spread out across all entries. As far as we know, this is the first result that reveals near-optimal entrywise error control for noisy matrix completion using nonconvex optimization, without resorting to sample splitting.

Remark 4 Theorem 2 remains valid if the total number T of iterations obeys $T=n^{O(1)}$. In the noiseless case where $\sigma=0$, the theory allows arbitrarily large T.

Finally, we report the empirical statistical accuracy of vanilla gradient descent in the presence of noise. Figure 6 displays the squared relative error of vanilla gradient descent as a function of the signal-to-noise ratio (SNR), where the SNR is defined to be

$$\mathsf{SNR}:=\frac{\sum_{(j,k)\in\Omega}\left(M^\star_{j,k}\right)^2}{\sum_{(j,k)\in\Omega}\mathsf{Var}\left(E_{j,k}\right)}\approx\frac{\|\boldsymbol{M}^\star\|_{\mathrm{F}}^2}{n^2\sigma^2},\tag{30}$$

and the relative error is measured in terms of the square of the metrics as in (28) as well as the squared entrywise prediction error. Both the relative error and the SNR are shown on a dB scale (i.e., $10\log_{10}(\mathsf{SNR})$ and $10\log_{10}$(squared relative error) are plotted). The results are averaged over 20 independent trials. As one can see from the plot, the squared relative error scales inversely proportional to the SNR, which is consistent with our theory.[6]

3.3 Blind Deconvolution

Suppose we have collected m bilinear measurements

$$y_j=\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j,\qquad 1\le j\le m,\tag{31}$$

where $\boldsymbol{a}_j$ follows a complex Gaussian distribution, i.e., $\boldsymbol{a}_j\overset{\text{i.i.d.}}{\sim}\mathcal{N}\left(\boldsymbol{0},\frac{1}{2}\boldsymbol{I}_K\right)+i\mathcal{N}\left(\boldsymbol{0},\frac{1}{2}\boldsymbol{I}_K\right)$ for $1\le j\le m$, and $\boldsymbol{B}:=[\boldsymbol{b}_1,\cdots,\boldsymbol{b}_m]^{\mathsf{H}}\in\mathbb{C}^{m\times K}$ is formed by the first

[6] Note that when $\boldsymbol{M}^\star$ is well conditioned and when $r=O(1)$, one can easily check that $\mathsf{SNR}\approx\left(\|\boldsymbol{M}^\star\|_{\mathrm{F}}^2\right)/\left(n^2\sigma^2\right)\asymp\sigma_{\min}^2/(n^2\sigma^2)$, and our theory says that the squared relative error bound is proportional to $\sigma^2/\sigma_{\min}^2$.

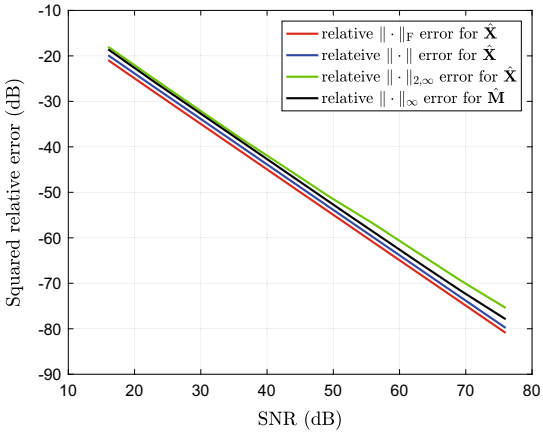


Fig. 6 Squared relative error of the estimate $\widehat{\boldsymbol{X}}$ (measured by $\|\cdot\|_{\mathrm{F}}$, $\|\cdot\|$, $\|\cdot\|_{2,\infty}$ modulo global transformation) and $\widehat{\boldsymbol{M}} = \widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top$ (measured by $\|\cdot\|_\infty$) versus SNR for noisy matrix completion, where $n = 500$, $r = 10$, $p = 0.1$, and $\eta_t = 0.2$. Here $\widehat{\boldsymbol{X}}$ denotes the estimate returned by Algorithm 2 after convergence. The results are averaged over 20 independent Monte Carlo trials

K columns of a unitary discrete Fourier transform (DFT) matrix $\boldsymbol{F} \in \mathbb{C}^{m\times m}$ obeying $\boldsymbol{F}\boldsymbol{F}^{\mathsf{H}} = \boldsymbol{I}_m$ (see Appendix D.3.2 for a brief introduction to DFT matrices). This setup models blind deconvolution, where the two signals under convolution belong to known low-dimensional subspaces of dimension K [3].[7] In particular, the partial DFT matrix $\boldsymbol{B}$ plays an important role in image blind deblurring. In this subsection, we consider solving the following nonconvex optimization problem

$$\text{minimize}_{\boldsymbol{h},\boldsymbol{x}\in\mathbb{C}^K} \quad f(\boldsymbol{h},\boldsymbol{x}) = \sum_{j=1}^{m} \left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}\boldsymbol{a}_j - y_j\right|^2. \tag{32}$$

The (Wirtinger) gradient descent algorithm (with spectral initialization) is summarized in Algorithm 3; here, $\nabla_{\boldsymbol{h}} f(\boldsymbol{h},\boldsymbol{x})$ and $\nabla_{\boldsymbol{x}} f(\boldsymbol{h},\boldsymbol{x})$ stand for the Wirtinger gradient and are given in (77) and (78), respectively; see [18, Section 6] for a brief introduction to Wirtinger calculus.

It is self-evident that $\boldsymbol{h}^\star$ and $\boldsymbol{x}^\star$ are only identifiable up to global scaling, that is, for any nonzero $\alpha \in \mathbb{C}$,

$$\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}} = \frac{1}{\alpha}\boldsymbol{h}^\star\left(\alpha\boldsymbol{x}^\star\right)^{\mathsf{H}}.$$

In light of this, we will measure the discrepancy between

$$z := \begin{bmatrix}\boldsymbol{h}\\ \boldsymbol{x}\end{bmatrix} \in \mathbb{C}^{2K} \qquad \text{and} \qquad z^\star := \begin{bmatrix}\boldsymbol{h}^\star\\ \boldsymbol{x}^\star\end{bmatrix} \in \mathbb{C}^{2K} \tag{33}$$

via the following function

$$\text{dist}\left(z, z^\star\right) := \min_{\alpha\in\mathbb{C}} \sqrt{\left\|\frac{1}{\alpha}\boldsymbol{h} - \boldsymbol{h}^\star\right\|_2^2 + \|\alpha\boldsymbol{x} - \boldsymbol{x}^\star\|_2^2}. \tag{34}$$

[7] For simplicity, we have set the dimensions of the two subspaces equal, and it is straightforward to extend our results to the case of unequal subspace dimensions.

Algorithm 3 Vanilla gradient descent for blind deconvolution (with spectral initialization)

Input: $\{\boldsymbol{a}_j\}_{1\le j\le m}$, $\{\boldsymbol{b}_j\}_{1\le j\le m}$ and $\{y_j\}_{1\le j\le m}$.

Spectral initialization: Let $\sigma_1(\boldsymbol{M})$, $\check{\boldsymbol{h}}^0$ and $\check{\boldsymbol{x}}^0$ be the leading singular value, left and right singular vectors of

$$\boldsymbol{M} := \sum_{j=1}^{m} y_j \boldsymbol{b}_j \boldsymbol{a}_j^{\mathsf{H}},$$

respectively. Set $\boldsymbol{h}^0 = \sqrt{\sigma_1(\boldsymbol{M})}\,\check{\boldsymbol{h}}^0$ and $\boldsymbol{x}^0 = \sqrt{\sigma_1(\boldsymbol{M})}\,\check{\boldsymbol{x}}^0$.

Gradient updates: **for** $t = 0, 1, 2, \ldots, T-1$ **do**

$$\begin{bmatrix} \boldsymbol{h}^{t+1} \\ \boldsymbol{x}^{t+1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{h}^{t} \\ \boldsymbol{x}^{t} \end{bmatrix} - \eta \begin{bmatrix} \frac{1}{\|\boldsymbol{x}^t\|_2^2} \nabla_{\boldsymbol{h}} f\left(\boldsymbol{h}^t, \boldsymbol{x}^t\right) \\ \frac{1}{\|\boldsymbol{h}^t\|_2^2} \nabla_{\boldsymbol{x}} f\left(\boldsymbol{h}^t, \boldsymbol{x}^t\right) \end{bmatrix}. \tag{35}$$

Before proceeding, we need to introduce the incoherence parameter [3,76], which is crucial for blind deconvolution, whose role is similar to the incoherence parameter (cf. Definition 3) in matrix completion.

Definition 4 *(Incoherence for blind deconvolution)* Let the incoherence parameter μ of $\boldsymbol{h}^\star$ be the smallest number such that

$$\max_{1\le j\le m} \left|\boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star\right| \le \frac{\mu}{\sqrt{m}} \left\|\boldsymbol{h}^\star\right\|_2. \tag{36}$$

The incoherence parameter describes the spectral flatness of the signal $\boldsymbol{h}^\star$. With this definition in place, we have the following theorem, where for identifiability we assume that $\left\|\boldsymbol{h}^\star\right\|_2 = \|\boldsymbol{x}^\star\|_2$.

Theorem 3 *Suppose the number of measurements obeys $m \ge C\mu^2 K \log^9 m$ for some sufficiently large constant $C > 0$, and suppose the step size $\eta > 0$ is taken to be some sufficiently small constant. Then there exist constants $c_1, c_2, C_1, C_3, C_4 > 0$ such that with probability exceeding $1 - c_1 m^{-5} - c_1 m e^{-c_2 K}$, the iterates in Algorithm 3 satisfy*

$$\operatorname{dist}\left(\boldsymbol{z}^t, \boldsymbol{z}^\star\right) \le C_1 \left(1 - \frac{\eta}{16}\right)^t \frac{1}{\log^2 m} \left\|\boldsymbol{z}^\star\right\|_2, \tag{37a}$$

$$\max_{1\le l\le m} \left|\boldsymbol{a}_l^{\mathsf{H}} \left(\alpha^t \boldsymbol{x}^t - \boldsymbol{x}^\star\right)\right| \le C_3 \frac{1}{\log^{1.5} m} \left\|\boldsymbol{x}^\star\right\|_2, \tag{37b}$$

$$\max_{1\le l\le m} \left|\boldsymbol{b}_l^{\mathsf{H}} \frac{1}{\overline{\alpha^t}} \boldsymbol{h}^t\right| \le C_4 \frac{\mu}{\sqrt{m}} \log^2 m \left\|\boldsymbol{h}^\star\right\|_2 \tag{37c}$$

for all $t \ge 0$. Here, we denote α^t as the alignment parameter*,*

$$\alpha^t := \arg\min_{\alpha\in\mathbb{C}} \left\| \frac{1}{\overline{\alpha}} \boldsymbol{h}^t - \boldsymbol{h}^\star \right\|_2^2 + \left\|\alpha \boldsymbol{x}^t - \boldsymbol{x}^\star\right\|_2^2. \tag{38}$$

Theorem 3 provides the first theoretical guarantee of unregularized gradient descent for blind deconvolution at a near-optimal statistical and computational complexity. A few remarks are in order.

- *Implicit regularization* Theorem 3 reveals that the unregularized gradient descent iterates remain incoherent with the sampling mechanism (see (37b) and (37c)). Recall that prior works operate upon a regularized cost function with an additional penalty term that regularizes the global scaling $\{\|\boldsymbol{h}\|_2, \|\boldsymbol{x}\|_2\}$ and the incoherence $\{|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}|\}_{1\le j\le m}$ [58,76,82]. In comparison, our theorem implies that it is unnecessary to regularize either the incoherence or the scaling ambiguity, which is somewhat surprising. This justifies the use of regularization-free (Wirtinger) gradient descent for blind deconvolution.
- *Constant step size* Compared to the step size $\eta_t \lesssim 1/m$ suggested in [76] for regularized gradient descent, our theory admits a substantially more aggressive step size (i.e., $\eta_t \asymp 1$) even without regularization. Similar to phase retrieval, the computational efficiency is boosted by a factor of m, attaining ϵ-accuracy within $O(\log(1/\epsilon))$ iterations (vs. $O(m\log(1/\epsilon))$ iterations in prior theory).
- *Near-optimal sample complexity* It is demonstrated that vanilla gradient descent succeeds at a near-optimal sample complexity up to logarithmic factors, although our requirement is slightly worse than [76] which uses explicit regularization. Notably, even under the sample complexity herein, the iteration complexity given in [76] is still $O(m/\mathrm{poly}\log(m))$.
- *Incoherence of spectral initialization* As in phase retrieval, Theorem 3 demonstrates that the estimates returned by the spectral method are incoherent with respect to both $\{\boldsymbol{a}_j\}$ and $\{\boldsymbol{b}_j\}$. In contrast, [76] recommends a projection operation (via a linear program) to enforce incoherence of the initial estimates, which is dispensable according to our theory.
- *Contraction in* $\|\cdot\|_{\mathrm{F}}$ It is easy to check that the Frobenius norm error satisfies $\left\|\boldsymbol{h}^t\boldsymbol{x}^{t\mathsf{H}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}} \lesssim \mathrm{dist}\left(\boldsymbol{z}^t, \boldsymbol{z}^\star\right)$, and therefore, Theorem 3 corroborates the empirical results shown in Fig. 1c.

4 Related Work

Solving nonlinear systems of equations has received much attention in the past decade. Rather than directly attacking the nonconvex formulation, convex relaxation lifts the object of interest into a higher-dimensional space and then attempts recovery via semidefinite programming (e.g., [3,19,20,94]). This has enjoyed great success in both theory and practice. Despite appealing statistical guarantees, semidefinite programming is in general prohibitively expensive when processing large-scale datasets.

Nonconvex approaches, on the other end, have been under extensive study in the last few years, due to their computational advantages. There is a growing list of statistical estimation problems for which nonconvex approaches are guaranteed to find global optimal solutions, including but not limited to phase retrieval [18,25,90], low-rank matrix sensing and completion [7,32,48,115,130], blind deconvolution and self-calibration [72,76,78,82], dictionary learning [106], tensor decomposition [49],

joint alignment [24], learning shallow neural networks [103,132], robust subspace learning [34,74,86,91]. In several problems [40,48,49,75,77,86,87,105,106], it is further suggested that the optimization landscape is benign under sufficiently large sample complexity, in the sense that all local minima are globally optimal, and hence nonconvex iterative algorithms become promising in solving such problems. See [37] for a recent overview. Below we review the three problems studied in this paper in more detail. Some state-of-the-art results are summarized in Table 1.

- *Phase retrieval* Candès et al. proposed *PhaseLift* [20] to solve the quadratic systems of equations based on convex programming. Specifically, it lifts the decision variable $\boldsymbol{x}^\star$ into a rank-one matrix $\boldsymbol{X}^\star = \boldsymbol{x}^\star\boldsymbol{x}^{\star\top}$ and translates the quadratic constraints of $\boldsymbol{x}^\star$ in (14) into linear constraints of $\boldsymbol{X}^\star$. By dropping the rank constraint, the problem becomes convex [11,16,20,29,113]. Another convex program PhaseMax [5,41,50,53] operates in the natural parameter space via linear programming, provided that an anchor vector is available. On the other hand, alternating minimization [90] with sample splitting has been shown to enjoy much better computational guarantee. In contrast, Wirtinger flow [18] provides the first global convergence result for nonconvex methods without sample splitting, whose statistical and computational guarantees are later improved by [25] via an adaptive truncation strategy. Several other variants of WF are also proposed [12,68,102], among which an amplitude-based loss function has been investigated [117–119,127]. In particular, [127] demonstrates that the amplitude-based loss function has a better curvature, and vanilla gradient descent can indeed converge with a constant step size at the orderwise optimal sample complexity. A small sample of other nonconvex phase retrieval methods include [6,10,22,36,43,47,92,98,100,109,122], which are beyond the scope of this paper.
- *Matrix completion* Nuclear norm minimization was studied in [19] as a convex relaxation paradigm to solve the matrix completion problem. Under certain incoherence conditions imposed upon the ground truth matrix, exact recovery is guaranteed under near-optimal sample complexity [14,23,38,51,93]. Concurrently, several works [54,55,60,61,63–65,71,110,123,129,129] tackled the matrix completion problem via nonconvex approaches. In particular, the seminal work by Keshavan et al. [64,65] pioneered the two-stage approach that is widely adopted by later works. Sun and Luo [107] demonstrated the convergence of gradient descent type methods for noiseless matrix completion with a regularized nonconvex loss function. Instead of penalizing the loss function, [32,131] employed projection to enforce the incoherence condition throughout the execution of the algorithm. To the best of our knowledge, no rigorous guarantees have been established for matrix completion without explicit regularization. A notable exception is [63], which uses unregularized stochastic gradient descent for matrix completion in the online setting. However, the analysis is performed with fresh samples in each iteration. Our work closes the gap and makes the first contribution toward understanding implicit regularization in gradient descent without sample splitting. In addition, entrywise eigenvector perturbation has been studied by [1,26,60] in order to analyze the spectral algorithms for matrix completion, which helps us establish theoretical guarantees for the spectral initialization step. Finally, it has recently been shown

that the analysis of nonconvex gradient descent in turn yields near-optimal statistical guarantees for convex relaxation in the context of noisy matrix completion; see [28,31].

- *Blind deconvolution* In [3], Ahmed et al. first proposed to invoke similar lifting ideas for blind deconvolution, which translates the bilinear measurements (31) into a system of linear measurements of a rank-one matrix $\boldsymbol{X}^\star = \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}}$. Near-optimal performance guarantees have been established for convex relaxation [3]. Under the same model, Li et al. [76] proposed a regularized gradient descent algorithm that directly optimizes the nonconvex loss function (32) with a few regularization terms that account for scaling ambiguity and incoherence. In [58], a Riemannian steepest descent method is developed that removes the regularization for scaling ambiguity, although they still need to regularize for incoherence. In [2], a linear program is proposed but requires exact knowledge of the signs of the signals. Blind deconvolution has also been studied for other models—interested readers are referred to [35,72,73,81,82,120,128].

On the other hand, our analysis framework is based on a leave-one-out perturbation argument. This technique has been widely used to analyze high-dimensional problems with random designs, including but not limited to robust M-estimation [44,45], statistical inference for sparse regression [62], likelihood ratio test in logistic regression [108], phase synchronization [1,133], ranking from pairwise comparisons [30], community recovery [1], and covariance sketching [79]. In particular, this technique results in tight performance guarantees for the generalized power method [133], the spectral method [1,30], and convex programming approaches [30,44,108,133]; however, it has not been applied to analyze nonconvex optimization algorithms.

Finally, we note that the notion of implicit regularization—broadly defined—arises in settings far beyond the models and algorithms considered herein. For instance, it has been conjectured that in matrix factorization, over-parameterized stochastic gradient descent effectively enforces certain norm constraints, allowing it to converge to a minimal-norm solution as long as it starts from the origin [52]. The stochastic gradient methods have also been shown to implicitly enforce Tikhonov regularization in several statistical learning settings [80]. More broadly, this phenomenon seems crucial in enabling efficient training of deep neural networks [104,125].

5 A General Recipe for Trajectory Analysis

In this section, we sketch a general recipe for establishing performance guarantees of gradient descent, which conveys the key idea for proving the main results of this paper. The main challenge is to demonstrate that appropriate incoherence conditions are preserved throughout the trajectory of the algorithm. This requires exploiting statistical independence of the samples in a careful manner, in conjunction with generic optimization theory. Central to our approach is a leave-one-out perturbation argument, which allows to decouple the statistical dependency while controlling the component-wise incoherence measures.

General Recipe (a leave-one-out analysis)

Step 1: characterize restricted strong convexity and smoothness of f, and identify the region of incoherence and contraction (RIC).

Step 2: introduce leave-one-out sequences $\{\boldsymbol{X}^{t,(l)}\}$ and $\{\boldsymbol{H}^{t,(l)}\}$ for each l, where $\{\boldsymbol{X}^{t,(l)}\}$ (resp. $\{\boldsymbol{H}^{t,(l)}\}$) is independent of any sample involving $\boldsymbol{\phi}_l$ (resp. $\boldsymbol{\psi}_l$);

Step 3: establish the incoherence condition for $\{\boldsymbol{X}^t\}$ and $\{\boldsymbol{H}^t\}$ via induction. Suppose the iterates satisfy the claimed conditions in the tth iteration:

(a) show, via restricted strong convexity, that the true iterates $(\boldsymbol{X}^{t+1}, \boldsymbol{H}^{t+1})$ and the leave-one-out version $(\boldsymbol{X}^{t+1,(l)}, \boldsymbol{H}^{t+1,(l)})$ are exceedingly close;

(b) use statistical independence to show that $\boldsymbol{X}^{t+1,(l)} - \boldsymbol{X}^\star$ (resp. $\boldsymbol{H}^{t+1,(l)} - \boldsymbol{H}^\star$) is incoherent w.r.t. $\boldsymbol{\phi}_l$ (resp. $\boldsymbol{\psi}_l$); namely, $\|\boldsymbol{\phi}_l^{\mathsf{H}}(\boldsymbol{X}^{t+1,(l)} - \boldsymbol{X}^\star)\|_2$ and $\|\boldsymbol{\psi}_l^{\mathsf{H}}(\boldsymbol{H}^{t+1,(l)} - \boldsymbol{H}^\star)\|_2$ are both well controlled;

(c) combine the bounds to establish the desired incoherence condition concerning $\max_l \|\boldsymbol{\phi}_l^{\mathsf{H}}(\boldsymbol{X}^{t+1} - \boldsymbol{X}^\star)\|_2$ and $\max_l \|\boldsymbol{\psi}_l^{\mathsf{H}}(\boldsymbol{H}^{t+1} - \boldsymbol{H}^\star)\|_2$.

5.1 General Model

Consider the following problem where the samples are collected in a bilinear/quadratic form as

$$y_j = \boldsymbol{\psi}_j^{\mathsf{H}} \boldsymbol{H}^\star \boldsymbol{X}^{\star\mathsf{H}} \boldsymbol{\phi}_j, \qquad 1 \le j \le m, \tag{39}$$

where the objects of interest $\boldsymbol{H}^\star, \boldsymbol{X}^\star \in \mathbb{C}^{n\times r}$ or $\mathbb{R}^{n\times r}$ might be vectors or tall matrices taking either real or complex values. The design vectors $\{\boldsymbol{\psi}_j\}$ and $\{\boldsymbol{\phi}_j\}$ are in either $\mathbb{C}^n$ or $\mathbb{R}^n$, and can be either random or deterministic. This model is quite general and entails all three examples in this paper as special cases:

- *Phase retrieval*: $\boldsymbol{H}^\star = \boldsymbol{X}^\star = \boldsymbol{x}^\star \in \mathbb{R}^n$, and $\boldsymbol{\psi}_j = \boldsymbol{\phi}_j = \boldsymbol{a}_j$;
- *Matrix completion*: $\boldsymbol{H}^\star = \boldsymbol{X}^\star \in \mathbb{R}^{n\times r}$ and $\boldsymbol{\psi}_j, \boldsymbol{\phi}_j \in \{\boldsymbol{e}_1, \cdots, \boldsymbol{e}_n\}$;
- *Blind deconvolution*: $\boldsymbol{H}^\star = \boldsymbol{h}^\star \in \mathbb{C}^K$, $\boldsymbol{X}^\star = \boldsymbol{x}^\star \in \mathbb{C}^K$, $\boldsymbol{\phi}_j = \boldsymbol{a}_j$, and $\boldsymbol{\psi}_j = \boldsymbol{b}_j$.

For this setting, the empirical loss function is given by

$$f(\boldsymbol{Z}) := f(\boldsymbol{H}, \boldsymbol{X}) = \frac{1}{m}\sum_{j=1}^{m} \left|\boldsymbol{\psi}_j^{\mathsf{H}} \boldsymbol{H} \boldsymbol{X}^{\mathsf{H}} \boldsymbol{\phi}_j - y_j\right|^2,$$

where we denote $\boldsymbol{Z} = (\boldsymbol{H}, \boldsymbol{X})$. To minimize $f(\boldsymbol{Z})$, we proceed with vanilla gradient descent

$$\boldsymbol{Z}^{t+1} = \boldsymbol{Z}^t - \eta \nabla f\left(\boldsymbol{Z}^t\right), \qquad \forall t \ge 0$$

following a standard spectral initialization, where η is the step size. As a remark, for complex-valued problems, the gradient (resp. Hessian) should be understood as the Wirtinger gradient (resp. Hessian).

It is clear from (39) that $\boldsymbol{Z}^\star = (\boldsymbol{H}^\star, \boldsymbol{X}^\star)$ can only be recovered up to certain global ambiguity. For clarity of presentation, we assume in this section that such ambiguity has already been taken care of via proper global transformation.

5.2 Outline of the Recipe

We are now positioned to outline the general recipe, which entails the following steps.

- *Step 1: characterizing local geometry in the RIC* Our first step is to characterize a region $\mathcal{R}$—which we term as the *region of incoherence and contraction* (RIC)—such that the Hessian matrix $\nabla^2 f(\boldsymbol{Z})$ obeys strong convexity and smoothness,

$$\boldsymbol{0} \prec \alpha \boldsymbol{I} \preceq \nabla^2 f(\boldsymbol{Z}) \preceq \beta \boldsymbol{I}, \qquad \forall \boldsymbol{Z} \in \mathcal{R}, \tag{40}$$

  or at least along certain directions (i.e., restricted strong convexity and smoothness), where β/α scales slowly (or even remains bounded) with the problem size. As revealed by optimization theory, this geometric property (40) immediately implies linear convergence with the contraction rate $1 - O(\alpha/\beta)$ for a properly chosen step size η, as long as all iterates stay within the RIC.
  A natural question then arises: What does the RIC $\mathcal{R}$ look like? As it turns out, the RIC typically contains all points such that the ℓ_2 error $\|\boldsymbol{Z} - \boldsymbol{Z}^\star\|_{\mathrm{F}}$ is not too large and

$$\text{(\textbf{incoherence})} \qquad \max_j \left\|\boldsymbol{\phi}_j^{\mathsf{H}}(\boldsymbol{X} - \boldsymbol{X}^\star)\right\|_2 \text{ and } \max_j \left\|\boldsymbol{\psi}_j^{\mathsf{H}}(\boldsymbol{H} - \boldsymbol{H}^\star)\right\|_2 \text{ are well controlled.} \tag{41}$$

  In the three examples, the above incoherence condition translates to:

  - *Phase retrieval*: $\max_j \left|\boldsymbol{a}_j^\top(\boldsymbol{x} - \boldsymbol{x}^\star)\right|$ is well controlled;
  - *Matrix completion*: $\left\|\boldsymbol{X} - \boldsymbol{X}^\star\right\|_{2,\infty}$ is well controlled;
  - *Blind deconvolution*: $\max_j \left|\boldsymbol{a}_j^\top(\boldsymbol{x} - \boldsymbol{x}^\star)\right|$ and $\max_j \left|\boldsymbol{b}_j^\top(\boldsymbol{h} - \boldsymbol{h}^\star)\right|$ are well controlled.

- *Step 2: introducing the leave-one-out sequences* To justify that no iterates leave the RIC, we rely on the construction of auxiliary sequences. Specifically, for each l, produce an auxiliary sequence $\{\boldsymbol{Z}^{t,(l)} = (\boldsymbol{X}^{t,(l)}, \boldsymbol{H}^{t,(l)})\}$ such that $\boldsymbol{X}^{t,(l)}$ (resp. $\boldsymbol{H}^{t,(l)}$) is independent of any sample involving $\boldsymbol{\phi}_l$ (resp. $\boldsymbol{\psi}_l$). As an example, suppose that the $\boldsymbol{\phi}_l$'s and the $\boldsymbol{\psi}_l$'s are independently and randomly generated. Then for each l, one can consider a leave-one-out loss function

$$f^{(l)}(\boldsymbol{Z}) := \frac{1}{m} \sum_{j: j \neq l} \left|\boldsymbol{\psi}_j^{\mathsf{H}} \boldsymbol{H} \boldsymbol{X}^{\mathsf{H}} \boldsymbol{\phi}_j - y_j\right|^2$$

that discards the lth sample. One further generates $\{\boldsymbol{Z}^{t,(l)}\}$ by running vanilla gradient descent w.r.t. this auxiliary loss function, with a spectral initialization that similarly discards the lth sample. Note that this procedure is only introduced to facilitate analysis and is never implemented in practice.

- *Step 3: establishing the incoherence condition* We are now ready to establish the incoherence condition with the assistance of the auxiliary sequences. Usually, the proof proceeds by induction, where our goal is to show that the next iterate remains within the RIC, given that the current one does.
  - *Step 3(a): proximity between the original and the leave-one-out iterates* As one can anticipate, $\{\boldsymbol{Z}^t\}$ and $\{\boldsymbol{Z}^{t,(l)}\}$ remain "glued" to each other along the whole trajectory, since their constructions differ by only a single sample. In fact, as long as the initial estimates stay sufficiently close, their gaps will never explode. To intuitively see why, use the fact $\nabla f(\boldsymbol{Z}^t) \approx \nabla f^{(l)}(\boldsymbol{Z}^t)$ to discover that

$$\begin{aligned}\boldsymbol{Z}^{t+1} - \boldsymbol{Z}^{t+1,(l)} &= \boldsymbol{Z}^t - \eta\nabla f(\boldsymbol{Z}^t) - \big(\boldsymbol{Z}^{t,(l)} - \eta\nabla f^{(l)}\big(\boldsymbol{Z}^{t,(l)}\big)\big)\\ &\approx \boldsymbol{Z}^t - \boldsymbol{Z}^{t,(l)} - \eta\nabla^2 f(\boldsymbol{Z}^t)\big(\boldsymbol{Z}^t - \boldsymbol{Z}^{t,(l)}\big),\end{aligned}$$

    which together with the strong convexity condition implies ℓ_2 contraction

$$\big\|\boldsymbol{Z}^{t+1} - \boldsymbol{Z}^{t+1,(l)}\big\|_{\mathrm{F}} \approx \Big\|\big(\boldsymbol{I} - \eta\nabla^2 f(\boldsymbol{Z}^t)\big)\big(\boldsymbol{Z}^t - \boldsymbol{Z}^{t,(l)}\big)\Big\|_{\mathrm{F}} \le \big\|\boldsymbol{Z}^t - \boldsymbol{Z}^{t,(l)}\big\|_2.$$

    Indeed, (restricted) strong convexity is crucial in controlling the size of leave-one-out perturbations.
  - *Step 3(b): incoherence condition of the leave-one-out iterates* The fact that $\boldsymbol{Z}^{t+1}$ and $\boldsymbol{Z}^{t+1,(l)}$ are exceedingly close motivates us to control the incoherence of $\boldsymbol{Z}^{t+1,(l)} - \boldsymbol{Z}^\star$ instead, for $1 \le l \le m$. By construction, $\boldsymbol{X}^{t+1,(l)}$ (resp. $\boldsymbol{H}^{t+1,(l)}$) is statistically *independent* of any sample involving the design vector $\boldsymbol{\phi}_l$ (resp. $\boldsymbol{\psi}_l$), a fact that typically leads to a more friendly analysis for controlling $\big\|\boldsymbol{\phi}_l^{\mathsf{H}}\big(\boldsymbol{X}^{t+1,(l)} - \boldsymbol{X}^\star\big)\big\|_2$ and $\big\|\boldsymbol{\psi}_l^{\mathsf{H}}\big(\boldsymbol{H}^{t+1,(l)} - \boldsymbol{H}^\star\big)\big\|_2$.
  - *Step 3(c): combining the bounds* With these results in place, apply the triangle inequality to obtain

$$\big\|\boldsymbol{\phi}_l^{\mathsf{H}}\big(\boldsymbol{X}^{t+1} - \boldsymbol{X}^\star\big)\big\|_2 \le \|\boldsymbol{\phi}_l\|_2\big\|\boldsymbol{X}^{t+1} - \boldsymbol{X}^{t+1,(l)}\big\|_{\mathrm{F}} + \big\|\boldsymbol{\phi}_l^{\mathsf{H}}\big(\boldsymbol{X}^{t+1,(l)} - \boldsymbol{X}^\star\big)\big\|_2,$$

    where the first term is controlled in Step 3(a) and the second term is controlled in Step 3(b). The term $\big\|\boldsymbol{\psi}_l^{\mathsf{H}}\big(\boldsymbol{H}^{t+1} - \boldsymbol{H}^\star\big)\big\|_2$ can be bounded similarly. By choosing the bounds properly, this establishes the incoherence condition for all $1 \le l \le m$ as desired.

6 Analysis for Phase Retrieval

In this section, we instantiate the general recipe presented in Sect. 5 to phase retrieval and prove Theorem 1. Similar to the section 7.1 in [18], we are going to use $\eta_t =$

$c_1/(\log n \cdot \|\boldsymbol{x}^\star\|_2^2)$ instead of $c_1/(\log n \cdot \|\boldsymbol{x}_0\|_2^2)$ as the step size for analysis. This is because with high probability, $\|\boldsymbol{x}_0\|_2$ and $\|\boldsymbol{x}^\star\|_2$ are rather close in the relative sense. Without loss of generality, we assume throughout this section that $\left\|\boldsymbol{x}^\star\right\|_2 = 1$ and

$$\operatorname{dist}(\boldsymbol{x}^0, \boldsymbol{x}^\star) = \|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|_2 \le \|\boldsymbol{x}^0 + \boldsymbol{x}^\star\|_2. \tag{42}$$

In addition, the gradient and the Hessian of $f(\cdot)$ for this problem (see (15)) are given, respectively, by

$$\nabla f(\boldsymbol{x}) = \frac{1}{m}\sum_{j=1}^{m}\left[\left(\boldsymbol{a}_j^\top \boldsymbol{x}\right)^2 - y_j\right]\left(\boldsymbol{a}_j^\top \boldsymbol{x}\right)\boldsymbol{a}_j, \tag{43}$$

$$\nabla^2 f(\boldsymbol{x}) = \frac{1}{m}\sum_{j=1}^{m}\left[3\left(\boldsymbol{a}_j^\top \boldsymbol{x}\right)^2 - y_j\right]\boldsymbol{a}_j\boldsymbol{a}_j^\top, \tag{44}$$

which are useful throughout the proof.

6.1 Step 1: Characterizing Local Geometry in the RIC

6.1.1 Local Geometry

We start by characterizing the region that enjoys both strong convexity and the desired level of smoothness. This is supplied in the following lemma, which plays a crucial role in the subsequent analysis.

Lemma 1 (Restricted strong convexity and smoothness for phase retrieval) *Fix any sufficiently small constant $C_1 > 0$ and any sufficiently large constant $C_2 > 0$, and suppose the sample complexity obeys $m \ge c_0 n \log n$ for some sufficiently large constant $c_0 > 0$. With probability at least $1 - O(mn^{-10})$,*

$$\nabla^2 f(\boldsymbol{x}) \succeq (1/2)\cdot \boldsymbol{I}_n$$

holds simultaneously for all $\boldsymbol{x} \in \mathbb{R}^n$ satisfying $\|\boldsymbol{x} - \boldsymbol{x}^\star\|_2 \le 2C_1$, and

$$\nabla^2 f(\boldsymbol{x}) \preceq (5C_2(10 + C_2)\log n)\cdot \boldsymbol{I}_n$$

holds simultaneously for all $\boldsymbol{x} \in \mathbb{R}^n$ obeying

$$\left\|\boldsymbol{x} - \boldsymbol{x}^\star\right\|_2 \le 2C_1, \tag{45a}$$

$$\max_{1\le j\le m}\left|\boldsymbol{a}_j^\top\left(\boldsymbol{x} - \boldsymbol{x}^\star\right)\right| \le C_2\sqrt{\log n}. \tag{45b}$$

Proof See Appendix A.1. □

In words, Lemma 1 reveals that the Hessian matrix is positive definite and (almost) well conditioned, if one restricts attention to the set of points that are (i) not far away from the truth (cf. (45a)) and (ii) incoherent with respect to the measurement vectors $\{\boldsymbol{a}_j\}_{1\le j\le m}$ (cf. (45b)).

6.1.2 Error Contraction

As we point out before, the nice local geometry enables ℓ_2 contraction, which we formalize below.

Lemma 2 *There exists an event that does not depend on t and has probability $1-O(mn^{-10})$, such that when it happens and $\boldsymbol{x}^t$ obeys conditions (45), one has*

$$\left\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^\star\right\|_2 \le (1-\eta/2)\left\|\boldsymbol{x}^t-\boldsymbol{x}^\star\right\|_2 \tag{46}$$

provided that the step size satisfies $0<\eta\le 1/\left[5C_2\,(10+C_2)\log n\right]$.

Proof This proof applies the standard argument when establishing the ℓ_2 error contraction of gradient descent for strongly convex and smooth functions. See Appendix A.2. □

With the help of Lemma 2, we can turn the proof of Theorem 1 into ensuring that the trajectory $\{\boldsymbol{x}^t\}_{0\le t\le n}$ lies in the RIC specified by (47).[8] This is formally stated in the next lemma.

Lemma 3 *Suppose for all $0\le t\le T_0:=n$, the trajectory $\{\boldsymbol{x}^t\}$ falls within the region of incoherence and contraction (termed the RIC), namely*

$$\left\|\boldsymbol{x}^t-\boldsymbol{x}^\star\right\|_2 \le C_1, \tag{47a}$$

$$\max_{1\le l\le m}\left|\boldsymbol{a}_l^\top\left(\boldsymbol{x}^t-\boldsymbol{x}^\star\right)\right| \le C_2\sqrt{\log n}, \tag{47b}$$

then the claims in Theorem 1 hold true. Here and throughout this section, $C_1, C_2>0$ are two absolute constants as specified in Lemma 1.

Proof See Appendix A.3. □

6.2 Step 2: Introducing the Leave-One-Out Sequences

In comparison with the ℓ_2 error bound (47a) that captures the overall loss, the incoherence hypothesis (47b)—which concerns sample-wise control of the empirical risk—is more complicated to establish. This is partly due to the statistical dependence between $\boldsymbol{x}^t$ and the sampling vectors $\{\boldsymbol{a}_l\}$. As described in the general recipe, the key idea is

[8] Here, we deliberately change $2C_1$ in (45a) to C_1 in the definition of the RIC (47a) to ensure the correctness of the analysis.

the introduction of a *leave-one-out* version of the WF iterates, which removes a single measurement from consideration.

To be precise, for each $1 \le l \le m$, we define the leave-one-out empirical loss function as

$$f^{(l)}(\boldsymbol{x}) := \frac{1}{4m} \sum_{j:j\neq l} \left[\left(\boldsymbol{a}_j^\top \boldsymbol{x} \right)^2 - y_j \right]^2, \tag{48}$$

and the auxiliary trajectory $\left\{ \boldsymbol{x}^{t,(l)} \right\}_{t\ge 0}$ is constructed by running WF w.r.t. $f^{(l)}(\boldsymbol{x})$. In addition, the spectral initialization $\boldsymbol{x}^{0,(l)}$ is computed based on the rescaled leading eigenvector of the leave-one-out data matrix

$$\boldsymbol{Y}^{(l)} := \frac{1}{m} \sum_{j:j\neq l} y_j \boldsymbol{a}_j \boldsymbol{a}_j^\top. \tag{49}$$

Clearly, the entire sequence $\left\{ \boldsymbol{x}^{t,(l)} \right\}_{t\ge 0}$ is independent of the lth sampling vector $\boldsymbol{a}_l$. This auxiliary procedure is formally described in Algorithm 4.

Algorithm 4 The lth leave-one-out sequence for phase retrieval

Input: $\{\boldsymbol{a}_j\}_{1\le j\le m, j\neq l}$ and $\{y_j\}_{1\le j\le m, j\neq l}$.

Spectral initialization: let $\lambda_1\left(\boldsymbol{Y}^{(l)}\right)$ and $\widetilde{\boldsymbol{x}}^{0,(l)}$ be the leading eigenvalue and eigenvector of

$$\boldsymbol{Y}^{(l)} = \frac{1}{m} \sum_{j:j\neq l} y_j \boldsymbol{a}_j \boldsymbol{a}_j^\top,$$

respectively, and set

$$\boldsymbol{x}^{0,(l)} = \begin{cases} \sqrt{\lambda_1\left(\boldsymbol{Y}^{(l)}\right)/3}\, \widetilde{\boldsymbol{x}}^{0,(l)}, & \text{if } \left\|\widetilde{\boldsymbol{x}}^{0,(l)} - \boldsymbol{x}^\star\right\|_2 \le \left\|\widetilde{\boldsymbol{x}}^{0,(l)} + \boldsymbol{x}^\star\right\|_2, \\ -\sqrt{\lambda_1\left(\boldsymbol{Y}^{(l)}\right)/3}\, \widetilde{\boldsymbol{x}}^{0,(l)}, & \text{else.} \end{cases}$$

Gradient updates: **for** $t = 0, 1, 2, \ldots, T-1$ **do**

$$\boldsymbol{x}^{t+1,(l)} = \boldsymbol{x}^{t,(l)} - \eta_t \nabla f^{(l)}\left(\boldsymbol{x}^{t,(l)}\right). \tag{50}$$

6.3 Step 3: Establishing the Incoherence Condition by Induction

As revealed by Lemma 3, it suffices to prove that the iterates $\{\boldsymbol{x}^t\}_{0\le t\le T_0}$ satisfies (47) with high probability. Our proof will be inductive in nature. For the sake of clarity, we list all the induction hypotheses:

$$\left\|\boldsymbol{x}^t - \boldsymbol{x}^\star\right\|_2 \le C_1, \tag{51a}$$

$$\max_{1\le l\le m} \left\|\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)}\right\|_2 \le C_3\sqrt{\frac{\log n}{n}} \tag{51b}$$

$$\max_{1\le j\le m} \left|\boldsymbol{a}_j^\top\left(\boldsymbol{x}^t - \boldsymbol{x}^\star\right)\right| \le C_2\sqrt{\log n}. \tag{51c}$$

Here $C_3 > 0$ is some universal constant. For any $t \ge 0$, define $\mathcal{E}_t$ to be the event where the conditions in (51) hold for the tth iteration. According to Lemma 2, there exists some event $\mathcal{E}$ with probability $1 - O(mn^{-10})$ such that on $\mathcal{E}_t \cap \mathcal{E}$ one has

$$\left\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^\star\right\|_2 \le C_1. \tag{52}$$

This subsection is devoted to establishing (51b) and (51c) for the $(t+1)$th iteration, assuming that (51) holds true up to the tth iteration. We defer the justification of the base case (i.e., initialization at $t = 0$) to Sect. 6.4.

- *Step 3(a): proximity between the original and the leave-one-out iterates* The leave-one-out sequence $\{\boldsymbol{x}^{t,(l)}\}$ behaves similarly to the true WF iterates $\{\boldsymbol{x}^t\}$ while maintaining statistical independence with $\boldsymbol{a}_l$, a key fact that allows us to control the incoherence of lth leave-one-out sequence w.r.t. $\boldsymbol{a}_l$. We will formally quantify the gap between $\boldsymbol{x}^{t+1}$ and $\boldsymbol{x}^{t+1,(l)}$ in the following lemma, which establishes the induction in (51b).

Lemma 4 *Suppose that the sample size obeys $m \ge Cn\log n$ for some sufficiently large constant $C > 0$ and that the step size obeys $0 < \eta < 1/[5C_2(10 + C_2)\log n]$. Then on some event $\mathcal{E}_{t+1,1} \subseteq \mathcal{E}_t$ obeying $\mathbb{P}(\mathcal{E}_t \cap \mathcal{E}_{t+1,1}^c) = O(mn^{-10})$, one has*

$$\max_{1\le l\le m} \left\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^{t+1,(l)}\right\|_2 \le C_3\sqrt{\frac{\log n}{n}}. \tag{53}$$

Proof The proof relies heavily on the restricted strong convexity (see Lemma 1) and is deferred to Appendix A.4. □

- *Step 3(b): incoherence of the leave-one-out iterates* By construction, $\boldsymbol{x}^{t+1,(l)}$ is statistically independent of the sampling vector $\boldsymbol{a}_l$. One can thus invoke the standard Gaussian concentration results and the union bound to derive that on an event $\mathcal{E}_{t+1,2} \subseteq \mathcal{E}_t$ obeying $\mathbb{P}(\mathcal{E}_t \cap \mathcal{E}_{t+1,2}^c) = O(mn^{-10})$,

$$\begin{aligned}
\max_{1\le l\le m} \left|\boldsymbol{a}_l^\top\left(\boldsymbol{x}^{t+1,(l)} - \boldsymbol{x}^\star\right)\right| &\le 5\sqrt{\log n}\left\|\boldsymbol{x}^{t+1,(l)} - \boldsymbol{x}^\star\right\|_2 \\
&\overset{\text{(i)}}{\le} 5\sqrt{\log n}\left(\left\|\boldsymbol{x}^{t+1,(l)} - \boldsymbol{x}^{t+1}\right\|_2 + \left\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^\star\right\|_2\right) \\
&\overset{\text{(ii)}}{\le} 5\sqrt{\log n}\left(C_3\sqrt{\frac{\log n}{n}} + C_1\right) \\
&\le C_4\sqrt{\log n}
\end{aligned} \tag{54}$$

holds for some constant $C_4 \geq 6C_1 > 0$ and n sufficiently large. Here, (i) comes from the triangle inequality and (ii) arises from the proximity bound (53) and the condition (52).

- *Step 3(c): combining the bounds* We are now prepared to establish (51c) for the $(t+1)$th iteration. Specifically,

$$\begin{aligned}\max_{1\leq l\leq m}\left|\boldsymbol{a}_l^\top\left(\boldsymbol{x}^{t+1}-\boldsymbol{x}^\star\right)\right| &\leq \max_{1\leq l\leq m}\left|\boldsymbol{a}_l^\top\left(\boldsymbol{x}^{t+1}-\boldsymbol{x}^{t+1,(l)}\right)\right| + \max_{1\leq l\leq m}\left|\boldsymbol{a}_l^\top\left(\boldsymbol{x}^{t+1,(l)}-\boldsymbol{x}^\star\right)\right| \\ &\overset{\text{(i)}}{\leq} \max_{1\leq l\leq m}\|\boldsymbol{a}_l\|_2\left\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^{t+1,(l)}\right\|_2 + C_4\sqrt{\log n} \\ &\overset{\text{(ii)}}{\leq} \sqrt{6n}\cdot C_3\sqrt{\frac{\log n}{n}} + C_4\sqrt{\log n} \leq C_2\sqrt{\log n}, \end{aligned} \tag{55}$$

where (i) follows from the Cauchy–Schwarz inequality and (54), the inequality (ii) is a consequence of (53) and (98), and the last inequality holds as long as $C_2/(C_3+C_4)$ is sufficiently large. From the deduction above we easily get $\mathbb{P}(\mathcal{E}_t \cap \mathcal{E}_{t+1}^c) = O(mn^{-10})$.

Using mathematical induction and the union bound, we establish (51) for all $t \leq T_0 = n$ with high probability. This in turn concludes the proof of Theorem 1, as long as the hypotheses are valid for the base case.

6.4 The Base Case: Spectral Initialization

In the end, we return to verify the induction hypotheses for the base case ($t = 0$), i.e., the spectral initialization obeys (51). The following lemma justifies (51a) by choosing δ sufficiently small.

Lemma 5 *Fix any small constant $\delta > 0$, and suppose $m > c_0 n \log n$ for some large constant $c_0 > 0$. Consider the two vectors $\boldsymbol{x}^0$ and $\widetilde{\boldsymbol{x}}^0$ as defined in Algorithm 1, and suppose without loss of generality that (42) holds. Then with probability exceeding $1 - O(n^{-10})$, one has*

$$\|\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}]\| \leq \delta, \tag{56}$$

$$\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|_2 \leq 2\delta \quad \text{and} \quad \left\|\widetilde{\boldsymbol{x}}^0 - \boldsymbol{x}^\star\right\|_2 \leq \sqrt{2}\delta. \tag{57}$$

Proof This result follows directly from the Davis–Kahan $\sin\Theta$ theorem. See Appendix A.5. □

We then move on to justifying (51b), the proximity between the original and leave-one-out iterates for $t = 0$.

Lemma 6 *Suppose $m > c_0 n \log n$ for some large constant $c_0 > 0$. Then with probability at least $1 - O(mn^{-10})$, one has*

$$\max_{1\leq l\leq m}\left\|\boldsymbol{x}^0 - \boldsymbol{x}^{0,(l)}\right\|_2 \leq C_3\sqrt{\frac{\log n}{n}}. \tag{58}$$

Proof This is also a consequence of the Davis–Kahan sinΘ theorem. See Appendix A.6. □

The final claim (51c) can be proved using the same argument as in deriving (55) and hence is omitted.

7 Analysis for Matrix Completion

In this section, we instantiate the general recipe presented in Sect. 5 to matrix completion and prove Theorem 2. Before continuing, we first gather a few useful facts regarding the loss function in (23). The gradient of it is given by

$$\nabla f\left(\boldsymbol{X}\right)=\frac{1}{p}\mathcal{P}_{\Omega}\left[\boldsymbol{X}\boldsymbol{X}^{\top}-\left(\boldsymbol{M}^{\star}+\boldsymbol{E}\right)\right]\boldsymbol{X}. \tag{59}$$

We define the expected gradient (with respect to the sampling set Ω) to be

$$\nabla F\left(\boldsymbol{X}\right)=\left[\boldsymbol{X}\boldsymbol{X}^{\top}-\left(\boldsymbol{M}^{\star}+\boldsymbol{E}\right)\right]\boldsymbol{X}$$

and also the (expected) gradient without noise to be

$$\nabla f_{\text{clean}}\left(\boldsymbol{X}\right)=\frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{X}\boldsymbol{X}^{\top}-\boldsymbol{M}^{\star}\right)\boldsymbol{X}\qquad\text{and}\qquad\nabla F_{\text{clean}}\left(\boldsymbol{X}\right)=\left(\boldsymbol{X}\boldsymbol{X}^{\top}-\boldsymbol{M}^{\star}\right)\boldsymbol{X}. \tag{60}$$

In addition, we need the Hessian $\nabla^2 f_{\text{clean}}(\boldsymbol{X})$, which is represented by an $nr\times nr$ matrix. Simple calculations reveal that for any $\boldsymbol{V}\in\mathbb{R}^{n\times r}$,

$$\begin{aligned}\text{vec}\left(\boldsymbol{V}\right)^{\top}\nabla^2 f_{\text{clean}}\left(\boldsymbol{X}\right)\text{vec}\left(\boldsymbol{V}\right)&=\frac{1}{2p}\left\|\mathcal{P}_{\Omega}\left(\boldsymbol{V}\boldsymbol{X}^{\top}+\boldsymbol{X}\boldsymbol{V}^{\top}\right)\right\|_{\text{F}}^{2}\\&\quad+\frac{1}{p}\left\langle\mathcal{P}_{\Omega}\left(\boldsymbol{X}\boldsymbol{X}^{\top}-\boldsymbol{M}^{\star}\right),\boldsymbol{V}\boldsymbol{V}^{\top}\right\rangle,\end{aligned} \tag{61}$$

where $\text{vec}(\boldsymbol{V})\in\mathbb{R}^{nr}$ denotes the vectorization of $\boldsymbol{V}$.

7.1 Step 1: Characterizing Local Geometry in the RIC

7.1.1 Local Geometry

The first step is to characterize the region where the empirical loss function enjoys restricted strong convexity and smoothness in an appropriate sense. This is formally stated in the following lemma.

Lemma 7 (Restricted strong convexity and smoothness for matrix completion) *Suppose that the sample size obeys $n^2 p\geq C\kappa^2\mu r n\log n$ for some sufficiently large*

constant $C > 0$. Then with probability at least $1 - O\left(n^{-10}\right)$, the Hessian $\nabla^2 f_{\text{clean}}(\boldsymbol{X})$ as defined in (61) obeys

$$\operatorname{vec}(\boldsymbol{V})^\top \nabla^2 f_{\text{clean}}(\boldsymbol{X}) \operatorname{vec}(\boldsymbol{V}) \geq \frac{\sigma_{\min}}{2} \|\boldsymbol{V}\|_{\mathrm{F}}^2 \quad \text{and} \quad \left\|\nabla^2 f_{\text{clean}}(\boldsymbol{X})\right\| \leq \frac{5}{2}\sigma_{\max} \tag{62}$$

for all $\boldsymbol{X}$ and $\boldsymbol{V} = \boldsymbol{Y}\boldsymbol{H}_{\boldsymbol{Y}} - \boldsymbol{Z}$, with $\boldsymbol{H}_{\boldsymbol{Y}} := \arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \|\boldsymbol{Y}\boldsymbol{R} - \boldsymbol{Z}\|_{\mathrm{F}}$, satisfying:

$$\left\|\boldsymbol{X} - \boldsymbol{X}^\star\right\|_{2,\infty} \leq \epsilon \left\|\boldsymbol{X}^\star\right\|_{2,\infty}, \tag{63a}$$

$$\|\boldsymbol{Z} - \boldsymbol{X}^\star\| \leq \delta \|\boldsymbol{X}^\star\|, \tag{63b}$$

where $\epsilon \ll 1/\sqrt{\kappa^3 \mu r \log^2 n}$ and $\delta \ll 1/\kappa$.

Proof See Appendix B.1. □

Lemma 7 reveals that the Hessian matrix is well conditioned in a neighborhood close to $\boldsymbol{X}^\star$ that remains incoherent measured in the ℓ_2/ℓ_∞ norm (cf. (63a)), and along directions that point toward points which are not far away from the truth in the spectral norm (cf. (63b)).

Remark 5 The second condition (63b) is characterized using the spectral norm $\|\cdot\|$, while in previous works this is typically presented in the Frobenius norm $\|\cdot\|_{\mathrm{F}}$. It is also worth noting that the Hessian matrix—even in the infinite-sample and noiseless case—is rank-deficient and cannot be positive definite. As a result, we resort to the form of strong convexity by restricting attention to certain directions (see the conditions on $\boldsymbol{V}$).

7.1.2 Error Contraction

Our goal is to demonstrate the error bounds (28) measured in three different norms. Notably, as long as the iterates satisfy (28) at the tth iteration, then $\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t - \boldsymbol{X}^\star\|_{2,\infty}$ is sufficiently small. Under our sample complexity assumption, $\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t$ satisfies the ℓ_2/ℓ_∞ condition (63a) required in Lemma 7. Consequently, we can invoke Lemma 7 to arrive at the following error contraction result.

Lemma 8 (Contraction w.r.t. the Frobenius norm) *Suppose that $n^2 p \geq C\kappa^3\mu^3 r^3 n \log^3 n$ for some sufficiently large constant $C > 0$, and the noise satisfies (27). There exists an event that does not depend on t and has probability $1 - O(n^{-10})$, such that when it happens and (28a), (28b) hold for the tth iteration, one has*

$$\left\|\boldsymbol{X}^{t+1}\widehat{\boldsymbol{H}}^{t+1} - \boldsymbol{X}^\star\right\|_{\mathrm{F}} \leq C_4\rho^{t+1}\mu r\frac{1}{\sqrt{np}}\left\|\boldsymbol{X}^\star\right\|_{\mathrm{F}} + C_1\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\left\|\boldsymbol{X}^\star\right\|_{\mathrm{F}}$$

provided that $0 < \eta \leq 2/(25\kappa\sigma_{\max})$, $1 - (\sigma_{\min}/4)\cdot\eta \leq \rho < 1$, and C_1 is sufficiently large.

Proof The proof is built upon Lemma 7. See Appendix B.2. □

Further, if the current iterate satisfies all three conditions in (28), then we can derive a stronger sense of error contraction, namely contraction in terms of the spectral norm.

Lemma 9 (Contraction w.r.t. the spectral norm) *Suppose $n^2 p \geq C\kappa^3\mu^3 r^3 n \log^3 n$ for some sufficiently large constant $C > 0$, and the noise satisfies (27). There exists an event that does not depend on t and has probability $1 - O(n^{-10})$, such that when it happens and (28) holds for the tth iteration, one has*

$$\left\| \boldsymbol{X}^{t+1}\widehat{\boldsymbol{H}}^{t+1} - \boldsymbol{X}^\star \right\| \leq C_9\rho^{t+1}\mu r\frac{1}{\sqrt{np}}\left\|\boldsymbol{X}^\star\right\| + C_{10}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\left\|\boldsymbol{X}^\star\right\| \tag{64}$$

provided that $0 < \eta \leq 1/(2\sigma_{\max})$ and $1 - (\sigma_{\min}/3)\cdot\eta \leq \rho < 1$.

Proof The key observation is this: the iterate that proceeds according to the population-level gradient reduces the error w.r.t. $\|\cdot\|$, namely

$$\left\| \boldsymbol{X}^t\widehat{\boldsymbol{H}}^t - \eta\nabla F_{\text{clean}}\left(\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t\right) - \boldsymbol{X}^\star \right\| < \left\| \boldsymbol{X}^t\widehat{\boldsymbol{H}}^t - \boldsymbol{X}^\star \right\|,$$

as long as $\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t$ is sufficiently close to the truth. Notably, the orthonormal matrix $\widehat{\boldsymbol{H}}^t$ is still chosen to be the one that minimizes the $\|\cdot\|_{\text{F}}$ distance (as opposed to $\|\cdot\|$), which yields a symmetry property $\boldsymbol{X}^{\star\top}\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t = \left(\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t\right)^\top\boldsymbol{X}^\star$, crucial for our analysis. See Appendix B.3 for details. □

7.2 Step 2: Introducing the Leave-One-Out Sequences

In order to establish the incoherence properties (28b) for the entire trajectory, which is difficult to deal with directly due to the complicated statistical dependence, we introduce a collection of *leave-one-out* versions of $\left\{\boldsymbol{X}^t\right\}_{t\geq 0}$, denoted by $\left\{\boldsymbol{X}^{t,(l)}\right\}_{t\geq 0}$ for each $1 \leq l \leq n$. Specifically, $\left\{\boldsymbol{X}^{t,(l)}\right\}_{t\geq 0}$ is the iterates of gradient descent operating on the auxiliary loss function

$$f^{(l)}(\boldsymbol{X}) := \frac{1}{4p}\left\|\mathcal{P}_{\Omega^{-l}}\left[\boldsymbol{X}\boldsymbol{X}^\top - \left(\boldsymbol{M}^\star + \boldsymbol{E}\right)\right]\right\|_{\text{F}}^2 + \frac{1}{4}\left\|\mathcal{P}_l\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M}^\star\right)\right\|_{\text{F}}^2. \tag{65}$$

Here, $\mathcal{P}_{\Omega_l}$ (resp. $\mathcal{P}_{\Omega^{-l}}$ and $\mathcal{P}_l$) represents the orthogonal projection onto the subspace of matrices which vanish outside of the index set $\Omega_l := \{(i,j) \in \Omega \mid i = l \text{ or } j = l\}$ (resp. $\Omega^{-l} := \{(i,j) \in \Omega \mid i \neq l, j \neq l\}$ and $\{(i,j) \mid i = l \text{ or } j = l\}$); that is, for any matrix $\boldsymbol{M}$,

$$\left[\mathcal{P}_{\Omega_l}(\boldsymbol{M})\right]_{i,j} = \begin{cases} M_{i,j}, & \text{if } (i = l \text{ or } j = l) \text{ and } (i,j) \in \Omega, \\ 0, & \text{else,} \end{cases} \tag{66}$$

$$\left[\mathcal{P}_{\Omega^{-l}}(\boldsymbol{M})\right]_{i,j} = \begin{cases} M_{i,j}, & \text{if } i \neq l \text{ and } j \neq l \text{ and } (i,j) \in \Omega \\ 0, & \text{else} \end{cases} \quad \text{and}$$

$$[\mathcal{P}_l(\boldsymbol{M})]_{i,j} = \begin{cases} 0, & \text{if } i \neq l \text{ and } j \neq l, \\ M_{i,j}, & \text{if } i = l \text{ or } j = l. \end{cases} \tag{67}$$

The gradient of the leave-one-out loss function (65) is given by

$$\nabla f^{(l)}(\boldsymbol{X}) = \frac{1}{p}\mathcal{P}_{\Omega^{-l}}\left[\boldsymbol{X}\boldsymbol{X}^\top - \left(\boldsymbol{M}^\star + \boldsymbol{E}\right)\right]\boldsymbol{X} + \mathcal{P}_l\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M}^\star\right)\boldsymbol{X}. \tag{68}$$

The full algorithm to obtain the leave-one-out sequence $\{\boldsymbol{X}^{t,(l)}\}_{t\geq 0}$ (including spectral initialization) is summarized in Algorithm 5.

Algorithm 5 The lth leave-one-out sequence for matrix completion

Input: $\boldsymbol{Y} = [Y_{i,j}]_{1\leq i,j\leq n}$, $\boldsymbol{M}^\star_{\cdot,l}$, $\boldsymbol{M}^\star_{l,\cdot}$, r, p.

Spectral initialization: Let $\boldsymbol{U}^{0,(l)}\boldsymbol{\Sigma}^{(l)}\boldsymbol{U}^{0,(l)\top}$ be the top-r eigendecomposition of

$$\boldsymbol{M}^{(l)} := \frac{1}{p}\mathcal{P}_{\Omega^{-l}}(\boldsymbol{Y}) + \mathcal{P}_l\left(\boldsymbol{M}^\star\right) = \frac{1}{p}\mathcal{P}_{\Omega^{-l}}\left(\boldsymbol{M}^\star + \boldsymbol{E}\right) + \mathcal{P}_l\left(\boldsymbol{M}^\star\right)$$

with $\mathcal{P}_{\Omega^{-l}}$ and $\mathcal{P}_l$ defined in (67), and set $\boldsymbol{X}^{0,(l)} = \boldsymbol{U}^{0,(l)}\left(\boldsymbol{\Sigma}^{(l)}\right)^{1/2}$.

Gradient updates: **for** $t = 0, 1, 2, \ldots, T-1$ **do**

$$\boldsymbol{X}^{t+1,(l)} = \boldsymbol{X}^{t,(l)} - \eta_t \nabla f^{(l)}\left(\boldsymbol{X}^{t,(l)}\right). \tag{69}$$

Remark 6 Rather than simply dropping all samples in the lth row/column, we replace the lth row/column with their respective population means. In other words, the leave-one-out gradient forms an unbiased surrogate for the true gradient, which is particularly important in ensuring high estimation accuracy.

7.3 Step 3: Establishing the Incoherence Condition by Induction

We will continue the proof of Theorem 2 in an inductive manner. As seen in Sect. 7.1.2, the induction hypotheses (28a) and (28c) hold for the $(t+1)$th iteration as long as (28) holds at the tth iteration. Therefore, we are left with proving the incoherence hypothesis (28b) for all $0 \leq t \leq T = O(n^5)$. For clarity of analysis, it is crucial to maintain a list of induction hypotheses, which includes a few more hypotheses that complement (28), and is given below.

$$\left\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t - \boldsymbol{X}^\star\right\|_{\mathrm{F}} \leq \left(C_4\rho^t\mu r\frac{1}{\sqrt{np}} + C_1\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\right)\left\|\boldsymbol{X}^\star\right\|_{\mathrm{F}}, \tag{70a}$$

$$\left\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t - \boldsymbol{X}^\star\right\|_{2,\infty} \leq \left(C_5\rho^t\mu r\sqrt{\frac{\log n}{np}} + C_8\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\right)\left\|\boldsymbol{X}^\star\right\|_{2,\infty}, \tag{70b}$$

$$\left\|\boldsymbol{X}^{t}\widehat{\boldsymbol{H}}^{t}-\boldsymbol{X}^{\star}\right\| \le \left(C_9\rho^{t}\mu r\frac{1}{\sqrt{np}}+C_{10}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\right)\left\|\boldsymbol{X}^{\star}\right\|, \tag{70c}$$

$$\max_{1\le l\le n}\left\|\boldsymbol{X}^{t}\widehat{\boldsymbol{H}}^{t}-\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}\right\|_{\mathrm{F}} \le \left(C_3\rho^{t}\mu r\sqrt{\frac{\log n}{np}}+C_7\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\right)\left\|\boldsymbol{X}^{\star}\right\|_{2,\infty}, \tag{70d}$$

$$\max_{1\le l\le n}\left\|\left(\boldsymbol{X}^{t,(l)}\widehat{\boldsymbol{H}}^{t,(l)}-\boldsymbol{X}^{\star}\right)_{l,\cdot}\right\|_{2} \le \left(C_2\rho^{t}\mu r\frac{1}{\sqrt{np}}+C_6\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\right)\left\|\boldsymbol{X}^{\star}\right\|_{2,\infty} \tag{70e}$$

hold for some absolute constants $0<\rho<1$ and $C_1,\cdots,C_{10}>0$. Here, $\widehat{\boldsymbol{H}}^{t,(l)}$ and $\boldsymbol{R}^{t,(l)}$ are orthonormal matrices defined by

$$\widehat{\boldsymbol{H}}^{t,(l)} := \arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}}\left\|\boldsymbol{X}^{t,(l)}\boldsymbol{R}-\boldsymbol{X}^{\star}\right\|_{\mathrm{F}}, \tag{71}$$

$$\boldsymbol{R}^{t,(l)} := \arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}}\left\|\boldsymbol{X}^{t,(l)}\boldsymbol{R}-\boldsymbol{X}^{t}\widehat{\boldsymbol{H}}^{t}\right\|_{\mathrm{F}}. \tag{72}$$

Clearly, the first three hypotheses (70a)–(70c) constitute the conclusion of Theorem 2, i.e., (28). The last two hypotheses (70d) and (70e) are auxiliary properties connecting the true iterates and the auxiliary leave-one-out sequences. Moreover, we summarize below several immediate consequences of (70), which will be useful throughout.

Lemma 10 *Suppose $n^2p\ge C\kappa^3\mu^2r^2n\log n$ for some sufficiently large constant $C>0$, and the noise satisfies (27). Under hypotheses (70), one has*

$$\left\|\boldsymbol{X}^{t}\widehat{\boldsymbol{H}}^{t}-\boldsymbol{X}^{t,(l)}\widehat{\boldsymbol{H}}^{t,(l)}\right\|_{\mathrm{F}} \le 5\kappa\left\|\boldsymbol{X}^{t}\widehat{\boldsymbol{H}}^{t}-\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}\right\|_{\mathrm{F}}, \tag{73a}$$

$$\left\|\boldsymbol{X}^{t,(l)}\widehat{\boldsymbol{H}}^{t,(l)}-\boldsymbol{X}^{\star}\right\|_{\mathrm{F}} \le \left\|\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}-\boldsymbol{X}^{\star}\right\|_{\mathrm{F}} \le \left\{2C_4\rho^{t}\mu r\frac{1}{\sqrt{np}}+2C_1\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\right\}\left\|\boldsymbol{X}^{\star}\right\|_{\mathrm{F}}, \tag{73b}$$

$$\left\|\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}-\boldsymbol{X}^{\star}\right\|_{2,\infty} \le \left\{(C_3+C_5)\,\rho^{t}\mu r\sqrt{\frac{\log n}{np}}+(C_8+C_7)\,\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\right\}\left\|\boldsymbol{X}^{\star}\right\|_{2,\infty}, \tag{73c}$$

$$\left\|\boldsymbol{X}^{t,(l)}\widehat{\boldsymbol{H}}^{t,(l)}-\boldsymbol{X}^{\star}\right\| \le \left\{2C_9\rho^{t}\mu r\frac{1}{\sqrt{np}}+2C_{10}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\right\}\left\|\boldsymbol{X}^{\star}\right\|. \tag{73d}$$

In particular, (73a) *follows from hypotheses* (70c) *and* (70d).

Proof See Appendix B.4. □

In the sequel, we follow the general recipe outlined in Sect. 5 to establish the induction hypotheses. We only need to establish (70b), (70d), and (70e) for the $(t+1)$th iteration, since (70a) and (70c) are established in Sect. 7.1.2. Specifically, we resort to the leave-one-out iterates by showing that: first, the true and the auxiliary iterates remain exceedingly close throughout; second, the lth leave-one-out sequence stays incoherent with $\boldsymbol{e}_l$ due to statistical independence.

- *Step 3(a): proximity between the original and the leave-one-out iterates* We demonstrate that $\boldsymbol{X}^{t+1}$ is well approximated by $\boldsymbol{X}^{t+1,(l)}$, up to proper orthonormal

transforms. This is precisely the induction hypothesis (70d) for the $(t+1)$th iteration.

Lemma 11 *Suppose the sample complexity satisfies $n^2 p \geq C\kappa^4\mu^3 r^3 n \log^3 n$ for some sufficiently large constant $C > 0$, and the noise satisfies (27). Let $\mathcal{E}_t$ be the event where the hypotheses in (70) hold for the tth iteration. Then on some event $\mathcal{E}_{t+1,1} \subseteq \mathcal{E}_t$ obeying $\mathbb{P}(\mathcal{E}_t \cap \mathcal{E}_{t+1,1}^c) = O(n^{-10})$, we have*

$$\left\| \boldsymbol{X}^{t+1}\widehat{\boldsymbol{H}}^{t+1} - \boldsymbol{X}^{t+1,(l)}\boldsymbol{R}^{t+1,(l)} \right\|_{\mathrm{F}} \leq C_3\rho^{t+1}\mu r\sqrt{\frac{\log n}{np}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty} + C_7\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty} \tag{74}$$

provided that $0 < \eta \leq 2/(25\kappa\sigma_{\max})$, $1 - (\sigma_{\min}/5)\cdot\eta \leq \rho < 1$, and $C_7 > 0$ is sufficiently large.

Proof The fact that this difference is well controlled relies heavily on the benign geometric property of the Hessian revealed by Lemma 7. Two important remarks are in order: (1) both points $\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t$ and $\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}$ satisfy (63a); (2) the difference $\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t - \boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}$ forms a valid direction for restricted strong convexity. These two properties together allow us to invoke Lemma 7. See Appendix B.5. □

- *Step 3(b): incoherence of the leave-one-out iterates* Given that $\boldsymbol{X}^{t+1,(l)}$ is sufficiently close to $\boldsymbol{X}^{t+1}$, we turn our attention to establishing the incoherence of this surrogate $\boldsymbol{X}^{t+1,(l)}$ w.r.t. $\boldsymbol{e}_l$. This amounts to proving the induction hypothesis (70e) for the $(t+1)$th iteration.

Lemma 12 *Suppose the sample complexity meets $n^2 p \geq C\kappa^3\mu^3 r^3 n\log^3 n$ for some sufficiently large constant $C > 0$, and the noise satisfies (27). Let $\mathcal{E}_t$ be the event where the hypotheses in (70) hold for the tth iteration. Then on some event $\mathcal{E}_{t+1,2} \subseteq \mathcal{E}_t$ obeying $\mathbb{P}(\mathcal{E}_t \cap \mathcal{E}_{t+1,2}^c) = O(n^{-10})$, we have*

$$\left\| \left(\boldsymbol{X}^{t+1,(l)}\widehat{\boldsymbol{H}}^{t+1,(l)} - \boldsymbol{X}^\star\right)_{l,\cdot} \right\|_2 \leq C_2\rho^{t+1}\mu r\frac{1}{\sqrt{np}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty} + C_6\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty} \tag{75}$$

so long as $0 < \eta \leq 1/\sigma_{\max}$, $1 - (\sigma_{\min}/3)\cdot\eta \leq \rho < 1$, $C_2 \gg \kappa C_9$, and $C_6 \gg \kappa C_{10}/\sqrt{\log n}$.

Proof The key observation is that $\boldsymbol{X}^{t+1,(l)}$ is statistically independent from any sample in the lth row/column of the matrix. Since there are an order of np samples in each row/column, we obtain enough information that helps establish the desired incoherence property. See Appendix B.6. □

- *Step 3(c): combining the bounds* Inequalities (70d) and (70e) taken collectively allow us to establish the induction hypothesis (70b). Specifically, for every $1 \leq l \leq n$, write

$$\left(\boldsymbol{X}^{t+1}\widehat{\boldsymbol{H}}^{t+1}-\boldsymbol{X}^{\star}\right)_{l,\cdot}=\left(\boldsymbol{X}^{t+1}\widehat{\boldsymbol{H}}^{t+1}-\boldsymbol{X}^{t+1,(l)}\widehat{\boldsymbol{H}}^{t+1,(l)}\right)_{l,\cdot}+\left(\boldsymbol{X}^{t+1,(l)}\widehat{\boldsymbol{H}}^{t+1,(l)}-\boldsymbol{X}^{\star}\right)_{l,\cdot},$$

and the triangle inequality gives

$$\left\|\left(\boldsymbol{X}^{t+1}\widehat{\boldsymbol{H}}^{t+1}-\boldsymbol{X}^{\star}\right)_{l,\cdot}\right\|_2\le\left\|\boldsymbol{X}^{t+1}\widehat{\boldsymbol{H}}^{t+1}-\boldsymbol{X}^{t+1,(l)}\widehat{\boldsymbol{H}}^{t+1,(l)}\right\|_{\mathrm{F}}+\left\|\left(\boldsymbol{X}^{t+1,(l)}\widehat{\boldsymbol{H}}^{t+1,(l)}-\boldsymbol{X}^{\star}\right)_{l,\cdot}\right\|_2. \tag{76}$$

The second term has already been bounded by (75). Since we have established the induction hypotheses (70c) and (70d) for the $(t+1)$th iteration, the first term can be bounded by (73a) for the $(t+1)$th iteration, i.e.,

$$\left\|\boldsymbol{X}^{t+1}\widehat{\boldsymbol{H}}^{t+1}-\boldsymbol{X}^{t+1,(l)}\widehat{\boldsymbol{H}}^{t+1,(l)}\right\|_{\mathrm{F}}\le 5\kappa\left\|\boldsymbol{X}^{t+1}\widehat{\boldsymbol{H}}^{t+1}-\boldsymbol{X}^{t+1,(l)}\boldsymbol{R}^{t+1,(l)}\right\|_{\mathrm{F}}.$$

Plugging the above inequality, (74), and (75) into (76), we have

$$\begin{aligned}&\left\|\boldsymbol{X}^{t+1}\widehat{\boldsymbol{H}}^{t+1}-\boldsymbol{X}^{\star}\right\|_{2,\infty}\\&\le 5\kappa\left(C_3\rho^{t+1}\mu r\sqrt{\frac{\log n}{np}}\left\|\boldsymbol{X}^{\star}\right\|_{2,\infty}+\frac{C_7}{\sigma_{\min}}\sigma\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{X}^{\star}\right\|_{2,\infty}\right)\\&\quad+C_2\rho^{t+1}\mu r\frac{1}{\sqrt{np}}\left\|\boldsymbol{X}^{\star}\right\|_{2,\infty}+\frac{C_6}{\sigma_{\min}}\sigma\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{X}^{\star}\right\|_{2,\infty}\\&\le C_5\rho^{t+1}\mu r\sqrt{\frac{\log n}{np}}\left\|\boldsymbol{X}^{\star}\right\|_{2,\infty}+\frac{C_8}{\sigma_{\min}}\sigma\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{X}^{\star}\right\|_{2,\infty}\end{aligned}$$

as long as $C_5/(\kappa C_3+C_2)$ and $C_8/(\kappa C_7+C_6)$ are sufficiently large. This establishes the induction hypothesis (70b). From the deduction above, we see $\mathcal{E}_t\cap\mathcal{E}_{t+1}^c=O(n^{-10})$ and thus finish the proof.

7.4 The Base Case: Spectral Initialization

Finally, we return to check the base case; namely, we aim to show that the spectral initialization satisfies the induction hypotheses (70a)–(70e) for $t=0$. This is accomplished via the following lemma.

Lemma 13 *Suppose the sample size obeys $n^2p\ge C\mu^2r^2n\log n$ for some sufficiently large constant $C>0$, the noise satisfies (27), and $\kappa=\sigma_{\max}/\sigma_{\min}\asymp 1$. Then with probability at least $1-O\left(n^{-10}\right)$, the claims in (70a)–(70e) hold simultaneously for $t=0$.*

Proof This follows by invoking the Davis–Kahan sinΘ theorem [39] as well as the entrywise eigenvector perturbation analysis in [1]. We defer the proof to Appendix B.7. □

8 Analysis for Blind Deconvolution

In this section, we instantiate the general recipe presented in Sect. 5 to blind deconvolution and prove Theorem 3. Without loss of generality, we assume throughout that $\left\|\boldsymbol{h}^{\star}\right\|_2 = \|\boldsymbol{x}^{\star}\|_2 = 1$.

Before presenting the analysis, we first gather some simple facts about the empirical loss function in (32). Recall the definition of z in (33), and for notational simplicity, we write $f(z) = f(\boldsymbol{h}, \boldsymbol{x})$. Since z is complex-valued, we need to resort to Wirtinger calculus; see [18, Section 6] for a brief introduction. The Wirtinger gradient of (32) with respect to $\boldsymbol{h}$ and $\boldsymbol{x}$ are given, respectively, by

$$\nabla_{\boldsymbol{h}} f(z) = \nabla_{\boldsymbol{h}} f(\boldsymbol{h}, \boldsymbol{x}) = \sum_{j=1}^{m} \left(\boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h} \boldsymbol{x}^{\mathsf{H}} \boldsymbol{a}_j - y_j\right) \boldsymbol{b}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}; \tag{77}$$

$$\nabla_{\boldsymbol{x}} f(z) = \nabla_{\boldsymbol{x}} f(\boldsymbol{h}, \boldsymbol{x}) = \sum_{j=1}^{m} \overline{(\boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h} \boldsymbol{x}^{\mathsf{H}} \boldsymbol{a}_j - y_j)} \boldsymbol{a}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}. \tag{78}$$

It is worth noting that the formal Wirtinger gradient contains $\nabla_{\overline{\boldsymbol{h}}} f(\boldsymbol{h}, \boldsymbol{x})$ and $\nabla_{\overline{\boldsymbol{x}}} f(\boldsymbol{h}, \boldsymbol{x})$ as well. Nevertheless, since $f(\boldsymbol{h}, \boldsymbol{x})$ is a real-valued function, the following identities always hold

$$\nabla_{\boldsymbol{h}} f(\boldsymbol{h}, \boldsymbol{x}) = \overline{\nabla_{\overline{\boldsymbol{h}}} f(\boldsymbol{h}, \boldsymbol{x})} \quad \text{and} \quad \nabla_{\boldsymbol{x}} f(\boldsymbol{h}, \boldsymbol{x}) = \overline{\nabla_{\overline{\boldsymbol{x}}} f(\boldsymbol{h}, \boldsymbol{x})}.$$

In light of these observations, one often omits the gradient with respect to the conjugates; correspondingly, the gradient update rule (35) can be written as

$$\boldsymbol{h}^{t+1} = \boldsymbol{h}^t - \frac{\eta}{\|\boldsymbol{x}^t\|_2^2} \sum_{j=1}^{m} \left(\boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^t \boldsymbol{x}^{t\mathsf{H}} \boldsymbol{a}_j - y_j\right) \boldsymbol{b}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^t, \tag{79a}$$

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \frac{\eta}{\left\|\boldsymbol{h}^t\right\|_2^2} \sum_{j=1}^{m} \overline{(\boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^t \boldsymbol{x}^{t\mathsf{H}} \boldsymbol{a}_j - y_j)} \boldsymbol{a}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^t. \tag{79b}$$

We can also compute the Wirtinger Hessian of $f(z)$ as follows,

$$\nabla^2 f(z) = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}^{\mathsf{H}} & \overline{\boldsymbol{A}} \end{bmatrix}, \tag{80}$$

where

$$\boldsymbol{A} = \begin{bmatrix} \sum_{j=1}^{m} \left|\boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}\right|^2 \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} & \sum_{j=1}^{m} \left(\boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h} \boldsymbol{x}^{\mathsf{H}} \boldsymbol{a}_j - y_j\right) \boldsymbol{b}_j \boldsymbol{a}_j^{\mathsf{H}} \\ \sum_{j=1}^{m} \left[\left(\boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h} \boldsymbol{x}^{\mathsf{H}} \boldsymbol{a}_j - y_j\right) \boldsymbol{b}_j \boldsymbol{a}_j^{\mathsf{H}}\right]^{\mathsf{H}} & \sum_{j=1}^{m} \left|\boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}\right|^2 \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \end{bmatrix} \in \mathbb{C}^{2K \times 2K};$$

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{0} & \sum_{j=1}^{m} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h} \left(\boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}\right)^{\top} \\ \sum_{j=1}^{m} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x} \left(\boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}\right)^{\top} & \boldsymbol{0} \end{bmatrix} \in \mathbb{C}^{2K \times 2K}.$$

Last but not least, we say $(\boldsymbol{h}_1, \boldsymbol{x}_1)$ is aligned with $(\boldsymbol{h}_2, \boldsymbol{x}_2)$, if the following holds,

$$\|\boldsymbol{h}_1 - \boldsymbol{h}_2\|_2^2 + \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2^2 = \min_{\alpha \in \mathbb{C}} \left\{ \left\| \frac{1}{\overline{\alpha}} \boldsymbol{h}_1 - \boldsymbol{h}_2 \right\|_2^2 + \|\alpha \boldsymbol{x}_1 - \boldsymbol{x}_2\|_2^2 \right\}.$$

To simplify notations, define $\widetilde{\boldsymbol{z}}^t$ as

$$\widetilde{\boldsymbol{z}}^t = \begin{bmatrix} \widetilde{\boldsymbol{h}}^t \\ \widetilde{\boldsymbol{x}}^t \end{bmatrix} := \begin{bmatrix} \frac{1}{\overline{\alpha^t}} \boldsymbol{h}^t \\ \alpha^t \boldsymbol{x}^t \end{bmatrix} \tag{81}$$

with the alignment parameter α^t given in (38). Then we can see that $\widetilde{\boldsymbol{z}}^t$ is aligned with $\boldsymbol{z}^\star$ and

$$\operatorname{dist}\left(\boldsymbol{z}^t, \boldsymbol{z}^\star\right) = \operatorname{dist}\left(\widetilde{\boldsymbol{z}}^t, \boldsymbol{z}^\star\right) = \left\|\widetilde{\boldsymbol{z}}^t - \boldsymbol{z}^\star\right\|_2.$$

8.1 Step 1: Characterizing Local Geometry in the RIC

8.1.1 Local Geometry

The first step is to characterize the region of incoherence and contraction (RIC), where the empirical loss function enjoys restricted strong convexity and smoothness properties. To this end, we have the following lemma.

Lemma 14 (Restricted strong convexity and smoothness for blind deconvolution) *Let $c > 0$ be a sufficiently small constant and*

$$\delta = c/\log^2 m.$$

Suppose the sample size satisfies $m \geq c_0 \mu^2 K \log^9 m$ for some sufficiently large constant $c_0 > 0$. Then with probability $1 - O\left(m^{-10} + e^{-K} \log m\right)$, the Wirtinger Hessian $\nabla^2 f(\boldsymbol{z})$ obeys

$$\boldsymbol{u}^{\mathsf{H}} \left[\boldsymbol{D} \nabla^2 f(\boldsymbol{z}) + \nabla^2 f(\boldsymbol{z}) \boldsymbol{D} \right] \boldsymbol{u} \geq (1/4) \cdot \|\boldsymbol{u}\|_2^2 \quad \text{and} \quad \left\| \nabla^2 f(\boldsymbol{z}) \right\| \leq 3$$

simultaneously for all

$$\boldsymbol{z} = \begin{bmatrix} \boldsymbol{h} \\ \boldsymbol{x} \end{bmatrix} \quad \text{and} \quad \boldsymbol{u} = \begin{bmatrix} \boldsymbol{h}_1 - \boldsymbol{h}_2 \\ \boldsymbol{x}_1 - \boldsymbol{x}_2 \\ \overline{\boldsymbol{h}_1 - \boldsymbol{h}_2} \\ \overline{\boldsymbol{x}_1 - \boldsymbol{x}_2} \end{bmatrix} \quad \text{and} \quad \boldsymbol{D} = \begin{bmatrix} \gamma_1 \boldsymbol{I}_K & & & \\ & \gamma_2 \boldsymbol{I}_K & & \\ & & \gamma_1 \boldsymbol{I}_K & \\ & & & \gamma_2 \boldsymbol{I}_K \end{bmatrix},$$

where $\boldsymbol{z}$ satisfies

$$\max\left\{ \left\|\boldsymbol{h} - \boldsymbol{h}^\star\right\|_2, \left\|\boldsymbol{x} - \boldsymbol{x}^\star\right\|_2 \right\} \leq \delta; \tag{82a}$$

$$\max_{1\le j\le m}\left|\boldsymbol{a}_j^{\mathsf{H}}\left(\boldsymbol{x}-\boldsymbol{x}^\star\right)\right| \le 2C_3\frac{1}{\log^{3/2} m}; \tag{82b}$$

$$\max_{1\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}\right| \le 2C_4\frac{\mu}{\sqrt{m}}\log^2 m; \tag{82c}$$

$(\boldsymbol{h}_1, \boldsymbol{x}_1)$ is aligned with $(\boldsymbol{h}_2, \boldsymbol{x}_2)$, and they satisfy

$$\max\left\{\left\|\boldsymbol{h}_1-\boldsymbol{h}^\star\right\|_2, \left\|\boldsymbol{h}_2-\boldsymbol{h}^\star\right\|_2, \left\|\boldsymbol{x}_1-\boldsymbol{x}^\star\right\|_2, \left\|\boldsymbol{x}_2-\boldsymbol{x}^\star\right\|_2\right\} \le \delta; \tag{83}$$

and finally, $\boldsymbol{D}$ satisfies for $\gamma_1, \gamma_2 \in \mathbb{R}$,

$$\max\left\{|\gamma_1-1|, |\gamma_2-1|\right\} \le \delta. \tag{84}$$

Here, $C_3, C_4 > 0$ are numerical constants.

Proof See Appendix C.1. □

Lemma 14 characterizes the restricted strong convexity and smoothness of the loss function used in blind deconvolution. To the best of our knowledge, this provides the first characterization regarding geometric properties of the Hessian matrix for blind deconvolution. A few interpretations are in order.

- Conditions (82) specify the region of incoherence and contraction (RIC). In particular, (82a) specifies a neighborhood that is close to the ground truth in ℓ_2 norm, and (82b) and (82c) specify the incoherence region with respect to the sensing vectors $\{\boldsymbol{a}_j\}$ and $\{\boldsymbol{b}_j\}$, respectively.
- Similar to matrix completion, the Hessian matrix is rank-deficient even at the population level. Consequently, we resort to a restricted form of strong convexity by focusing on certain directions. More specifically, these directions can be viewed as the difference between two pre-aligned points that are not far from the truth, which is characterized by (83).
- Finally, the diagonal matrix $\boldsymbol{D}$ accounts for scaling factors that are not too far from 1 (see (84)), which allows us to account for different step sizes employed for $\boldsymbol{h}$ and $\boldsymbol{x}$.

8.1.2 Error Contraction

The restricted strong convexity and smoothness allow us to establish the contraction of the error measured in terms of $\mathrm{dist}(\cdot, \boldsymbol{z}^\star)$ as defined in (34) as long as the iterates stay in the RIC.

Lemma 15 *Suppose the number of measurements satisfies $m \ge C\mu^2 K \log^9 m$ for some sufficiently large constant $C > 0$, and the step size $\eta > 0$ is some sufficiently small constant. There exists an event that does not depend on t and has probability $1 - O\left(m^{-10} + e^{-K}\log m\right)$, such that when it happens and*

$$\mathrm{dist}\left(\boldsymbol{z}^t, \boldsymbol{z}^\star\right) \le \xi, \tag{85a}$$

$$\max_{1\le j\le m}\left|\boldsymbol{a}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^t-\boldsymbol{x}^\star\right)\right|\le C_3\frac{1}{\log^{1.5}m}, \tag{85b}$$

$$\max_{1\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right|\le C_4\frac{\mu}{\sqrt{m}}\log^2 m \tag{85c}$$

hold for some constants $C_3, C_4 > 0$, one has

$$\operatorname{dist}\left(\boldsymbol{z}^{t+1},\boldsymbol{z}^\star\right)\le(1-\eta/16)\operatorname{dist}\left(\boldsymbol{z}^t,\boldsymbol{z}^\star\right).$$

Here, $\widetilde{\boldsymbol{h}}^t$ and $\widetilde{\boldsymbol{x}}^t$ are defined in (81), and $\xi \ll 1/\log^2 m$.

Proof See Appendix C.2. □

As a result, if $\boldsymbol{z}^t$ satisfies condition (85) for all $0\le t\le T$, then

$$\operatorname{dist}\left(\boldsymbol{z}^t,\boldsymbol{z}^\star\right)\le\rho\operatorname{dist}\left(\boldsymbol{z}^{t-1},\boldsymbol{z}^\star\right)\le\rho^t\operatorname{dist}\left(\boldsymbol{z}^0,\boldsymbol{z}^\star\right)\le\rho^t c_1,\qquad 0<t\le T,$$

where $\rho := 1-\eta/16$. Furthermore, similar to the case of phase retrieval (i.e., Lemma 3), as soon as we demonstrate that conditions (85) hold for all $0\le t\le m$, then Theorem 3 holds true. The proof of this claim is exactly the same as for Lemma 3 and is thus omitted for conciseness. In what follows, we focus on establishing (85) for all $0\le t\le m$.

Before concluding this subsection, we make note of another important result that concerns the alignment parameter α^t, which will be useful in the subsequent analysis. Specifically, the alignment parameter sequence $\{\alpha^t\}$ converges linearly to a constant whose magnitude is fairly close to 1, as long as the two initial vectors $\boldsymbol{h}^0$ and $\boldsymbol{x}^0$ have similar ℓ_2 norms and are close to the truth. Given that α^t determines the global scaling of the iterates, this reveals rapid convergence of both $\|\boldsymbol{h}^t\|_2$ and $\|\boldsymbol{x}^t\|_2$, which explains why there is no need to impose extra terms to regularize the ℓ_2 norm as employed in [58,76].

Lemma 16 *When $m>1$ is sufficiently large, the following two claims hold true.*

- *If $\left||\alpha^t|-1\right|\le 1/2$ and $\operatorname{dist}(\boldsymbol{z}^t,\boldsymbol{z}^\star)\le C_1/\log^2 m$, then*

$$\left|\frac{\alpha^{t+1}}{\alpha^t}-1\right|\le c\operatorname{dist}(\boldsymbol{z}^t,\boldsymbol{z}^\star)\le\frac{cC_1}{\log^2 m}$$

  for some absolute constant $c>0$;
- *If $\left||\alpha^0|-1\right|\le 1/4$ and $\operatorname{dist}(\boldsymbol{z}^s,\boldsymbol{z}^\star)\le C_1(1-\eta/16)^s/\log^2 m$ for all $0\le s\le t$, then one has*

$$\left||\alpha^{s+1}|-1\right|\le 1/2,\qquad 0\le s\le t.$$

Proof See Appendix C.2. □

The initial condition $\left||\alpha^0|-1\right|<1/4$ will be guaranteed to hold with high probability by Lemma 19.

8.2 Step 2: Introducing the Leave-One-Out Sequences

As demonstrated by the assumptions in Lemma 15, the key is to show that the whole trajectory lies in the region specified by (85a)–(85c). Once again, the difficulty lies in the statistical dependency between the iterates $\{z^t\}$ and the measurement vectors $\{a_j\}$. We follow the general recipe and introduce the *leave-one-out* sequences, denoted by $\left\{h^{t,(l)}, x^{t,(l)}\right\}_{t\geq 0}$ for each $1 \leq l \leq m$. Specifically, $\left\{h^{t,(l)}, x^{t,(l)}\right\}_{t\geq 0}$ is the gradient sequence operating on the loss function

$$f^{(l)}(h, x) := \sum_{j:j\neq l} \left| b_j^{\mathsf{H}} \left(hx^{\mathsf{H}} - h^{\star}x^{\star\mathsf{H}} \right) a_j \right|^2. \tag{86}$$

The whole sequence is constructed by running gradient descent with spectral initialization on the leave-one-out loss (86). The precise description is supplied in Algorithm 6.

For notational simplicity, we denote $z^{t,(l)} = \begin{bmatrix} h^{t,(l)} \\ x^{t,(l)} \end{bmatrix}$ and use $f(z^{t,(l)}) = f(h^{t,(l)}, x^{t,(l)})$ interchangeably. Define similarly the alignment parameters

$$\alpha^{t,(l)} := \arg\min_{\alpha\in\mathbb{C}} \left\| \frac{1}{\overline{\alpha}} h^{t,(l)} - h^{\star} \right\|_2^2 + \left\| \alpha x^{t,(l)} - x^{\star} \right\|_2^2, \tag{87}$$

and denote $\widetilde{z}^{t,(l)} = \begin{bmatrix} \widetilde{h}^{t,(l)} \\ \widetilde{x}^{t,(l)} \end{bmatrix}$ where

$$\widetilde{h}^{t,(l)} = \frac{1}{\overline{\alpha^{t,(l)}}} h^{t,(l)} \quad \text{and} \quad \widetilde{x}^{t,(l)} = \alpha^{t,(l)} x^{t,(l)}. \tag{88}$$

Algorithm 6 The lth leave-one-out sequence for blind deconvolution

Input: $\{a_j\}_{1\leq j\leq m, j\neq l}$, $\{b_j\}_{1\leq j\leq m, j\neq l}$ and $\{y_j\}_{1\leq j\leq m, j\neq l}$.

Spectral initialization: Let $\sigma_1(M^{(l)})$, $\check{h}^{0,(l)}$ and $\check{x}^{0,(l)}$ be the leading singular value, left and right singular vectors of

$$M^{(l)} := \sum_{j:j\neq l} y_j b_j a_j^{\mathsf{H}},$$

respectively. Set $h^{0,(l)} = \sqrt{\sigma_1(M^{(l)})}\, \check{h}^{0,(l)}$ and $x^{0,(l)} = \sqrt{\sigma_1(M^{(l)})}\, \check{x}^{0,(l)}$.

Gradient updates: **for** $t = 0, 1, 2, \ldots, T-1$ **do**

$$\begin{bmatrix} h^{t+1,(l)} \\ x^{t+1,(l)} \end{bmatrix} = \begin{bmatrix} h^{t,(l)} \\ x^{t,(l)} \end{bmatrix} - \eta \begin{bmatrix} \frac{1}{\|x^{t,(l)}\|_2^2} \nabla_h f^{(l)}(h^{t,(l)}, x^{t,(l)}) \\ \frac{1}{\|h^{t,(l)}\|_2^2} \nabla_x f^{(l)}(h^{t,(l)}, x^{t,(l)}) \end{bmatrix}. \tag{89}$$

8.3 Step 3: Establishing the Incoherence Condition by Induction

As usual, we continue the proof in an inductive manner. For clarity of presentation, we list below the set of induction hypotheses underlying our analysis:

$$\operatorname{dist}\left(\boldsymbol{z}^t, \boldsymbol{z}^\star\right) \le C_1 \frac{1}{\log^2 m}, \tag{90a}$$

$$\max_{1\le l\le m} \operatorname{dist}\left(\boldsymbol{z}^{t,(l)}, \widetilde{\boldsymbol{z}}^t\right) \le C_2 \frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^9 m}{m}}, \tag{90b}$$

$$\max_{1\le l\le m} \left|\boldsymbol{a}_l^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^t - \boldsymbol{x}^\star\right)\right| \le C_3 \frac{1}{\log^{1.5} m}, \tag{90c}$$

$$\max_{1\le l\le m} \left|\boldsymbol{b}_l^{\mathsf{H}} \widetilde{\boldsymbol{h}}^t\right| \le C_4 \frac{\mu}{\sqrt{m}} \log^2 m, \tag{90d}$$

where $\widetilde{\boldsymbol{h}}^t$, $\widetilde{\boldsymbol{x}}^t$, and $\widetilde{\boldsymbol{z}}^t$ are defined in (81). Here, $C_1, C_3 > 0$ are some sufficiently small constants, while $C_2, C_4 > 0$ are some sufficiently large constants. We aim to show that if these hypotheses (90) hold up to the tth iteration, then the same would hold for the $(t+1)$th iteration with exceedingly high probability (e.g., $1 - O(m^{-10})$). The first hypothesis (90a) has already been established in Lemma 15, and hence, the rest of this section focuses on establishing the remaining three. To justify the incoherence hypotheses (90c) and (90d) for the $(t+1)$th iteration, we need to leverage the nice properties of the leave-one-out sequences and establish (90b) first. In the sequel, we follow the steps suggested in the general recipe.

- *Step 3(a): proximity between the original and the leave-one-out iterates* We first justify hypothesis (90b) for the $(t+1)$th iteration via the following lemma.

Lemma 17 *Suppose the sample complexity obeys $m \ge C\mu^2 K \log^9 m$ for some sufficiently large constant $C > 0$. Let $\mathcal{E}_t$ be the event where hypotheses (90a)–(90d) hold for the tth iteration. Then on an event $\mathcal{E}_{t+1,1} \subseteq \mathcal{E}_t$ obeying $\mathbb{P}(\mathcal{E}_t \cap \mathcal{E}_{t+1,1}^c) = O(m^{-10} + me^{-cK})$ for some constant $c > 0$, one has*

$$\max_{1\le l\le m} \operatorname{dist}\left(\boldsymbol{z}^{t+1,(l)}, \widetilde{\boldsymbol{z}}^{t+1}\right) \le C_2 \frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^9 m}{m}}$$

and

$$\max_{1\le l\le m} \left\|\widetilde{\boldsymbol{z}}^{t+1,(l)} - \widetilde{\boldsymbol{z}}^{t+1}\right\|_2 \lesssim C_2 \frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^9 m}{m}},$$

provided that the step size $\eta > 0$ is some sufficiently small constant.

Proof As usual, this result follows from the restricted strong convexity, which forces the distance between the two sequences of interest to be contractive. See Appendix C.3. □

- *Step 3(b): incoherence of the leave-one-out iterate $\boldsymbol{x}^{t+1,(l)}$ w.r.t. $\boldsymbol{a}_l$* Next, we show that the leave-one-out iterate $\widetilde{\boldsymbol{x}}^{t+1,(l)}$—which is independent of $\boldsymbol{a}_l$—is incoherent w.r.t. $\boldsymbol{a}_l$ in the sense that

$$\left|\boldsymbol{a}_l^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^{t+1,(l)}-\boldsymbol{x}^\star\right)\right| \le 10C_1\frac{1}{\log^{3/2} m} \tag{91}$$

with probability exceeding $1-O\left(m^{-10}+e^{-K}\log m\right)$. To see why, use the statistical independence and the standard Gaussian concentration inequality to show that

$$\max_{1\le l\le m}\left|\boldsymbol{a}_l^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^{t+1,(l)}-\boldsymbol{x}^\star\right)\right| \le 5\sqrt{\log m}\max_{1\le l\le m}\left\|\widetilde{\boldsymbol{x}}^{t+1,(l)}-\boldsymbol{x}^\star\right\|_2$$

with probability exceeding $1-O(m^{-10})$. It then follows from the triangle inequality that

$$\begin{aligned}\left\|\widetilde{\boldsymbol{x}}^{t+1,(l)}-\boldsymbol{x}^\star\right\|_2 &\le \left\|\widetilde{\boldsymbol{x}}^{t+1,(l)}-\widetilde{\boldsymbol{x}}^{t+1}\right\|_2+\left\|\widetilde{\boldsymbol{x}}^{t+1}-\boldsymbol{x}^\star\right\|_2\\ &\overset{\text{(i)}}{\le} CC_2\frac{\mu}{\sqrt{m}}\sqrt{\frac{\mu^2K\log^9 m}{m}}+C_1\frac{1}{\log^2 m}\\ &\overset{\text{(ii)}}{\le} 2C_1\frac{1}{\log^2 m},\end{aligned}$$

where (i) follows from Lemmas 15 and 17 and (ii) holds as soon as $m/(\mu^2\sqrt{K}\log^{13/2} m)$ is sufficiently large. Combining the preceding two bounds establishes (91).

- *Step 3(c): combining the bounds to show incoherence of* $\boldsymbol{x}^{t+1}$ *w.r.t.* $\{\boldsymbol{a}_l\}$ The above bounds immediately allow us to conclude that

$$\max_{1\le l\le m}\left|\boldsymbol{a}_l^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^{t+1}-\boldsymbol{x}^\star\right)\right| \le C_3\frac{1}{\log^{3/2} m}$$

with probability at least $1-O\left(m^{-10}+e^{-K}\log m\right)$, which is exactly hypothesis (90c) for the $(t+1)$th iteration. Specifically, for each $1\le l\le m$, the triangle inequality yields

$$\begin{aligned}\left|\boldsymbol{a}_l^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^{t+1}-\boldsymbol{x}^\star\right)\right| &\le \left|\boldsymbol{a}_l^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^{t+1}-\widetilde{\boldsymbol{x}}^{t+1,(l)}\right)\right|+\left|\boldsymbol{a}_l^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^{t+1,(l)}-\boldsymbol{x}^\star\right)\right|\\ &\overset{\text{(i)}}{\le} \|\boldsymbol{a}_l\|_2\left\|\widetilde{\boldsymbol{x}}^{t+1}-\widetilde{\boldsymbol{x}}^{t+1,(l)}\right\|_2+\left|\boldsymbol{a}_l^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^{t+1,(l)}-\boldsymbol{x}^\star\right)\right|\\ &\overset{\text{(ii)}}{\le} 3\sqrt{K}\cdot CC_2\frac{\mu}{\sqrt{m}}\sqrt{\frac{\mu^2K\log^9 m}{m}}+10C_1\frac{1}{\log^{3/2} m}\\ &\overset{\text{(iii)}}{\le} C_3\frac{1}{\log^{3/2} m}.\end{aligned}$$

Here (i) follows from Cauchy–Schwarz; (ii) is a consequence of (190), Lemma 17, and bound (91); and the last inequality holds as long as $m/(\mu^2K\log^6 m)$ is sufficiently large and $C_3\ge 11C_1$.

- *Step 3(d): incoherence of $\boldsymbol{h}^{t+1}$ w.r.t. $\{\boldsymbol{b}_l\}$* It remains to justify that $\boldsymbol{h}^{t+1}$ is also incoherent w.r.t. its associated design vectors $\{\boldsymbol{b}_l\}$. This proof of this step, however, is much more involved and challenging, due to the deterministic nature of the $\boldsymbol{b}_l$'s. As a result, we would need to "propagate" the randomness brought about by $\{\boldsymbol{a}_l\}$ to $\boldsymbol{h}^{t+1}$ in order to facilitate the analysis. The result is summarized as follows.

Lemma 18 *Suppose that the sample complexity obeys $m \geq C\mu^2 K \log^9 m$ for some sufficiently large constant $C > 0$. Let $\mathcal{E}_t$ be the event where hypotheses (90a)–(90d) hold for the tth iteration. Then on an event $\mathcal{E}_{t+1,2} \subseteq \mathcal{E}_t$ obeying $\mathbb{P}(\mathcal{E}_t \cap \mathcal{E}_{t+1,2}^c) = O(m^{-10})$, one has*

$$\max_{1\leq l\leq m} \left|\boldsymbol{b}_l^{\mathsf{H}} \widetilde{\boldsymbol{h}}^{t+1}\right| \leq C_4 \frac{\mu}{\sqrt{m}} \log^2 m$$

as long as C_4 is sufficiently large and $\eta > 0$ is taken to be some sufficiently small constant.

Proof The key idea is to divide $\{1, \cdots, m\}$ into consecutive bins each of size poly $\log(m)$, and to exploit the randomness (namely, the randomness from $\boldsymbol{a}_l$) within each bin. This binning idea is crucial in ensuring that the incoherence measure of interest does not blow up as t increases. See Appendix C.4. □

With these steps in place, we conclude the proof of Theorem 3 via induction and the union bound.

8.4 The Base Case: Spectral Initialization

In order to finish the induction steps, we still need to justify the induction hypotheses for the base cases; namely, we need to show that the spectral initializations $\boldsymbol{z}^0$ and $\left\{\boldsymbol{z}^{0,(l)}\right\}_{1\leq l\leq m}$ satisfy the induction hypotheses (90) at $t = 0$.

To start with, the initializations are sufficiently close to the truth when measured by the ℓ_2 norm, as summarized by the following lemma.

Lemma 19 *Fix any small constant $\xi > 0$. Suppose the sample size obeys $m \geq C\mu^2 K \log^2 m/\xi^2$ for some sufficiently large constant $C > 0$. Then with probability at least $1 - O(m^{-10})$, we have*

$$\min_{\alpha\in\mathbb{C},|\alpha|=1} \left\{\left\|\alpha\boldsymbol{h}^0 - \boldsymbol{h}^\star\right\|_2 + \left\|\alpha\boldsymbol{x}^0 - \boldsymbol{x}^\star\right\|_2\right\} \leq \xi \quad \text{and} \tag{92}$$

$$\min_{\alpha\in\mathbb{C},|\alpha|=1} \left\{\left\|\alpha\boldsymbol{h}^{0,(l)} - \boldsymbol{h}^\star\right\|_2 + \left\|\alpha\boldsymbol{x}^{0,(l)} - \boldsymbol{x}^\star\right\|_2\right\} \leq \xi, \quad 1 \leq l \leq m, \tag{93}$$

and $||\alpha_0| - 1| \leq 1/4$.

Proof This follows from Wedin's sinΘ theorem [121] and [76, Lemma 5.20]. See Appendix C.5. □

From the definition of $\mathrm{dist}(\cdot,\cdot)$ (cf. (34)), we immediately have

$$\begin{aligned}\mathrm{dist}\big(z^0, z^\star\big) &= \min_{\alpha\in\mathbb{C}} \sqrt{\left\|\frac{1}{\overline{\alpha}}\boldsymbol{h}-\boldsymbol{h}^\star\right\|_2^2 + \|\alpha\boldsymbol{x}-\boldsymbol{x}^\star\|_2^2} \\ &\overset{\text{(i)}}{\le} \min_{\alpha\in\mathbb{C}} \left\{\left\|\frac{1}{\overline{\alpha}}\boldsymbol{h}-\boldsymbol{h}^\star\right\|_2 + \left\|\alpha\boldsymbol{x}-\boldsymbol{x}^\star\right\|_2\right\} \\ &\overset{\text{(ii)}}{\le} \min_{\alpha\in\mathbb{C},|\alpha|=1} \left\{\left\|\alpha\boldsymbol{h}^0-\boldsymbol{h}^\star\right\|_2 + \left\|\alpha\boldsymbol{x}^0-\boldsymbol{x}^\star\right\|_2\right\} \overset{\text{(iii)}}{\le} C_1\frac{1}{\log^2 m}, \end{aligned} \tag{94}$$

as long as $m \ge C\mu^2 K\log^6 m$ for some sufficiently large constant $C > 0$. Here (i) follows from the elementary inequality that $a^2+b^2 \le (a+b)^2$ for positive a and b, (ii) holds since the feasible set of the latter one is strictly smaller, and (iii) follows directly from Lemma 19. This finishes the proof of (90a) for $t = 0$. Similarly, with high probability we have

$$\begin{aligned}&\mathrm{dist}\big(z^{0,(l)}, z^\star\big) \quad \le \min_{\alpha\in\mathbb{C},|\alpha|=1} \left\{\left\|\alpha\boldsymbol{h}^{0,(l)}-\boldsymbol{h}^\star\right\|_2 + \left\|\alpha\boldsymbol{x}^{0,(l)}-\boldsymbol{x}^\star\right\|_2\right\} \lesssim \frac{1}{\log^2 m}, \\ &1\le l\le m. \end{aligned} \tag{95}$$

Next, when properly aligned, the true initial estimate z^0 and the leave-one-out estimate $z^{0,(l)}$ are expected to be sufficiently close, as claimed by the following lemma. Along the way, we show that $\boldsymbol{h}^0$ is incoherent w.r.t. the sampling vectors $\{\boldsymbol{b}_l\}$. This establishes (90b) and (90d) for $t = 0$.

Lemma 20 *Suppose that $m \ge C\mu^2 K\log^3 m$ for some sufficiently large constant $C > 0$. Then with probability at least $1 - O(m^{-10})$, one has*

$$\max_{1\le l\le m} \mathrm{dist}\big(z^{0,(l)}, \widetilde{z}^0\big) \le C_2\frac{\mu}{\sqrt{m}}\sqrt{\frac{\mu^2 K\log^5 m}{m}} \tag{96}$$

and

$$\max_{1\le l\le m} \left|\boldsymbol{b}_l^{\mathsf{H}}\widetilde{\boldsymbol{h}}^0\right| \le C_4\frac{\mu\log^2 m}{\sqrt{m}}. \tag{97}$$

Proof The key is to establish that $\mathrm{dist}\big(z^{0,(l)}, \widetilde{z}^0\big)$ can be upper bounded by some linear scaling of $\left|\boldsymbol{b}_l^{\mathsf{H}}\widetilde{\boldsymbol{h}}^0\right|$, and vice versa. This allows us to derive bounds simultaneously for both quantities. See Appendix C.6. □

Finally, we establish (90c) regarding the incoherence of $\boldsymbol{x}^0$ with respect to the design vectors $\{\boldsymbol{a}_l\}$.

Lemma 21 *Suppose that $m \ge C\mu^2 K\log^6 m$ for some sufficiently large constant $C > 0$. Then with probability exceeding $1 - O(m^{-10})$, we have*

$$\max_{1\le l\le m}\left|\boldsymbol{a}_l^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^0-\boldsymbol{x}^\star\right)\right|\le C_3\frac{1}{\log^{1.5}m}.$$

Proof See Appendix C.7. □

9 Discussion

This paper showcases an important phenomenon in nonconvex optimization: even without explicit enforcement of regularization, the vanilla form of gradient descent effectively achieves implicit regularization for a large family of statistical estimation problems. We believe this phenomenon arises in problems far beyond the three cases studied herein, and our results are initial steps toward understanding this fundamental phenomenon. There are numerous avenues open for future investigation, and we point out a few of them.

- *Improving sample complexity* In the current paper, the required sample complexity $O\left(\mu^3 r^3 n\log^3 n\right)$ for matrix completion is suboptimal when the rank r of the underlying matrix is large. While this allows us to achieve a dimension-free iteration complexity, it is slightly higher than the sample complexity derived for regularized gradient descent in [32]. We expect our results continue to hold under lower sample complexity $O\left(\mu^2 r^2 n\log n\right)$, but it calls for a more refined analysis (e.g., a generic chaining argument).
- *Leave-one-out tricks for more general designs* So far, our focus is on independent designs, including the i.i.d. Gaussian design adopted in phase retrieval and partially in blind deconvolution, as well as the independent sampling mechanism in matrix completion. Such independence property creates some sort of "statistical homogeneity," for which the leave-one-out argument works beautifully. It remains unclear how to generalize such leave-one-out tricks for more general designs (e.g., more general sampling patterns in matrix completion and more structured Fourier designs in phase retrieval and blind deconvolution). In fact, the readers can already get a flavor of this issue in the analysis of blind deconvolution, where the Fourier design vectors require much more delicate treatments than purely Gaussian designs.
- *Uniform stability* The leave-one-out perturbation argument is established upon a basic fact: when we exclude one sample from consideration, the resulting estimates/predictions do not deviate much from the original ones. This leave-one-out stability bears similarity to the notion of uniform stability studied in statistical learning theory [8]. We expect our analysis framework to be helpful for analyzing other learning algorithms that are uniformly stable.
- *Other iterative methods and other loss functions* The focus of the current paper has been the analysis of vanilla GD tailored to the natural squared loss. This is by no means to advocate GD as the top-performing algorithm in practice; rather, we are using this simple algorithm to isolate some seemingly pervasive phenomena (i.e., implicit regularization) that generic optimization theory fails to account for. The simplicity of vanilla GD makes it an ideal object to initiate such discussions. That being said, practitioners should definitely explore as many algorithmic alter-

natives as possible before settling on a particular algorithm. Take phase retrieval for example: iterative methods other than GD and/or algorithms tailored to other loss functions have been proposed in the nonconvex optimization literature, including but not limited to alternating minimization, block coordinate descent, and subgradient methods and prox-linear methods tailed to nonsmooth losses. It would be interesting to develop a full theoretical understanding of a broader class of iterative algorithms, and to conduct a careful comparison regarding which loss functions lead to the most desirable practical performance.

- *Connections to deep learning?* We have focused on nonlinear systems that are bilinear or quadratic in this paper. Deep learning formulations/architectures, highly nonlinear, are notorious for their daunting nonconvex geometry. However, iterative methods including stochastic gradient descent have enjoyed enormous practical success in learning neural networks (e.g., [46,103,132]), even when the architecture is significantly over-parameterized without explicit regularization. We hope the message conveyed in this paper for several simple statistical models can shed light on why simple forms of gradient descent and variants work so well in learning complicated neural networks.

Finally, while the present paper provides a general recipe for problem-specific analyses of nonconvex algorithms, we acknowledge that a unified theory of this kind has yet to be developed. As a consequence, each problem requires delicate and somewhat lengthy analyses of its own. It would certainly be helpful if one could single out a few stylized structural properties/elements (like sparsity and incoherence in compressed sensing [13]) that enable near-optimal performance guarantees through an overarching method of analysis; with this in place, one would not need to start each problem from scratch. Having said that, we believe that our current theory elucidates a few ingredients (e.g., the region of incoherence and leave-one-out stability) that might serve as crucial building blocks for such a general theory. We invite the interested readers to contribute toward this path forward.

Acknowledgements Y. Chen is supported in part by the AFOSR YIP Award FA9550-19-1-0030, ONR Grant N00014-19-1-2120, ARO Grant W911NF-18-1-0303, NSF Grant CCF-1907661, and the Princeton SEAS innovation award. Y. Chi is supported in part by the Grants AFOSR FA9550-15-1-0205, ONR N00014-18-1-2142 and N00014-19-1-2404, ARO W911NF-18-1-0303, NSF CCF-1826519, ECCS-1818571, CCF-1806154. Y. Chen thanks Yudong Chen for inspiring discussions about matrix completion.

A Proofs for Phase Retrieval

Before proceeding, we gather a few simple facts. The standard concentration inequality for χ^2 random variables together with the union bound reveals that the sampling vectors $\{\boldsymbol{a}_j\}$ obey

$$\max_{1\le j\le m} \left\|\boldsymbol{a}_j\right\|_2 \le \sqrt{6n} \tag{98}$$

with probability at least $1 - O(me^{-1.5n})$. In addition, standard Gaussian concentration inequalities give

$$\max_{1\le j\le m} \left|\boldsymbol{a}_j^\top \boldsymbol{x}^\star\right| \le 5\sqrt{\log n} \tag{99}$$

with probability exceeding $1 - O(mn^{-10})$.

A.1 Proof of Lemma 1

We start with the smoothness bound, namely $\nabla^2 f(\boldsymbol{x}) \preceq O(\log n)\cdot \boldsymbol{I}_n$. It suffices to prove the upper bound $\left\|\nabla^2 f(\boldsymbol{x})\right\| \lesssim \log n$. To this end, we first decompose the Hessian (cf. (44)) into three components as follows:

$$\nabla^2 f(\boldsymbol{x}) = \underbrace{\frac{3}{m}\sum_{j=1}^m \left[\left(\boldsymbol{a}_j^\top \boldsymbol{x}\right)^2 - \left(\boldsymbol{a}_j^\top \boldsymbol{x}^\star\right)^2\right]\boldsymbol{a}_j\boldsymbol{a}_j^\top}_{:=\boldsymbol{\Lambda}_1}$$
$$+ \underbrace{\frac{2}{m}\sum_{j=1}^m \left(\boldsymbol{a}_j^\top \boldsymbol{x}^\star\right)^2 \boldsymbol{a}_j\boldsymbol{a}_j^\top - 2\left(\boldsymbol{I}_n + 2\boldsymbol{x}^\star\boldsymbol{x}^{\star\top}\right)}_{:=\boldsymbol{\Lambda}_2} + \underbrace{2\left(\boldsymbol{I}_n + 2\boldsymbol{x}^\star\boldsymbol{x}^{\star\top}\right)}_{:=\boldsymbol{\Lambda}_3},$$

where we have used $y_j = (\boldsymbol{a}_j^\top \boldsymbol{x}^\star)^2$. In the sequel, we control the three terms $\boldsymbol{\Lambda}_1$, $\boldsymbol{\Lambda}_2$, and $\boldsymbol{\Lambda}_3$ in reverse order.

- The third term $\boldsymbol{\Lambda}_3$ can be easily bounded by

$$\|\boldsymbol{\Lambda}_3\| \le 2\left(\|\boldsymbol{I}_n\| + 2\left\|\boldsymbol{x}^\star\boldsymbol{x}^{\star\top}\right\|\right) = 6.$$

- The second term $\boldsymbol{\Lambda}_2$ can be controlled by means of Lemma 32:

$$\|\boldsymbol{\Lambda}_2\| \le 2\delta$$

  for an arbitrarily small constant $\delta > 0$, as long as $m \ge c_0 n\log n$ for c_0 sufficiently large.
- It thus remains to control $\boldsymbol{\Lambda}_1$. Toward this, we discover that

$$\|\boldsymbol{\Lambda}_1\| \le \left\|\frac{3}{m}\sum_{j=1}^m \left|\boldsymbol{a}_j^\top\left(\boldsymbol{x}-\boldsymbol{x}^\star\right)\right|\left|\boldsymbol{a}_j^\top\left(\boldsymbol{x}+\boldsymbol{x}^\star\right)\right|\boldsymbol{a}_j\boldsymbol{a}_j^\top\right\|. \tag{100}$$

  Under the assumption $\max_{1\le j\le m}\left|\boldsymbol{a}_j^\top(\boldsymbol{x}-\boldsymbol{x}^\star)\right| \le C_2\sqrt{\log n}$ and fact (99), we can also obtain

$$\max_{1\le j\le m}\left|\boldsymbol{a}_j^\top\left(\boldsymbol{x}+\boldsymbol{x}^\star\right)\right| \le 2\max_{1\le j\le m}\left|\boldsymbol{a}_j^\top\boldsymbol{x}^\star\right| + \max_{1\le j\le m}\left|\boldsymbol{a}_j^\top(\boldsymbol{x}-\boldsymbol{x}^\star)\right|$$
$$\le (10 + C_2)\sqrt{\log n}.$$

Substitution into (100) leads to

$$\|\boldsymbol{\Lambda}_1\| \le 3C_2\,(10+C_2)\log n \cdot \left\|\frac{1}{m}\sum_{j=1}^{m}\boldsymbol{a}_j\boldsymbol{a}_j^\top\right\| \le 4C_2\,(10+C_2)\log n,$$

where the last inequality is a direct consequence of Lemma 31.

Combining the above bounds on $\boldsymbol{\Lambda}_1$, $\boldsymbol{\Lambda}_2$, and $\boldsymbol{\Lambda}_3$ yields

$$\begin{aligned}\left\|\nabla^2 f(\boldsymbol{x})\right\| &\le \|\boldsymbol{\Lambda}_1\| + \|\boldsymbol{\Lambda}_2\| + \|\boldsymbol{\Lambda}_3\| \le 4C_2\,(10+C_2)\log n + 2\delta + 6\\ &\le 5C_2\,(10+C_2)\log n,\end{aligned}$$

as long as n is sufficiently large. This establishes the claimed smoothness property.

Next, we move on to the strong convexity lower bound. Picking a constant $C > 0$ and enforcing proper truncation, we get

$$\begin{aligned}\nabla^2 f(\boldsymbol{x}) &= \frac{1}{m}\sum_{j=1}^{m}\left[3\left(\boldsymbol{a}_j^\top\boldsymbol{x}\right)^2 - y_j\right]\boldsymbol{a}_j\boldsymbol{a}_j^\top\\ &\succeq \underbrace{\frac{3}{m}\sum_{j=1}^{m}\left(\boldsymbol{a}_j^\top\boldsymbol{x}\right)^2\mathbb{1}_{\left\{\left|\boldsymbol{a}_j^\top\boldsymbol{x}\right|\le C\right\}}\boldsymbol{a}_j\boldsymbol{a}_j^\top}_{:=\boldsymbol{\Lambda}_4} - \underbrace{\frac{1}{m}\sum_{j=1}^{m}\left(\boldsymbol{a}_j^\top\boldsymbol{x}^\star\right)^2\boldsymbol{a}_j\boldsymbol{a}_j^\top}_{:=\boldsymbol{\Lambda}_5}.\end{aligned}$$

We begin with the simpler term $\boldsymbol{\Lambda}_5$. Lemma 32 implies that with probability at least $1 - O(n^{-10})$,

$$\left\|\boldsymbol{\Lambda}_5 - \left(\boldsymbol{I}_n + 2\boldsymbol{x}^\star\boldsymbol{x}^{\star\top}\right)\right\| \le \delta$$

holds for any small constant $\delta > 0$, as long as $m/(n\log n)$ is sufficiently large. This reveals that

$$\boldsymbol{\Lambda}_5 \preceq (1+\delta)\cdot\boldsymbol{I}_n + 2\boldsymbol{x}^\star\boldsymbol{x}^{\star\top}.$$

To bound $\boldsymbol{\Lambda}_4$, invoke Lemma 33 to conclude that with probability at least $1 - c_3e^{-c_2m}$ (for some constants $c_2, c_3 > 0$),

$$\left\|\boldsymbol{\Lambda}_4 - 3\left(\beta_1\boldsymbol{x}\boldsymbol{x}^\top + \beta_2\|\boldsymbol{x}\|_2^2\boldsymbol{I}_n\right)\right\| \le \delta\|\boldsymbol{x}\|_2^2$$

for any small constant $\delta > 0$, provided that m/n is sufficiently large. Here,

$$\beta_1 := \mathbb{E}\left[\xi^4\,\mathbb{1}_{\{|\xi|\le C\}}\right] - \mathbb{E}\left[\xi^2\,\mathbb{1}_{|\xi|\le C}\right] \quad\text{and}\quad \beta_2 := \mathbb{E}\left[\xi^2\,\mathbb{1}_{|\xi|\le C}\right],$$

where the expectation is taken with respect to $\xi \sim \mathcal{N}(0,1)$. By the assumption $\|\boldsymbol{x} - \boldsymbol{x}^\star\|_2 \le 2C_1$, one has

$$\|\boldsymbol{x}\|_2 \le 1 + 2C_1, \quad \left|\|\boldsymbol{x}\|_2^2 - \|\boldsymbol{x}^\star\|_2^2\right| \le 2C_1\,(4C_1+1)\,, \quad \left\|\boldsymbol{x}^\star\boldsymbol{x}^{\star\top} - \boldsymbol{x}\boldsymbol{x}^\top\right\| \le 6C_1\,(4C_1+1)\,,$$

which leads to

$$\begin{aligned}
&\left\|\boldsymbol{\Lambda}_4 - 3\left(\beta_1\boldsymbol{x}^\star\boldsymbol{x}^{\star\top} + \beta_2\boldsymbol{I}_n\right)\right\| \\
&\quad\le \left\|\boldsymbol{\Lambda}_4 - 3\left(\beta_1\boldsymbol{x}\boldsymbol{x}^\top + \beta_2\|\boldsymbol{x}\|_2^2\boldsymbol{I}_n\right)\right\| \\
&\qquad + 3\left\|\left(\beta_1\boldsymbol{x}^\star\boldsymbol{x}^{\star\top} + \beta_2\boldsymbol{I}_n\right) - \left(\beta_1\boldsymbol{x}\boldsymbol{x}^\top + \beta_2\|\boldsymbol{x}\|_2^2\boldsymbol{I}_n\right)\right\| \\
&\quad\le \delta\|\boldsymbol{x}\|_2^2 + 3\beta_1\left\|\boldsymbol{x}^\star\boldsymbol{x}^{\star\top} - \boldsymbol{x}\boldsymbol{x}^\top\right\| + 3\beta_2\left\|\boldsymbol{I}_n - \|\boldsymbol{x}\|_2^2\boldsymbol{I}_n\right\| \\
&\quad\le \delta\,(1+2C_1)^2 + 18\beta_1 C_1\,(4C_1+1) + 6\beta_2 C_1\,(4C_1+1)\,.
\end{aligned}$$

This further implies

$$\begin{aligned}
\boldsymbol{\Lambda}_4 &\succeq 3\left(\beta_1\boldsymbol{x}^\star\boldsymbol{x}^{\star\top} + \beta_2\boldsymbol{I}_n\right) \\
&\quad - \left[\delta\,(1+2C_1)^2 + 18\beta_1 C_1\,(4C_1+1) + 6\beta_2 C_1\,(4C_1+1)\right]\boldsymbol{I}_n.
\end{aligned}$$

Recognizing that β_1 (resp. β_2) approaches 2 (resp. 1) as C grows, we can thus take C_1 small enough and C large enough to guarantee that

$$\boldsymbol{\Lambda}_4 \succeq 5\boldsymbol{x}^\star\boldsymbol{x}^{\star\top} + 2\boldsymbol{I}_n.$$

Putting the preceding two bounds on $\boldsymbol{\Lambda}_4$ and $\boldsymbol{\Lambda}_5$ together yields

$$\nabla^2 f\,(\boldsymbol{x}) \succeq 5\boldsymbol{x}^\star\boldsymbol{x}^{\star\top} + 2\boldsymbol{I}_n - \left[(1+\delta)\cdot\boldsymbol{I}_n + 2\boldsymbol{x}^\star\boldsymbol{x}^{\star\top}\right] \succeq (1/2)\cdot\boldsymbol{I}_n$$

as claimed.

A.2 Proof of Lemma 2

Using the update rule (cf. (17)) as well as the fundamental theorem of calculus [70, Chapter XIII, Theorem 4.2], we get

$$\begin{aligned}
\boldsymbol{x}^{t+1} - \boldsymbol{x}^\star &= \boldsymbol{x}^t - \eta\nabla f\left(\boldsymbol{x}^t\right) - \left[\boldsymbol{x}^\star - \eta\nabla f\left(\boldsymbol{x}^\star\right)\right] \\
&= \left[\boldsymbol{I}_n - \eta\int_0^1 \nabla^2 f\left(\boldsymbol{x}\left(\tau\right)\right)\mathrm{d}\tau\right]\left(\boldsymbol{x}^t - \boldsymbol{x}^\star\right),
\end{aligned}$$

where we denote $\boldsymbol{x}\left(\tau\right) = \boldsymbol{x}^\star + \tau(\boldsymbol{x}^t - \boldsymbol{x}^\star)$, $0 \le \tau \le 1$. Here, the first equality makes use of the fact that $\nabla f(\boldsymbol{x}^\star) = \boldsymbol{0}$. Under condition (45), it is self-evident that for all $0 \le \tau \le 1$,

$$\left\|\boldsymbol{x}\left(\tau\right)-\boldsymbol{x}^{\star}\right\|_2=\|\tau(\boldsymbol{x}^t-\boldsymbol{x}^{\star})\|_2\le 2C_1\qquad\text{and}$$
$$\max_{1\le l\le m}\left|\boldsymbol{a}_l^{\top}\left(\boldsymbol{x}(\tau)-\boldsymbol{x}^{\star}\right)\right|\le\max_{1\le l\le m}\left|\boldsymbol{a}_l^{\top}\tau\left(\boldsymbol{x}^t-\boldsymbol{x}^{\star}\right)\right|\le C_2\sqrt{\log n}.$$

This means that for all $0\le\tau\le 1$,

$$(1/2)\cdot\boldsymbol{I}_n\preceq\nabla^2 f\left(\boldsymbol{x}(\tau)\right)\preceq\left[5C_2\left(10+C_2\right)\log n\right]\cdot\boldsymbol{I}_n$$

in view of Lemma 1. Picking $\eta\le 1/\left[5C_2\left(10+C_2\right)\log n\right]$ (and hence $\|\eta\nabla^2 f(\boldsymbol{x}(\tau))\|\le 1$), one sees that

$$\boldsymbol{0}\preceq\boldsymbol{I}_n-\eta\int_0^1\nabla^2 f\left(\boldsymbol{x}\left(\tau\right)\right)\mathrm{d}\tau\preceq(1-\eta/2)\cdot\boldsymbol{I}_n,$$

which immediately yields

$$\left\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^{\star}\right\|_2\le\left\|\boldsymbol{I}_n-\eta\int_0^1\nabla^2 f\left(\boldsymbol{x}\left(\tau\right)\right)\mathrm{d}\tau\right\|\cdot\left\|\boldsymbol{x}^t-\boldsymbol{x}^{\star}\right\|_2\le(1-\eta/2)\left\|\boldsymbol{x}^t-\boldsymbol{x}^{\star}\right\|_2.$$

A.3 Proof of Lemma 3

We start with proving (19a). For all $0\le t\le T_0$, invoke Lemma 2 recursively with conditions (47) to reach

$$\left\|\boldsymbol{x}^t-\boldsymbol{x}^{\star}\right\|_2\le(1-\eta/2)^t\left\|\boldsymbol{x}^0-\boldsymbol{x}^{\star}\right\|_2\le C_1(1-\eta/2)^t\left\|\boldsymbol{x}^{\star}\right\|_2.\tag{101}$$

This finishes the proof of (19a) for $0\le t\le T_0$ and also reveals that

$$\left\|\boldsymbol{x}^{T_0}-\boldsymbol{x}^{\star}\right\|_2\le C_1(1-\eta/2)^{T_0}\left\|\boldsymbol{x}^{\star}\right\|_2\ll\frac{1}{n}\left\|\boldsymbol{x}^{\star}\right\|_2,\tag{102}$$

provided that $\eta\asymp 1/\log n$. Applying the Cauchy–Schwarz inequality and fact (98) indicate that

$$\max_{1\le l\le m}\left|\boldsymbol{a}_l^{\top}\left(\boldsymbol{x}^{T_0}-\boldsymbol{x}^{\star}\right)\right|\le\max_{1\le l\le m}\|\boldsymbol{a}_l\|_2\|\boldsymbol{x}^{T_0}-\boldsymbol{x}^{\star}\|_2\le\sqrt{6n}\cdot\frac{1}{n}\|\boldsymbol{x}^{\star}\|_2\ll C_2\sqrt{\log n},$$

leading to the satisfaction of (45). Therefore, invoking Lemma 2 yields

$$\left\|\boldsymbol{x}^{T_0+1}-\boldsymbol{x}^{\star}\right\|_2\le(1-\eta/2)\left\|\boldsymbol{x}^{T_0}-\boldsymbol{x}^{\star}\right\|_2\ll\frac{1}{n}\|\boldsymbol{x}^{\star}\|_2.$$

One can then repeat this argument to arrive at for all $t>T_0$

$$\left\|\boldsymbol{x}^t-\boldsymbol{x}^{\star}\right\|_2\le(1-\eta/2)^t\left\|\boldsymbol{x}^0-\boldsymbol{x}^{\star}\right\|_2\le C_1\left(1-\eta/2\right)^t\left\|\boldsymbol{x}^{\star}\right\|_2\ll\frac{1}{n}\|\boldsymbol{x}^{\star}\|_2.\tag{103}$$

We are left with (19b). It is self-evident that the iterates from $0 \le t \le T_0$ satisfy (19b) by assumptions. For $t > T_0$, we can use the Cauchy–Schwarz inequality to obtain

$$\max_{1\le j\le m} \left|\boldsymbol{a}_j^\top \left(\boldsymbol{x}^t - \boldsymbol{x}^\star\right)\right| \le \max_{1\le j\le m} \left\|\boldsymbol{a}_j\right\|_2 \left\|\boldsymbol{x}^t - \boldsymbol{x}^\star\right\|_2 \ll \sqrt{n}\cdot\frac{1}{n} \le C_2\sqrt{\log n},$$

where the penultimate relation uses conditions (98) and (103).

A.4 Proof of Lemma 4

First, going through the same derivation as in (54) and (55) will result in

$$\max_{1\le l\le m} \left|\boldsymbol{a}_l^\top \left(\boldsymbol{x}^{t,(l)} - \boldsymbol{x}^\star\right)\right| \le C_4\sqrt{\log n} \tag{104}$$

for some $C_4 < C_2$, which will be helpful for our analysis.

We use the gradient update rules once again to decompose

$$\begin{aligned}
\boldsymbol{x}^{t+1} - \boldsymbol{x}^{t+1,(l)} &= \boldsymbol{x}^t - \eta\nabla f\left(\boldsymbol{x}^t\right) - \left[\boldsymbol{x}^{t,(l)} - \eta\nabla f^{(l)}\left(\boldsymbol{x}^{t,(l)}\right)\right] \\
&= \boldsymbol{x}^t - \eta\nabla f\left(\boldsymbol{x}^t\right) - \left[\boldsymbol{x}^{t,(l)} - \eta\nabla f\left(\boldsymbol{x}^{t,(l)}\right)\right] \\
&\quad -\eta\left[\nabla f\left(\boldsymbol{x}^{t,(l)}\right) - \nabla f^{(l)}\left(\boldsymbol{x}^{t,(l)}\right)\right] \\
&= \underbrace{\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)} - \eta\left[\nabla f\left(\boldsymbol{x}^t\right) - \nabla f\left(\boldsymbol{x}^{t,(l)}\right)\right]}_{:=\boldsymbol{v}_1^{(l)}} \\
&\quad \underbrace{-\eta\frac{1}{m}\left[\left(\boldsymbol{a}_l^\top\boldsymbol{x}^{t,(l)}\right)^2 - \left(\boldsymbol{a}_l^\top\boldsymbol{x}^\star\right)^2\right]\left(\boldsymbol{a}_l^\top\boldsymbol{x}^{t,(l)}\right)\boldsymbol{a}_l}_{:=\boldsymbol{v}_2^{(l)}},
\end{aligned}$$

where the last line comes from the definition of $\nabla f(\cdot)$ and $\nabla f^{(l)}(\cdot)$.

1. We first control the term $\boldsymbol{v}_2^{(l)}$, which is easier to deal with. Specifically,

$$\begin{aligned}
\|\boldsymbol{v}_2^{(l)}\|_2 &\le \eta\frac{\|\boldsymbol{a}_l\|_2}{m}\left|\left(\boldsymbol{a}_l^\top\boldsymbol{x}^{t,(l)}\right)^2 - \left(\boldsymbol{a}_l^\top\boldsymbol{x}^\star\right)^2\right|\left|\boldsymbol{a}_l^\top\boldsymbol{x}^{t,(l)}\right| \\
&\overset{\text{(i)}}{\lesssim} C_4(C_4+5)(C_4+10)\eta\frac{n\log n}{m}\sqrt{\frac{\log n}{n}} \overset{\text{(ii)}}{\le} c\eta\sqrt{\frac{\log n}{n}},
\end{aligned}$$

for any small constant $c > 0$. Here (i) follows since (98) and, in view of (99) and (104),

$$\begin{aligned}
\left|\left(\boldsymbol{a}_l^\top\boldsymbol{x}^{t,(l)}\right)^2 - \left(\boldsymbol{a}_l^\top\boldsymbol{x}^\star\right)^2\right| &\le \left|\boldsymbol{a}_l^\top\left(\boldsymbol{x}^{t,(l)} - \boldsymbol{x}^\star\right)\right|\left(\left|\boldsymbol{a}_l^\top\left(\boldsymbol{x}^{t,(l)} - \boldsymbol{x}^\star\right)\right| + 2\left|\boldsymbol{a}_l^\top\boldsymbol{x}^\star\right|\right) \\
&\le C_4(C_4+10)\log n,
\end{aligned}$$

$$\text{and} \qquad \left|\boldsymbol{a}_l^\top \boldsymbol{x}^{t,(l)}\right| \le \left|\boldsymbol{a}_l^\top \left(\boldsymbol{x}^{t,(l)} - \boldsymbol{x}^\star\right)\right| + \left|\boldsymbol{a}_l^\top \boldsymbol{x}^\star\right| \le (C_4 + 5)\sqrt{\log n}.$$

And (ii) holds as long as $m \gg n \log n$.

2. For the term $\boldsymbol{v}_1^{(l)}$, the fundamental theorem of calculus [70, Chapter XIII, Theorem 4.2] tells us that

$$\boldsymbol{v}_1^{(l)} = \left[\boldsymbol{I}_n - \eta \int_0^1 \nabla^2 f\left(\boldsymbol{x}\left(\tau\right)\right) \mathrm{d}\tau\right]\left(\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)}\right),$$

where we abuse the notation and denote $\boldsymbol{x}\left(\tau\right) = \boldsymbol{x}^{t,(l)} + \tau(\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)})$. By the induction hypotheses (51) and condition (104), one can verify that

$$\begin{aligned} \left\|\boldsymbol{x}\left(\tau\right) - \boldsymbol{x}^\star\right\|_2 \le \tau\left\|\boldsymbol{x}^t - \boldsymbol{x}^\star\right\|_2 + (1-\tau)\left\|\boldsymbol{x}^{t,(l)} - \boldsymbol{x}^\star\right\|_2 \le 2C_1 \qquad \text{and} \\ \max\nolimits_{1\le l\le m}\left|\boldsymbol{a}_l^\top\left(\boldsymbol{x}\left(\tau\right) - \boldsymbol{x}^\star\right)\right| \le \tau \max\nolimits_{1\le l\le m}\left|\boldsymbol{a}_l^\top\left(\boldsymbol{x}^t - \boldsymbol{x}^\star\right)\right| + (1-\tau) \\ \max\nolimits_{1\le l\le m}\left|\boldsymbol{a}_l^\top\left(\boldsymbol{x}^{t,(l)} - \boldsymbol{x}^\star\right)\right| \le C_2\sqrt{\log n} \end{aligned} \tag{105}$$

for all $0 \le \tau \le 1$, as long as $C_4 \le C_2$. The second line follows directly from (104). To see why (105) holds, we note that

$$\left\|\boldsymbol{x}^{t,(l)} - \boldsymbol{x}^\star\right\|_2 \le \left\|\boldsymbol{x}^{t,(l)} - \boldsymbol{x}^t\right\|_2 + \left\|\boldsymbol{x}^t - \boldsymbol{x}^\star\right\|_2 \le C_3\sqrt{\frac{\log n}{n}} + C_1,$$

where the second inequality follows from the induction hypotheses (51b) and (51a). This combined with (51a) gives

$$\left\|\boldsymbol{x}\left(\tau\right) - \boldsymbol{x}^\star\right\|_2 \le \tau C_1 + (1-\tau)\left(C_3\sqrt{\frac{\log n}{n}} + C_1\right) \le 2C_1$$

as long as n is large enough, thus justifying (105). Hence, by Lemma 1, $\nabla^2 f\left(\boldsymbol{x}\left(\tau\right)\right)$ is positive definite and almost well conditioned. By choosing $0 < \eta \le 1/\left[5C_2\left(10 + C_2\right)\log n\right]$, we get

$$\left\|\boldsymbol{v}_1^{(l)}\right\|_2 \le (1 - \eta/2)\left\|\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)}\right\|_2.$$

3. Combine the preceding bounds on $\boldsymbol{v}_1^{(l)}$ and $\boldsymbol{v}_2^{(l)}$ as well as the induction bound (51b) to arrive at

$$\left\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^{t+1,(l)}\right\|_2 \le (1-\eta/2)\left\|\boldsymbol{x}^t - \boldsymbol{x}^{t,(l)}\right\|_2 + c\eta\sqrt{\frac{\log n}{n}} \le C_3\sqrt{\frac{\log n}{n}}. \tag{106}$$

This establishes (53) for the $(t+1)$th iteration.

A.5 Proof of Lemma 5

In view of the assumption (42) that $\left\|\boldsymbol{x}^0-\boldsymbol{x}^\star\right\|_2 \leq \left\|\boldsymbol{x}^0+\boldsymbol{x}^\star\right\|_2$ and the fact that $\boldsymbol{x}^0 = \sqrt{\lambda_1(\boldsymbol{Y})/3}\,\widetilde{\boldsymbol{x}}^0$ for some $\lambda_1(\boldsymbol{Y}) > 0$ (which we will verify below), it is straightforward to see that

$$\left\|\widetilde{\boldsymbol{x}}^0-\boldsymbol{x}^\star\right\|_2 \leq \left\|\widetilde{\boldsymbol{x}}^0+\boldsymbol{x}^\star\right\|_2.$$

One can then invoke the Davis–Kahan sinΘ theorem [124, Corollary 1] to obtain

$$\left\|\widetilde{\boldsymbol{x}}^0-\boldsymbol{x}^\star\right\|_2 \leq 2\sqrt{2}\frac{\|\boldsymbol{Y}-\mathbb{E}[\boldsymbol{Y}]\|}{\lambda_1(\mathbb{E}[\boldsymbol{Y}])-\lambda_2(\mathbb{E}[\boldsymbol{Y}])}.$$

Note that (56)—$\|\boldsymbol{Y}-\mathbb{E}[\boldsymbol{Y}]\| \leq \delta$—is a direct consequence of Lemma 32. Additionally, the fact that $\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{I}+2\boldsymbol{x}^\star\boldsymbol{x}^{\star\top}$ gives $\lambda_1(\mathbb{E}[\boldsymbol{Y}]) = 3$, $\lambda_2(\mathbb{E}[\boldsymbol{Y}]) = 1$, and $\lambda_1(\mathbb{E}[\boldsymbol{Y}]) - \lambda_2(\mathbb{E}[\boldsymbol{Y}]) = 2$. Combining this spectral gap and the inequality $\|\boldsymbol{Y}-\mathbb{E}[\boldsymbol{Y}]\| \leq \delta$, we arrive at

$$\left\|\widetilde{\boldsymbol{x}}^0-\boldsymbol{x}^\star\right\|_2 \leq \sqrt{2}\delta.$$

To connect this bound with $\boldsymbol{x}^0$, we need to take into account the scaling factor $\sqrt{\lambda_1(\boldsymbol{Y})/3}$. To this end, it follows from Weyl's inequality and (56) that

$$|\lambda_1(\boldsymbol{Y})-3| = |\lambda_1(\boldsymbol{Y})-\lambda_1(\mathbb{E}[\boldsymbol{Y}])| \leq \|\boldsymbol{Y}-\mathbb{E}[\boldsymbol{Y}]\| \leq \delta$$

and, as a consequence, $\lambda_1(\boldsymbol{Y}) \geq 3-\delta > 0$ when $\delta \leq 1$. This further implies that

$$\left|\sqrt{\frac{\lambda_1(\boldsymbol{Y})}{3}}-1\right| = \left|\frac{\frac{\lambda_1(\boldsymbol{Y})}{3}-1}{\sqrt{\frac{\lambda_1(\boldsymbol{Y})}{3}}+1}\right| \leq \left|\frac{\lambda_1(\boldsymbol{Y})}{3}-1\right| \leq \frac{1}{3}\delta, \tag{107}$$

where we have used the elementary identity $\sqrt{a}-\sqrt{b} = (a-b)/(\sqrt{a}+\sqrt{b})$. With these bounds in place, we can use the triangle inequality to get

$$\begin{aligned}\left\|\boldsymbol{x}^0-\boldsymbol{x}^\star\right\|_2 &= \left\|\sqrt{\frac{\lambda_1(\boldsymbol{Y})}{3}}\widetilde{\boldsymbol{x}}^0-\boldsymbol{x}^\star\right\|_2 = \left\|\sqrt{\frac{\lambda_1(\boldsymbol{Y})}{3}}\widetilde{\boldsymbol{x}}^0-\widetilde{\boldsymbol{x}}^0+\widetilde{\boldsymbol{x}}^0-\boldsymbol{x}^\star\right\|_2\\ &\leq \left|\sqrt{\frac{\lambda_1(\boldsymbol{Y})}{3}}-1\right| + \left\|\widetilde{\boldsymbol{x}}^0-\boldsymbol{x}^\star\right\|_2\\ &\leq \frac{1}{3}\delta+\sqrt{2}\delta \leq 2\delta.\end{aligned}$$

A.6 Proof of Lemma 6

To begin with, repeating the same argument as in Lemma 5 (which we omit here for conciseness), we see that for any fixed constant $\delta > 0$,

$$\left\| \boldsymbol{Y}^{(l)} - \mathbb{E}\left[\boldsymbol{Y}^{(l)}\right] \right\| \le \delta, \quad \|\boldsymbol{x}^{0,(l)} - \boldsymbol{x}^\star\|_2 \le 2\delta, \quad \left\|\widetilde{\boldsymbol{x}}^{0,(l)} - \boldsymbol{x}^\star\right\|_2 \le \sqrt{2}\delta, \qquad 1 \le l \le m \tag{108}$$

holds with probability at least $1 - O(mn^{-10})$ as long as $m \gg n \log n$. The ℓ_2 bound on $\|\boldsymbol{x}^0 - \boldsymbol{x}^{0,(l)}\|_2$ is derived as follows.

1. We start by controlling $\left\|\widetilde{\boldsymbol{x}}^0 - \widetilde{\boldsymbol{x}}^{0,(l)}\right\|_2$. Combining (57) and (108) yields

$$\left\|\widetilde{\boldsymbol{x}}^0 - \widetilde{\boldsymbol{x}}^{0,(l)}\right\|_2 \le \left\|\widetilde{\boldsymbol{x}}^0 - \boldsymbol{x}^\star\right\|_2 + \left\|\widetilde{\boldsymbol{x}}^{0,(l)} - \boldsymbol{x}^\star\right\|_2 \le 2\sqrt{2}\delta.$$

   For δ sufficiently small, this implies that $\left\|\widetilde{\boldsymbol{x}}^0 - \widetilde{\boldsymbol{x}}^{0,(l)}\right\|_2 \le \left\|\widetilde{\boldsymbol{x}}^0 + \widetilde{\boldsymbol{x}}^{0,(l)}\right\|_2$, and hence, the Davis–Kahan sinΘ theorem [39] gives

$$\left\|\widetilde{\boldsymbol{x}}^0 - \widetilde{\boldsymbol{x}}^{0,(l)}\right\|_2 \le \frac{\left\|\left(\boldsymbol{Y} - \boldsymbol{Y}^{(l)}\right)\widetilde{\boldsymbol{x}}^{0,(l)}\right\|_2}{\lambda_1\left(\boldsymbol{Y}\right) - \lambda_2\left(\boldsymbol{Y}^{(l)}\right)} \le \left\|\left(\boldsymbol{Y} - \boldsymbol{Y}^{(l)}\right)\widetilde{\boldsymbol{x}}^{0,(l)}\right\|_2. \tag{109}$$

   Here, the second inequality uses Weyl's inequality:

$$\begin{aligned}\lambda_1\left(\boldsymbol{Y}\right) - \lambda_2\left(\boldsymbol{Y}^{(l)}\right) &\ge \lambda_1(\mathbb{E}[\boldsymbol{Y}]) - \left\|\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}]\right\| - \lambda_2(\mathbb{E}[\boldsymbol{Y}^{(l)}]) - \left\|\boldsymbol{Y}^{(l)} - \mathbb{E}[\boldsymbol{Y}^{(l)}]\right\| \\ &\ge 3 - \delta - 1 - \delta \ge 1,\end{aligned}$$

   with the proviso that $\delta \le 1/2$.
2. We now connect $\|\boldsymbol{x}^0 - \boldsymbol{x}^{0,(l)}\|_2$ with $\|\widetilde{\boldsymbol{x}}^0 - \widetilde{\boldsymbol{x}}^{0,(l)}\|_2$. Applying the Weyl's inequality and (56) yields

$$|\lambda_1\left(\boldsymbol{Y}\right) - 3| \le \|\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}]\| \le \delta \qquad \Longrightarrow \qquad \lambda_1(\boldsymbol{Y}) \in [3 - \delta, 3 + \delta] \subseteq [2, 4] \tag{110}$$

   and, similarly, $\lambda_1(\boldsymbol{Y}^{(l)})$, $\|\boldsymbol{Y}\|$, $\|\boldsymbol{Y}^{(l)}\| \in [2, 4]$. Invoke Lemma 34 to arrive at

$$\begin{aligned}\frac{1}{\sqrt{3}}\left\|\boldsymbol{x}^0 - \boldsymbol{x}^{0,(l)}\right\|_2 &\le \frac{\left\|\left(\boldsymbol{Y} - \boldsymbol{Y}^{(l)}\right)\widetilde{\boldsymbol{x}}^{0,(l)}\right\|_2}{2\sqrt{2}} + \left(2 + \frac{4}{\sqrt{2}}\right)\left\|\widetilde{\boldsymbol{x}}^0 - \widetilde{\boldsymbol{x}}^{0,(l)}\right\|_2 \\ &\le 6\left\|\left(\boldsymbol{Y} - \boldsymbol{Y}^{(l)}\right)\widetilde{\boldsymbol{x}}^{0,(l)}\right\|_2,\end{aligned} \tag{111}$$

   where the last inequality comes from (109).
3. Everything then boils down to controlling $\left\|\left(\boldsymbol{Y} - \boldsymbol{Y}^{(l)}\right)\widetilde{\boldsymbol{x}}^{0,(l)}\right\|_2$. Toward this, we observe that

$$\begin{aligned}\max_{1\le l\le m}\left\|\left(\boldsymbol{Y} - \boldsymbol{Y}^{(l)}\right)\widetilde{\boldsymbol{x}}^{0,(l)}\right\|_2 &= \max_{1\le l\le m}\frac{1}{m}\left\|\left(\boldsymbol{a}_l^\top \boldsymbol{x}^\star\right)^2 \boldsymbol{a}_l\boldsymbol{a}_l^\top \widetilde{\boldsymbol{x}}^{0,(l)}\right\|_2 \\ &\le \max_{1\le l\le m}\frac{\left(\boldsymbol{a}_l^\top \boldsymbol{x}^\star\right)^2 \left|\boldsymbol{a}_l^\top \widetilde{\boldsymbol{x}}^{0,(l)}\right| \|\boldsymbol{a}_l\|_2}{m}\end{aligned}$$

$$\overset{\text{(i)}}{\lesssim} \frac{\log n \cdot \sqrt{\log n} \cdot \sqrt{n}}{m}$$
$$\asymp \sqrt{\frac{\log n}{n}} \cdot \frac{n \log n}{m}. \tag{112}$$

Inequality (i) makes use of the fact $\max_l \left|\boldsymbol{a}_l^\top \boldsymbol{x}^\star\right| \le 5\sqrt{\log n}$ (cf. (99)), the bound $\max_l \|\boldsymbol{a}_l\|_2 \le 6\sqrt{n}$ (cf. (98)), and $\max_l \left|\boldsymbol{a}_l^\top \widetilde{\boldsymbol{x}}^{0,(l)}\right| \le 5\sqrt{\log n}$ (due to statistical independence and standard Gaussian concentration). As long as $m/(n \log n)$ is sufficiently large, substituting the above bound (112) into (111) leads us to conclude that

$$\max_{1\le l\le m} \left\|\boldsymbol{x}^0 - \boldsymbol{x}^{0,(l)}\right\|_2 \le C_3 \sqrt{\frac{\log n}{n}} \tag{113}$$

for any constant $C_3 > 0$.

B Proofs for Matrix Completion

Before proceeding to the proofs, let us record an immediate consequence of the incoherence property (25):

$$\left\|\boldsymbol{X}^\star\right\|_{2,\infty} \le \sqrt{\frac{\kappa\mu}{n}} \left\|\boldsymbol{X}^\star\right\|_{\mathrm{F}} \le \sqrt{\frac{\kappa\mu r}{n}} \left\|\boldsymbol{X}^\star\right\|. \tag{114}$$

where $\kappa = \sigma_{\max}/\sigma_{\min}$ is the condition number of $\boldsymbol{M}^\star$. This follows since

$$\begin{aligned}\left\|\boldsymbol{X}^\star\right\|_{2,\infty} &= \left\|\boldsymbol{U}^\star \left(\boldsymbol{\Sigma}^\star\right)^{1/2}\right\|_{2,\infty} \le \left\|\boldsymbol{U}^\star\right\|_{2,\infty} \left\|\left(\boldsymbol{\Sigma}^\star\right)^{1/2}\right\| \\ &\le \sqrt{\frac{\mu}{n}} \left\|\boldsymbol{U}^\star\right\|_{\mathrm{F}} \left\|\left(\boldsymbol{\Sigma}^\star\right)^{1/2}\right\| \le \sqrt{\frac{\mu}{n}} \left\|\boldsymbol{U}^\star\right\|_{\mathrm{F}} \sqrt{\kappa\sigma_{\min}} \\ &\le \sqrt{\frac{\kappa\mu}{n}} \left\|\boldsymbol{X}^\star\right\|_{\mathrm{F}} \le \sqrt{\frac{\kappa\mu r}{n}} \left\|\boldsymbol{X}^\star\right\|.\end{aligned}$$

Unless otherwise specified, we use the indicator variable $\delta_{j,k}$ to denote whether the entry in the location (j,k) is included in Ω. Under our model, $\delta_{j,k}$ is a Bernoulli random variable with mean p.

B.1 Proof of Lemma 7

By the expression of the Hessian in (61), one can decompose

$$\operatorname{vec}(\boldsymbol{V})^\top \nabla^2 f_{\text{clean}}(\boldsymbol{X}) \operatorname{vec}(\boldsymbol{V}) = \frac{1}{2p} \left\|\mathcal{P}_\Omega\left(\boldsymbol{V}\boldsymbol{X}^\top + \boldsymbol{X}\boldsymbol{V}^\top\right)\right\|_{\mathrm{F}}^2$$
$$+ \frac{1}{p} \left\langle \mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M}^\star\right), \boldsymbol{V}\boldsymbol{V}^\top \right\rangle$$

$$
\begin{aligned}
&= \underbrace{\frac{1}{2p}\left\|\mathcal{P}_{\Omega}\left(\boldsymbol{V}\boldsymbol{X}^{\top}+\boldsymbol{X}\boldsymbol{V}^{\top}\right)\right\|_{\mathrm{F}}^{2}-\frac{1}{2p}\left\|\mathcal{P}_{\Omega}\left(\boldsymbol{V}\boldsymbol{X}^{\star\top}+\boldsymbol{X}^{\star}\boldsymbol{V}^{\top}\right)\right\|_{\mathrm{F}}^{2}}_{:=\alpha_1} \\
&\quad+\underbrace{\frac{1}{p}\left\langle\mathcal{P}_{\Omega}\left(\boldsymbol{X}\boldsymbol{X}^{\top}-\boldsymbol{M}^{\star}\right),\boldsymbol{V}\boldsymbol{V}^{\top}\right\rangle}_{:=\alpha_2} \\
&\quad+\underbrace{\frac{1}{2p}\left\|\mathcal{P}_{\Omega}\left(\boldsymbol{V}\boldsymbol{X}^{\star\top}+\boldsymbol{X}^{\star}\boldsymbol{V}^{\top}\right)\right\|_{\mathrm{F}}^{2}-\frac{1}{2}\left\|\boldsymbol{V}\boldsymbol{X}^{\star\top}+\boldsymbol{X}^{\star}\boldsymbol{V}^{\top}\right\|_{\mathrm{F}}^{2}}_{:=\alpha_3} \\
&\quad+\underbrace{\frac{1}{2}\left\|\boldsymbol{V}\boldsymbol{X}^{\star\top}+\boldsymbol{X}^{\star}\boldsymbol{V}^{\top}\right\|_{\mathrm{F}}^{2}}_{:=\alpha_4}.
\end{aligned}
$$

The basic idea is to demonstrate that: (1) α_4 is bounded both from above and from below, and (2) the first three terms are sufficiently small in size compared to α_4.

1. We start by controlling α_4. It is immediate to derive the following upper bound

$$
\alpha_4 \le \left\|\boldsymbol{V}\boldsymbol{X}^{\star\top}\right\|_{\mathrm{F}}^{2}+\left\|\boldsymbol{X}^{\star}\boldsymbol{V}^{\top}\right\|_{\mathrm{F}}^{2} \le 2\|\boldsymbol{X}^{\star}\|^{2}\|\boldsymbol{V}\|_{\mathrm{F}}^{2}=2\sigma_{\max}\|\boldsymbol{V}\|_{\mathrm{F}}^{2}.
$$

When it comes to the lower bound, one discovers that

$$
\begin{aligned}
\alpha_4 &= \frac{1}{2}\left\{\left\|\boldsymbol{V}\boldsymbol{X}^{\star\top}\right\|_{\mathrm{F}}^{2}+\left\|\boldsymbol{X}^{\star}\boldsymbol{V}^{\top}\right\|_{\mathrm{F}}^{2}+2\mathrm{Tr}\left(\boldsymbol{X}^{\star\top}\boldsymbol{V}\boldsymbol{X}^{\star\top}\boldsymbol{V}\right)\right\} \\
&\ge \sigma_{\min}\|\boldsymbol{V}\|_{\mathrm{F}}^{2}+\mathrm{Tr}\left[\left(\boldsymbol{Z}+\boldsymbol{X}^{\star}-\boldsymbol{Z}\right)^{\top}\boldsymbol{V}\left(\boldsymbol{Z}+\boldsymbol{X}^{\star}-\boldsymbol{Z}\right)^{\top}\boldsymbol{V}\right] \\
&\ge \sigma_{\min}\|\boldsymbol{V}\|_{\mathrm{F}}^{2}+\mathrm{Tr}\left(\boldsymbol{Z}^{\top}\boldsymbol{V}\boldsymbol{Z}^{\top}\boldsymbol{V}\right)-2\left\|\boldsymbol{Z}-\boldsymbol{X}^{\star}\right\|\|\boldsymbol{Z}\|\|\boldsymbol{V}\|_{\mathrm{F}}^{2}-\left\|\boldsymbol{Z}-\boldsymbol{X}^{\star}\right\|^{2}\|\boldsymbol{V}\|_{\mathrm{F}}^{2} \\
&\ge (\sigma_{\min}-5\delta\sigma_{\max})\|\boldsymbol{V}\|_{\mathrm{F}}^{2}+\mathrm{Tr}\left(\boldsymbol{Z}^{\top}\boldsymbol{V}\boldsymbol{Z}^{\top}\boldsymbol{V}\right), \qquad (115)
\end{aligned}
$$

where the last line comes from the assumptions that

$$
\left\|\boldsymbol{Z}-\boldsymbol{X}^{\star}\right\| \le \delta\left\|\boldsymbol{X}^{\star}\right\| \le \left\|\boldsymbol{X}^{\star}\right\| \qquad \text{and} \qquad \|\boldsymbol{Z}\| \le \left\|\boldsymbol{Z}-\boldsymbol{X}^{\star}\right\|+\left\|\boldsymbol{X}^{\star}\right\| \le 2\left\|\boldsymbol{X}^{\star}\right\|.
$$

With our assumption $\boldsymbol{V}=\boldsymbol{Y}\boldsymbol{H}_{Y}-\boldsymbol{Z}$ in mind, it comes down to controlling

$$
\mathrm{Tr}\left(\boldsymbol{Z}^{\top}\boldsymbol{V}\boldsymbol{Z}^{\top}\boldsymbol{V}\right)=\mathrm{Tr}\left[\boldsymbol{Z}^{\top}\left(\boldsymbol{Y}\boldsymbol{H}_{Y}-\boldsymbol{Z}\right)\boldsymbol{Z}^{\top}\left(\boldsymbol{Y}\boldsymbol{H}_{Y}-\boldsymbol{Z}\right)\right].
$$

From the definition of $\boldsymbol{H}_{Y}$, we see from Lemma 35 that $\boldsymbol{Z}^{\top}\boldsymbol{Y}\boldsymbol{H}_{Y}$ (and hence $\boldsymbol{Z}^{\top}\left(\boldsymbol{Y}\boldsymbol{H}_{Y}-\boldsymbol{Z}\right)$) is a symmetric matrix, which implies that

$$
\mathrm{Tr}\left[\boldsymbol{Z}^{\top}\left(\boldsymbol{Y}\boldsymbol{H}_{Y}-\boldsymbol{Z}\right)\boldsymbol{Z}^{\top}\left(\boldsymbol{Y}\boldsymbol{H}_{Y}-\boldsymbol{Z}\right)\right] \ge 0.
$$

Substitution into (115) gives

$$\alpha_4 \geq (\sigma_{\min} - 5\delta\sigma_{\max}) \|\boldsymbol{V}\|_{\mathrm{F}}^2 \geq \frac{9}{10}\sigma_{\min} \|\boldsymbol{V}\|_{\mathrm{F}}^2,$$

provided that $\kappa\delta \leq 1/50$.

2. For α_1, we consider the following quantity

$$\begin{aligned}
&\left\|\mathcal{P}_\Omega\left(\boldsymbol{V}\boldsymbol{X}^\top + \boldsymbol{X}\boldsymbol{V}^\top\right)\right\|_{\mathrm{F}}^2 \\
&= \left\langle \mathcal{P}_\Omega\left(\boldsymbol{V}\boldsymbol{X}^\top\right), \mathcal{P}_\Omega\left(\boldsymbol{V}\boldsymbol{X}^\top\right)\right\rangle + \left\langle \mathcal{P}_\Omega\left(\boldsymbol{V}\boldsymbol{X}^\top\right), \mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{V}^\top\right)\right\rangle \\
&\quad + \left\langle \mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{V}^\top\right), \mathcal{P}_\Omega\left(\boldsymbol{V}\boldsymbol{X}^\top\right)\right\rangle + \left\langle \mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{V}^\top\right), \mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{V}^\top\right)\right\rangle \\
&= 2\left\langle \mathcal{P}_\Omega\left(\boldsymbol{V}\boldsymbol{X}^\top\right), \mathcal{P}_\Omega\left(\boldsymbol{V}\boldsymbol{X}^\top\right)\right\rangle + 2\left\langle \mathcal{P}_\Omega\left(\boldsymbol{V}\boldsymbol{X}^\top\right), \mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{V}^\top\right)\right\rangle.
\end{aligned}$$

Similar decomposition can be performed on $\left\|\mathcal{P}_\Omega\left(\boldsymbol{V}\boldsymbol{X}^{\star\top} + \boldsymbol{X}^\star\boldsymbol{V}^\top\right)\right\|_{\mathrm{F}}^2$ as well. These identities yield

$$\begin{aligned}
\alpha_1 = &\underbrace{\frac{1}{p}\left[\left\langle \mathcal{P}_\Omega(\boldsymbol{V}\boldsymbol{X}^\top), \mathcal{P}_\Omega\left(\boldsymbol{V}\boldsymbol{X}^\top\right)\right\rangle - \left\langle \mathcal{P}_\Omega\left(\boldsymbol{V}\boldsymbol{X}^{\star\top}\right), \mathcal{P}_\Omega\left(\boldsymbol{V}\boldsymbol{X}^{\star\top}\right)\right\rangle\right]}_{:=\beta_1} \\
&+ \underbrace{\frac{1}{p}\left[\left\langle \mathcal{P}_\Omega(\boldsymbol{V}\boldsymbol{X}^\top), \mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{V}^\top\right)\right\rangle - \left\langle \mathcal{P}_\Omega\left(\boldsymbol{V}\boldsymbol{X}^{\star\top}\right), \mathcal{P}_\Omega\left(\boldsymbol{X}^\star\boldsymbol{V}^\top\right)\right\rangle\right]}_{:=\beta_2}.
\end{aligned}$$

For β_2, one has

$$\begin{aligned}
\beta_2 = &\frac{1}{p}\left\langle \mathcal{P}_\Omega\left(\boldsymbol{V}\left(\boldsymbol{X} - \boldsymbol{X}^\star\right)^\top\right), \mathcal{P}_\Omega\left(\left(\boldsymbol{X} - \boldsymbol{X}^\star\right)\boldsymbol{V}^\top\right)\right\rangle \\
&+ \frac{1}{p}\left\langle \mathcal{P}_\Omega\left(\boldsymbol{V}\left(\boldsymbol{X} - \boldsymbol{X}^\star\right)^\top\right), \mathcal{P}_\Omega\left(\boldsymbol{X}^\star\boldsymbol{V}^\top\right)\right\rangle \\
&+ \frac{1}{p}\left\langle \mathcal{P}_\Omega\left(\boldsymbol{V}\boldsymbol{X}^{\star\top}\right), \mathcal{P}_\Omega\left(\left(\boldsymbol{X} - \boldsymbol{X}^\star\right)\boldsymbol{V}^\top\right)\right\rangle
\end{aligned}$$

which together with the inequality $|\langle \boldsymbol{A}, \boldsymbol{B}\rangle| \leq \|\boldsymbol{A}\|_{\mathrm{F}}\|\boldsymbol{B}\|_{\mathrm{F}}$ gives

$$|\beta_2| \leq \frac{1}{p}\left\|\mathcal{P}_\Omega\left(\boldsymbol{V}\left(\boldsymbol{X} - \boldsymbol{X}^\star\right)^\top\right)\right\|_{\mathrm{F}}^2 + \frac{2}{p}\left\|\mathcal{P}_\Omega\left(\boldsymbol{V}\left(\boldsymbol{X} - \boldsymbol{X}^\star\right)^\top\right)\right\|_{\mathrm{F}}\left\|\mathcal{P}_\Omega\left(\boldsymbol{X}^\star\boldsymbol{V}^\top\right)\right\|_{\mathrm{F}}. \tag{116}$$

This then calls for upper bounds on the following two terms

$$\frac{1}{\sqrt{p}}\left\|\mathcal{P}_\Omega\left(\boldsymbol{V}\left(\boldsymbol{X} - \boldsymbol{X}^\star\right)^\top\right)\right\|_{\mathrm{F}} \quad \text{and} \quad \frac{1}{\sqrt{p}}\left\|\mathcal{P}_\Omega\left(\boldsymbol{X}^\star\boldsymbol{V}^\top\right)\right\|_{\mathrm{F}}.$$

The injectivity of $\mathcal{P}_\Omega$ (cf. [19, Section 4.2] or Lemma 38)—when restricted to the tangent space of $\boldsymbol{M}^\star$—gives: for any fixed constant $\gamma > 0$,

$$\frac{1}{\sqrt{p}} \left\| \mathcal{P}_\Omega \left(\boldsymbol{X}^\star \boldsymbol{V}^\top \right) \right\|_{\mathrm{F}} \le (1+\gamma) \left\| \boldsymbol{X}^\star \boldsymbol{V}^\top \right\|_{\mathrm{F}} \le (1+\gamma) \left\| \boldsymbol{X}^\star \right\| \|\boldsymbol{V}\|_{\mathrm{F}}$$

with probability at least $1 - O\left(n^{-10}\right)$, provided that $n^2 p/(\mu n r \log n)$ is sufficiently large. In addition,

$$\begin{aligned}
&\frac{1}{p} \left\| \mathcal{P}_\Omega \left(\boldsymbol{V} \left(\boldsymbol{X} - \boldsymbol{X}^\star \right)^\top \right) \right\|_{\mathrm{F}}^2 \\
&= \frac{1}{p} \sum_{1 \le j,k \le n} \delta_{j,k} \left[\boldsymbol{V}_{j,\cdot} \left(\boldsymbol{X}_{k,\cdot} - \boldsymbol{X}^\star_{k,\cdot} \right)^\top \right]^2 \\
&= \sum_{1 \le j \le n} \boldsymbol{V}_{j,\cdot} \left[\frac{1}{p} \sum_{1 \le k \le n} \delta_{j,k} \left(\boldsymbol{X}_{k,\cdot} - \boldsymbol{X}^\star_{k,\cdot} \right)^\top \left(\boldsymbol{X}_{k,\cdot} - \boldsymbol{X}^\star_{k,\cdot} \right) \right] \boldsymbol{V}_{j,\cdot}^\top \\
&\le \max_{1 \le j \le n} \left\| \frac{1}{p} \sum_{1 \le k \le n} \delta_{j,k} \left(\boldsymbol{X}_{k,\cdot} - \boldsymbol{X}^\star_{k,\cdot} \right)^\top \left(\boldsymbol{X}_{k,\cdot} - \boldsymbol{X}^\star_{k,\cdot} \right) \right\| \|\boldsymbol{V}\|_{\mathrm{F}}^2 \\
&\le \left\{ \frac{1}{p} \max_{1 \le j \le n} \sum_{1 \le k \le n} \delta_{j,k} \right\} \left\{ \max_{1 \le k \le n} \left\| \boldsymbol{X}_{k,\cdot} - \boldsymbol{X}^\star_{k,\cdot} \right\|_2^2 \right\} \|\boldsymbol{V}\|_{\mathrm{F}}^2 \\
&\le (1+\gamma)\, n \left\| \boldsymbol{X} - \boldsymbol{X}^\star \right\|_{2,\infty}^2 \|\boldsymbol{V}\|_{\mathrm{F}}^2 ,
\end{aligned}$$

with probability exceeding $1 - O\left(n^{-10}\right)$, which holds as long as $np/\log n$ is sufficiently large. Taken collectively, the above bounds yield that for any small constant $\gamma > 0$,

$$\begin{aligned}
|\beta_2| &\le (1+\gamma)\, n \left\| \boldsymbol{X} - \boldsymbol{X}^\star \right\|_{2,\infty}^2 \|\boldsymbol{V}\|_{\mathrm{F}}^2 \\
&\quad + 2\sqrt{(1+\gamma)\, n \left\| \boldsymbol{X} - \boldsymbol{X}^\star \right\|_{2,\infty}^2 \|\boldsymbol{V}\|_{\mathrm{F}}^2 \cdot (1+\gamma)^2 \left\| \boldsymbol{X}^\star \right\|^2 \|\boldsymbol{V}\|_{\mathrm{F}}^2} \\
&\lesssim \left(\epsilon^2 n \left\| \boldsymbol{X}^\star \right\|_{2,\infty}^2 + \epsilon \sqrt{n} \left\| \boldsymbol{X}^\star \right\|_{2,\infty} \left\| \boldsymbol{X}^\star \right\| \right) \|\boldsymbol{V}\|_{\mathrm{F}}^2 ,
\end{aligned}$$

where the last inequality makes use of the assumption $\|\boldsymbol{X} - \boldsymbol{X}^\star\|_{2,\infty} \le \epsilon \|\boldsymbol{X}^\star\|_{2,\infty}$. The same analysis can be repeated to control β_1. Altogether, we obtain

$$\begin{aligned}
|\alpha_1| \le |\beta_1| + |\beta_2| &\lesssim \left(n\epsilon^2 \left\| \boldsymbol{X}^\star \right\|_{2,\infty}^2 + \sqrt{n}\epsilon \left\| \boldsymbol{X}^\star \right\|_{2,\infty} \left\| \boldsymbol{X}^\star \right\| \right) \|\boldsymbol{V}\|_{\mathrm{F}}^2 \\
&\overset{\text{(i)}}{\le} \left(n\epsilon^2 \frac{\kappa \mu r}{n} + \sqrt{n}\epsilon \sqrt{\frac{\kappa \mu r}{n}} \right) \sigma_{\max} \|\boldsymbol{V}\|_{\mathrm{F}}^2 \overset{\text{(ii)}}{\le} \frac{1}{10} \sigma_{\min} \|\boldsymbol{V}\|_{\mathrm{F}}^2 ,
\end{aligned}$$

where (i) utilizes the incoherence condition (114) and (ii) holds with the proviso that $\epsilon \sqrt{\kappa^3 \mu r} \ll 1$.

3. To bound α_2, apply the Cauchy–Schwarz inequality to get

$$|\alpha_2| = \left|\left\langle \boldsymbol{V},\ \frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M}^\star\right)\boldsymbol{V}\right\rangle\right| \le \left\|\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M}^\star\right)\right\| \|\boldsymbol{V}\|_{\mathrm{F}}^2 .$$

In view of Lemma 43, with probability at least $1 - O\left(n^{-10}\right)$,

$$\begin{aligned}\left\|\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M}^\star\right)\right\| &\le 2n\epsilon^2 \left\|\boldsymbol{X}^\star\right\|_{2,\infty}^2 + 4\epsilon\sqrt{n}\log n \left\|\boldsymbol{X}^\star\right\|_{2,\infty} \left\|\boldsymbol{X}^\star\right\| \\ &\le \left(2n\epsilon^2\frac{\kappa\mu r}{n} + 4\epsilon\sqrt{n}\log n\sqrt{\frac{\kappa\mu r}{n}}\right)\sigma_{\max} \le \frac{1}{10}\sigma_{\min}\end{aligned}$$

as soon as $\epsilon\sqrt{\kappa^3\mu r}\log n \ll 1$, where we utilize the incoherence condition (114). This in turn implies that

$$|\alpha_2| \le \frac{1}{10}\sigma_{\min}\|\boldsymbol{V}\|_{\mathrm{F}}^2 .$$

Notably, this bound holds uniformly over all $\boldsymbol{X}$ satisfying the condition in Lemma 7, regardless of the statistical dependence between $\boldsymbol{X}$ and the sampling set Ω.

4. The last term α_3 can also be controlled using the injectivity of $\mathcal{P}_\Omega$ when restricted to the tangent space of $\boldsymbol{M}^\star$. Specifically, it follows from the bounds in [19, Section 4.2] or Lemma 38 that

$$|\alpha_3| \le \gamma\left\|\boldsymbol{V}\boldsymbol{X}^{\star\top} + \boldsymbol{X}^\star\boldsymbol{V}^\top\right\|_{\mathrm{F}}^2 \le 4\gamma\sigma_{\max}\|\boldsymbol{V}\|_{\mathrm{F}}^2 \le \frac{1}{10}\sigma_{\min}\|\boldsymbol{V}\|_{\mathrm{F}}^2$$

for any $\gamma > 0$ such that $\kappa\gamma$ is a small constant, as soon as $n^2 p \gg \kappa^2\mu r n\log n$.

5. Taking all the preceding bounds collectively yields

$$\begin{aligned}\mathrm{vec}\left(\boldsymbol{V}\right)^\top \nabla^2 f_{\mathrm{clean}}\left(\boldsymbol{X}\right)\mathrm{vec}\left(\boldsymbol{V}\right) &\ge \alpha_4 - |\alpha_1| - |\alpha_2| - |\alpha_3| \\ &\ge \left(\frac{9}{10} - \frac{3}{10}\right)\sigma_{\min}\|\boldsymbol{V}\|_{\mathrm{F}}^2 \ge \frac{1}{2}\sigma_{\min}\|\boldsymbol{V}\|_{\mathrm{F}}^2\end{aligned}$$

for all $\boldsymbol{V}$ satisfying our assumptions, and

$$\begin{aligned}\left|\mathrm{vec}\left(\boldsymbol{V}\right)^\top \nabla^2 f_{\mathrm{clean}}\left(\boldsymbol{X}\right)\mathrm{vec}\left(\boldsymbol{V}\right)\right| &\le \alpha_4 + |\alpha_1| + |\alpha_2| + |\alpha_3| \\ &\le \left(2\sigma_{\max} + \frac{3}{10}\sigma_{\min}\right)\|\boldsymbol{V}\|_{\mathrm{F}}^2 \le \frac{5}{2}\sigma_{\max}\|\boldsymbol{V}\|_{\mathrm{F}}^2\end{aligned}$$

for all $\boldsymbol{V}$. Since this upper bound holds uniformly over all $\boldsymbol{V}$, we conclude that

$$\left\|\nabla^2 f_{\mathrm{clean}}\left(\boldsymbol{X}\right)\right\| \le \frac{5}{2}\sigma_{\max}$$

as claimed.

B.2 Proof of Lemma 8

Given that $\widehat{\boldsymbol{H}}^{t+1}$ is chosen to minimize the error in terms of the Frobenius norm (cf. (26)), we have

$$\begin{aligned}
\left\|\boldsymbol{X}^{t+1}\widehat{\boldsymbol{H}}^{t+1}-\boldsymbol{X}^{\star}\right\|_{\mathrm{F}} &\leq \left\|\boldsymbol{X}^{t+1}\widehat{\boldsymbol{H}}^{t}-\boldsymbol{X}^{\star}\right\|_{\mathrm{F}} = \left\|\left[\boldsymbol{X}^{t}-\eta\nabla f\left(\boldsymbol{X}^{t}\right)\right]\widehat{\boldsymbol{H}}^{t}-\boldsymbol{X}^{\star}\right\|_{\mathrm{F}} \\
&\overset{\text{(i)}}{=} \left\|\boldsymbol{X}^{t}\widehat{\boldsymbol{H}}^{t}-\eta\nabla f\big(\boldsymbol{X}^{t}\widehat{\boldsymbol{H}}^{t}\big)-\boldsymbol{X}^{\star}\right\|_{\mathrm{F}} \\
&\overset{\text{(ii)}}{=} \left\|\boldsymbol{X}^{t}\widehat{\boldsymbol{H}}^{t}-\eta\left[\nabla f_{\text{clean}}\big(\boldsymbol{X}^{t}\widehat{\boldsymbol{H}}^{t}\big)-\frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{E}\right)\boldsymbol{X}^{t}\widehat{\boldsymbol{H}}^{t}\right]-\boldsymbol{X}^{\star}\right\|_{\mathrm{F}} \\
&\leq \underbrace{\left\|\boldsymbol{X}^{t}\widehat{\boldsymbol{H}}^{t}-\eta\nabla f_{\text{clean}}\big(\boldsymbol{X}^{t}\widehat{\boldsymbol{H}}^{t}\big)-\big(\boldsymbol{X}^{\star}-\eta\nabla f_{\text{clean}}\big(\boldsymbol{X}^{\star}\big)\big)\right\|_{\mathrm{F}}}_{:=\alpha_1}+\underbrace{\eta\left\|\frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{E}\right)\boldsymbol{X}^{t}\widehat{\boldsymbol{H}}^{t}\right\|_{\mathrm{F}}}_{:=\alpha_2},
\end{aligned} \tag{117}$$

where (i) follows from the identity $\nabla f(\boldsymbol{X}^{t}\boldsymbol{R}) = \nabla f\left(\boldsymbol{X}^{t}\right)\boldsymbol{R}$ for any orthonormal matrix $\boldsymbol{R}\in\mathcal{O}^{r\times r}$, (ii) arises from the definitions of $\nabla f\left(\boldsymbol{X}\right)$ and $\nabla f_{\text{clean}}\left(\boldsymbol{X}\right)$ (see (59) and (60), respectively), and the last inequality (117) utilizes the triangle inequality and the fact that $\nabla f_{\text{clean}}(\boldsymbol{X}^{\star})=\boldsymbol{0}$. It thus suffices to control α_1 and α_2.

1. For the second term α_2 in (117), it is easy to see that with probability at least $1-O\left(n^{-10}\right)$,

$$\alpha_2 \leq \eta\left\|\frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{E}\right)\right\|\left\|\boldsymbol{X}^{t}\widehat{\boldsymbol{H}}^{t}\right\|_{\mathrm{F}} \leq 2\eta\left\|\frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{E}\right)\right\|\left\|\boldsymbol{X}^{\star}\right\|_{\mathrm{F}} \leq 2\eta C\sigma\sqrt{\frac{n}{p}}\|\boldsymbol{X}^{\star}\|_{\mathrm{F}}$$

for some absolute constant $C>0$. Here, the second inequality holds because $\left\|\boldsymbol{X}^{t}\widehat{\boldsymbol{H}}^{t}\right\|_{\mathrm{F}} \leq \left\|\boldsymbol{X}^{t}\widehat{\boldsymbol{H}}^{t}-\boldsymbol{X}^{\star}\right\|_{\mathrm{F}}+\left\|\boldsymbol{X}^{\star}\right\|_{\mathrm{F}} \leq 2\left\|\boldsymbol{X}^{\star}\right\|_{\mathrm{F}}$, following hypothesis (28a) together with our assumptions on the noise and the sample complexity. The last inequality makes use of Lemma 40.

2. For the first term α_1 in (117), the fundamental theorem of calculus [70, Chapter XIII, Theorem 4.2] reveals

$$\begin{aligned}
&\operatorname{vec}\left[\boldsymbol{X}^{t}\widehat{\boldsymbol{H}}^{t}-\eta\nabla f_{\text{clean}}\big(\boldsymbol{X}^{t}\widehat{\boldsymbol{H}}^{t}\big)-\big(\boldsymbol{X}^{\star}-\eta\nabla f_{\text{clean}}\big(\boldsymbol{X}^{\star}\big)\big)\right] \\
&\quad=\operatorname{vec}\left[\boldsymbol{X}^{t}\widehat{\boldsymbol{H}}^{t}-\boldsymbol{X}^{\star}\right]-\eta\cdot\operatorname{vec}\left[\nabla f_{\text{clean}}\big(\boldsymbol{X}^{t}\widehat{\boldsymbol{H}}^{t}\big)-\nabla f_{\text{clean}}\left(\boldsymbol{X}^{\star}\right)\right] \\
&\quad=\Bigg(\boldsymbol{I}_{nr}-\eta\underbrace{\int_{0}^{1}\nabla^{2}f_{\text{clean}}\left(\boldsymbol{X}(\tau)\right)\mathrm{d}\tau}_{:=\boldsymbol{A}}\Bigg)\operatorname{vec}\left(\boldsymbol{X}^{t}\widehat{\boldsymbol{H}}^{t}-\boldsymbol{X}^{\star}\right),
\end{aligned} \tag{118}$$

where we denote $\boldsymbol{X}(\tau):=\boldsymbol{X}^{\star}+\tau(\boldsymbol{X}^{t}\widehat{\boldsymbol{H}}^{t}-\boldsymbol{X}^{\star})$. Taking the squared Euclidean norm of both sides of equality (118) leads to

$$\begin{aligned}
(\alpha_1)^2 &= \operatorname{vec}\big(\boldsymbol{X}^{t}\widehat{\boldsymbol{H}}^{t}-\boldsymbol{X}^{\star}\big)^{\top}\left(\boldsymbol{I}_{nr}-\eta\boldsymbol{A}\right)^{2}\operatorname{vec}\big(\boldsymbol{X}^{t}\widehat{\boldsymbol{H}}^{t}-\boldsymbol{X}^{\star}\big) \\
&= \operatorname{vec}\big(\boldsymbol{X}^{t}\widehat{\boldsymbol{H}}^{t}-\boldsymbol{X}^{\star}\big)^{\top}\left(\boldsymbol{I}_{nr}-2\eta\boldsymbol{A}+\eta^{2}\boldsymbol{A}^{2}\right)\operatorname{vec}\big(\boldsymbol{X}^{t}\widehat{\boldsymbol{H}}^{t}-\boldsymbol{X}^{\star}\big)
\end{aligned}$$

$$\leq \left\| \boldsymbol{X}^t \widehat{\boldsymbol{H}}^t - \boldsymbol{X}^\star \right\|_{\mathrm{F}}^2 + \eta^2 \|\boldsymbol{A}\|^2 \left\| \boldsymbol{X}^t \widehat{\boldsymbol{H}}^t - \boldsymbol{X}^\star \right\|_{\mathrm{F}}^2 - 2\eta \operatorname{vec}\left(\boldsymbol{X}^t \widehat{\boldsymbol{H}}^t - \boldsymbol{X}^\star\right)^\top \boldsymbol{A} \operatorname{vec}\left(\boldsymbol{X}^t \widehat{\boldsymbol{H}}^t - \boldsymbol{X}^\star\right), \tag{119}$$

where in (119) we have used the fact that

$$\begin{aligned} \operatorname{vec}\left(\boldsymbol{X}^t \widehat{\boldsymbol{H}}^t - \boldsymbol{X}^\star\right)^\top \boldsymbol{A}^2 \operatorname{vec}\left(\boldsymbol{X}^t \widehat{\boldsymbol{H}}^t - \boldsymbol{X}^\star\right) &\leq \|\boldsymbol{A}\|^2 \left\| \operatorname{vec}\left(\boldsymbol{X}^t \widehat{\boldsymbol{H}}^t - \boldsymbol{X}^\star\right) \right\|_2^2 \\ &= \|\boldsymbol{A}\|^2 \left\| \boldsymbol{X}^t \widehat{\boldsymbol{H}}^t - \boldsymbol{X}^\star \right\|_{\mathrm{F}}^2. \end{aligned}$$

Based on condition (28b), it is easily seen that $\forall \tau \in [0, 1]$,

$$\left\| \boldsymbol{X}(\tau) - \boldsymbol{X}^\star \right\|_{2,\infty} \leq \left(C_5 \mu r \sqrt{\frac{\log n}{np}} + \frac{C_8}{\sigma_{\min}} \sigma \sqrt{\frac{n \log n}{p}} \right) \left\| \boldsymbol{X}^\star \right\|_{2,\infty}.$$

Taking $\boldsymbol{X} = \boldsymbol{X}(\tau)$, $\boldsymbol{Y} = \boldsymbol{X}^t$, and $\boldsymbol{Z} = \boldsymbol{X}^\star$ in Lemma 7, one can easily verify the assumptions therein given our sample size condition $n^2 p \gg \kappa^3 \mu^3 r^3 n \log^3 n$ and the noise condition (27). As a result,

$$\operatorname{vec}\left(\boldsymbol{X}^t \widehat{\boldsymbol{H}}^t - \boldsymbol{X}^\star\right)^\top \boldsymbol{A} \operatorname{vec}\left(\boldsymbol{X}^t \widehat{\boldsymbol{H}}^t - \boldsymbol{X}^\star\right) \geq \frac{\sigma_{\min}}{2} \left\| \boldsymbol{X}^t \widehat{\boldsymbol{H}}^t - \boldsymbol{X}^\star \right\|_{\mathrm{F}}^2$$

$$\text{and} \qquad \|\boldsymbol{A}\| \leq \frac{5}{2} \sigma_{\max}.$$

Substituting these two inequalities into (119) yields

$$\begin{aligned} (\alpha_1)^2 &\leq \left(1 + \frac{25}{4}\eta^2 \sigma_{\max}^2 - \sigma_{\min}\eta\right) \left\| \boldsymbol{X}^t \widehat{\boldsymbol{H}}^t - \boldsymbol{X}^\star \right\|_{\mathrm{F}}^2 \\ &\leq \left(1 - \frac{\sigma_{\min}}{2}\eta\right) \left\| \boldsymbol{X}^t \widehat{\boldsymbol{H}}^t - \boldsymbol{X}^\star \right\|_{\mathrm{F}}^2 \end{aligned}$$

as long as $0 < \eta \leq (2\sigma_{\min})/(25\sigma_{\max}^2)$, which further implies that

$$\alpha_1 \leq \left(1 - \frac{\sigma_{\min}}{4}\eta\right) \left\| \boldsymbol{X}^t \widehat{\boldsymbol{H}}^t - \boldsymbol{X}^\star \right\|_{\mathrm{F}}.$$

3. Combining the preceding bounds on both α_1 and α_2 and making use of hypothesis (28a), we have

$$\begin{aligned} \left\| \boldsymbol{X}^{t+1} \widehat{\boldsymbol{H}}^{t+1} - \boldsymbol{X}^\star \right\|_{\mathrm{F}} &\leq \left(1 - \tfrac{\sigma_{\min}}{4}\eta\right) \left\| \boldsymbol{X}^t \widehat{\boldsymbol{H}}^t - \boldsymbol{X}^\star \right\|_{\mathrm{F}} + 2\eta C \sigma \sqrt{\frac{n}{p}} \left\| \boldsymbol{X}^\star \right\|_{\mathrm{F}} \\ &\leq \left(1 - \tfrac{\sigma_{\min}}{4}\eta\right) \left(C_4 \rho^t \mu r \frac{1}{\sqrt{np}} \left\| \boldsymbol{X}^\star \right\|_{\mathrm{F}} + C_1 \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \left\| \boldsymbol{X}^\star \right\|_{\mathrm{F}} \right) + 2\eta C \sigma \sqrt{\frac{n}{p}} \left\| \boldsymbol{X}^\star \right\|_{\mathrm{F}} \\ &\leq \left(1 - \tfrac{\sigma_{\min}}{4}\eta\right) C_4 \rho^t \mu r \frac{1}{\sqrt{np}} \left\| \boldsymbol{X}^\star \right\|_{\mathrm{F}} + \left[\left(1 - \tfrac{\sigma_{\min}}{4}\eta\right) \frac{C_1}{\sigma_{\min}} + 2\eta C \right] \sigma \sqrt{\frac{n}{p}} \left\| \boldsymbol{X}^\star \right\|_{\mathrm{F}} \\ &\leq C_4 \rho^{t+1} \mu r \frac{1}{\sqrt{np}} \left\| \boldsymbol{X}^\star \right\|_{\mathrm{F}} + C_1 \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \left\| \boldsymbol{X}^\star \right\|_{\mathrm{F}} \end{aligned}$$

as long as $0 < \eta \leq (2\sigma_{\min})/(25\sigma_{\max}^2)$, $1 - (\sigma_{\min}/4) \cdot \eta \leq \rho < 1$, and C_1 is sufficiently large. This completes the proof of the contraction with respect to the Frobenius norm.

B.3 Proof of Lemma 9

To facilitate analysis, we construct an auxiliary matrix defined as follows

$$\widetilde{\boldsymbol{X}}^{t+1} := \boldsymbol{X}^t \widehat{\boldsymbol{H}}^t - \eta \frac{1}{p} \mathcal{P}_\Omega \Big[\boldsymbol{X}^t \boldsymbol{X}^{t\top} - \big(\boldsymbol{M}^\star + \boldsymbol{E}\big) \Big] \boldsymbol{X}^\star. \tag{120}$$

With this auxiliary matrix in place, we invoke the triangle inequality to bound

$$\big\| \boldsymbol{X}^{t+1} \widehat{\boldsymbol{H}}^{t+1} - \boldsymbol{X}^\star \big\| \leq \underbrace{\big\| \boldsymbol{X}^{t+1} \widehat{\boldsymbol{H}}^{t+1} - \widetilde{\boldsymbol{X}}^{t+1} \big\|}_{:=\alpha_1} + \underbrace{\big\| \widetilde{\boldsymbol{X}}^{t+1} - \boldsymbol{X}^\star \big\|}_{:=\alpha_2}. \tag{121}$$

1. We start with the second term α_2 and show that the auxiliary matrix $\widetilde{\boldsymbol{X}}^{t+1}$ is also not far from the truth. The definition of $\widetilde{\boldsymbol{X}}^{t+1}$ allows one to express

$$\begin{aligned}
\alpha_2 &= \left\| \boldsymbol{X}^t \widehat{\boldsymbol{H}}^t - \eta \frac{1}{p} \mathcal{P}_\Omega \Big[\boldsymbol{X}^t \boldsymbol{X}^{t\top} - \big(\boldsymbol{M}^\star + \boldsymbol{E}\big) \Big] \boldsymbol{X}^\star - \boldsymbol{X}^\star \right\| \\
&\leq \eta \left\| \frac{1}{p} \mathcal{P}_\Omega (\boldsymbol{E}) \right\| \big\| \boldsymbol{X}^\star \big\| + \left\| \boldsymbol{X}^t \widehat{\boldsymbol{H}}^t - \eta \frac{1}{p} \mathcal{P}_\Omega \left(\boldsymbol{X}^t \boldsymbol{X}^{t\top} - \boldsymbol{X}^\star \boldsymbol{X}^{\star\top} \right) \boldsymbol{X}^\star - \boldsymbol{X}^\star \right\| \qquad (122) \\
&\leq \eta \left\| \frac{1}{p} \mathcal{P}_\Omega (\boldsymbol{E}) \right\| \big\| \boldsymbol{X}^\star \big\| + \underbrace{\left\| \boldsymbol{X}^t \widehat{\boldsymbol{H}}^t - \eta \left(\boldsymbol{X}^t \boldsymbol{X}^{t\top} - \boldsymbol{X}^\star \boldsymbol{X}^{\star\top} \right) \boldsymbol{X}^\star - \boldsymbol{X}^\star \right\|}_{:=\beta_1} \\
&\quad + \underbrace{\eta \left\| \frac{1}{p} \mathcal{P}_\Omega \left(\boldsymbol{X}^t \boldsymbol{X}^{t\top} - \boldsymbol{X}^\star \boldsymbol{X}^{\star\top} \right) \boldsymbol{X}^\star - \left(\boldsymbol{X}^t \boldsymbol{X}^{t\top} - \boldsymbol{X}^\star \boldsymbol{X}^{\star\top} \right) \boldsymbol{X}^\star \right\|}_{:=\beta_2}, \qquad (123)
\end{aligned}$$

where we have used the triangle inequality to separate the population-level component (i.e., β_1), the perturbation (i.e., β_2), and the noise component. In what follows, we will denote

$$\boldsymbol{\Delta}^t := \boldsymbol{X}^t \widehat{\boldsymbol{H}}^t - \boldsymbol{X}^\star$$

which, by Lemma 35, satisfies the following symmetry property

$$\widehat{\boldsymbol{H}}^{t\top} \boldsymbol{X}^{t\top} \boldsymbol{X}^\star = \boldsymbol{X}^{\star\top} \boldsymbol{X}^t \widehat{\boldsymbol{H}}^t \quad \Longrightarrow \quad \boldsymbol{\Delta}^{t\top} \boldsymbol{X}^\star = \boldsymbol{X}^{\star\top} \boldsymbol{\Delta}^t. \tag{124}$$

(a) The population-level component β_1 is easier to control. Specifically, we first simplify its expression as

$$\begin{aligned}\beta_1 &= \left\| \boldsymbol{\Delta}^t - \eta \left(\boldsymbol{\Delta}^t \boldsymbol{\Delta}^{t\top} + \boldsymbol{\Delta}^t \boldsymbol{X}^{\star\top} + \boldsymbol{X}^\star \boldsymbol{\Delta}^{t\top} \right) \boldsymbol{X}^\star \right\| \\ &\le \underbrace{\left\| \boldsymbol{\Delta}^t - \eta \left(\boldsymbol{\Delta}^t \boldsymbol{X}^{\star\top} + \boldsymbol{X}^\star \boldsymbol{\Delta}^{t\top} \right) \boldsymbol{X}^\star \right\|}_{:=\gamma_1} + \underbrace{\eta \left\| \boldsymbol{\Delta}^t \boldsymbol{\Delta}^{t\top} \boldsymbol{X}^\star \right\|}_{:=\gamma_2}.\end{aligned}$$

The leading term γ_1 can be upper bounded by

$$\begin{aligned}\gamma_1 &= \left\| \boldsymbol{\Delta}^t - \eta \boldsymbol{\Delta}^t \boldsymbol{\Sigma}^\star - \eta \boldsymbol{X}^\star \boldsymbol{\Delta}^{t\top} \boldsymbol{X}^\star \right\| = \left\| \boldsymbol{\Delta}^t - \eta \boldsymbol{\Delta}^t \boldsymbol{\Sigma}^\star - \eta \boldsymbol{X}^\star \boldsymbol{X}^{\star\top} \boldsymbol{\Delta}^t \right\| \\ &= \left\| \frac{1}{2} \boldsymbol{\Delta}^t \left(\boldsymbol{I}_r - 2\eta \boldsymbol{\Sigma}^\star \right) + \frac{1}{2} \left(\boldsymbol{I}_n - 2\eta \boldsymbol{M}^\star \right) \boldsymbol{\Delta}^t \right\| \\ &\le \frac{1}{2} \left(\left\| \boldsymbol{I}_r - 2\eta \boldsymbol{\Sigma}^\star \right\| + \left\| \boldsymbol{I}_n - 2\eta \boldsymbol{M}^\star \right\| \right) \left\| \boldsymbol{\Delta}^t \right\|\end{aligned}$$

where the second identity follows from the symmetry property (124). By choosing $\eta \le 1/(2\sigma_{\max})$, one has $\boldsymbol{0} \preceq \boldsymbol{I}_r - 2\eta \boldsymbol{\Sigma}^\star \preceq (1 - 2\eta\sigma_{\min}) \boldsymbol{I}_r$ and $\boldsymbol{0} \preceq \boldsymbol{I}_n - 2\eta \boldsymbol{M}^\star \preceq \boldsymbol{I}_n$, and further one can ensure

$$\gamma_1 \le \frac{1}{2} \left[(1 - 2\eta\sigma_{\min}) + 1 \right] \left\| \boldsymbol{\Delta}^t \right\| = (1 - \eta\sigma_{\min}) \left\| \boldsymbol{\Delta}^t \right\|. \tag{125}$$

Next, regarding the higher-order term γ_2, we can easily obtain

$$\gamma_2 \le \eta \left\| \boldsymbol{\Delta}^t \right\|^2 \left\| \boldsymbol{X}^\star \right\|. \tag{126}$$

Bounds (125) and (126) taken collectively give

$$\beta_1 \le (1 - \eta\sigma_{\min}) \left\| \boldsymbol{\Delta}^t \right\| + \eta \left\| \boldsymbol{\Delta}^t \right\|^2 \left\| \boldsymbol{X}^\star \right\|. \tag{127}$$

(b) We now turn to the perturbation part β_2 by showing that

$$\begin{aligned}\frac{1}{\eta}\beta_2 &= \left\| \frac{1}{p} \mathcal{P}_\Omega \left(\boldsymbol{\Delta}^t \boldsymbol{\Delta}^{t\top} + \boldsymbol{\Delta}^t \boldsymbol{X}^{\star\top} + \boldsymbol{X}^\star \boldsymbol{\Delta}^{t\top} \right) \boldsymbol{X}^\star \right. \\ &\qquad \left. - \left[\boldsymbol{\Delta}^t \boldsymbol{\Delta}^{t\top} + \boldsymbol{\Delta}^t \boldsymbol{X}^{\star\top} + \boldsymbol{X}^\star \boldsymbol{\Delta}^{t\top} \right] \boldsymbol{X}^\star \right\| \\ &\le \underbrace{\left\| \frac{1}{p} \mathcal{P}_\Omega \left(\boldsymbol{\Delta}^t \boldsymbol{X}^{\star\top} \right) \boldsymbol{X}^\star - \left(\boldsymbol{\Delta}^t \boldsymbol{X}^{\star\top} \right) \boldsymbol{X}^\star \right\|_{\mathrm{F}}}_{:=\theta_1}\end{aligned}$$

$$+\underbrace{\left\|\frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{X}^{\star}\boldsymbol{\Delta}^{t\top}\right)\boldsymbol{X}^{\star}-\left(\boldsymbol{X}^{\star}\boldsymbol{\Delta}^{t\top}\right)\boldsymbol{X}^{\star}\right\|_{\mathrm{F}}}_{:=\theta_2}$$

$$+\underbrace{\left\|\frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{\Delta}^{t}\boldsymbol{\Delta}^{t\top}\right)\boldsymbol{X}^{\star}-\left(\boldsymbol{\Delta}^{t}\boldsymbol{\Delta}^{t\top}\right)\boldsymbol{X}^{\star}\right\|_{\mathrm{F}}}_{:=\theta_3}, \tag{128}$$

where the last inequality holds due to the triangle inequality as well as the fact that $\|\boldsymbol{A}\| \le \|\boldsymbol{A}\|_{\mathrm{F}}$. In the sequel, we shall bound the three terms separately.

- For the first term θ_1 in (128), the lth row of $\frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{\Delta}^{t}\boldsymbol{X}^{\star\top}\right)\boldsymbol{X}^{\star}-\left(\boldsymbol{\Delta}^{t}\boldsymbol{X}^{\star\top}\right)\boldsymbol{X}^{\star}$ is given by

$$\frac{1}{p}\sum_{j=1}^{n}\left(\delta_{l,j}-p\right)\boldsymbol{\Delta}_{l,\cdot}^{t}\boldsymbol{X}_{j,\cdot}^{\star\top}\boldsymbol{X}_{j,\cdot}^{\star}=\boldsymbol{\Delta}_{l,\cdot}^{t}\left[\frac{1}{p}\sum_{j=1}^{n}\left(\delta_{l,j}-p\right)\boldsymbol{X}_{j,\cdot}^{\star\top}\boldsymbol{X}_{j,\cdot}^{\star}\right]$$

where, as usual, $\delta_{l,j}=\mathbb{1}_{\{(l,j)\in\Omega\}}$. Lemma 41 together with the union bound reveals that

$$\begin{aligned}&\left\|\frac{1}{p}\sum_{j=1}^{n}\left(\delta_{l,j}-p\right)\boldsymbol{X}_{j,\cdot}^{\star\top}\boldsymbol{X}_{j,\cdot}^{\star}\right\|\\&\lesssim\frac{1}{p}\left(\sqrt{p\left\|\boldsymbol{X}^{\star}\right\|_{2,\infty}^{2}\left\|\boldsymbol{X}^{\star}\right\|^{2}\log n}+\left\|\boldsymbol{X}^{\star}\right\|_{2,\infty}^{2}\log n\right)\\&\asymp\sqrt{\frac{\|\boldsymbol{X}^{\star}\|_{2,\infty}^{2}\sigma_{\max}\log n}{p}}+\frac{\|\boldsymbol{X}^{\star}\|_{2,\infty}^{2}\log n}{p}\end{aligned}$$

for all $1\le l\le n$ with high probability. This gives

$$\begin{aligned}&\left\|\boldsymbol{\Delta}_{l,\cdot}^{t}\left[\frac{1}{p}\sum_{j=1}^{n}\left(\delta_{l,j}-p\right)\boldsymbol{X}_{j,\cdot}^{\star\top}\boldsymbol{X}_{j,\cdot}^{\star}\right]\right\|_{2}\\&\le\left\|\boldsymbol{\Delta}_{l,\cdot}^{t}\right\|_{2}\left\|\frac{1}{p}\sum_{j}\left(\delta_{l,j}-p\right)\boldsymbol{X}_{j,\cdot}^{\star\top}\boldsymbol{X}_{j,\cdot}^{\star}\right\|\\&\lesssim\left\|\boldsymbol{\Delta}_{l,\cdot}^{t}\right\|_{2}\left\{\sqrt{\frac{\|\boldsymbol{X}^{\star}\|_{2,\infty}^{2}\sigma_{\max}\log n}{p}}+\frac{\|\boldsymbol{X}^{\star}\|_{2,\infty}^{2}\log n}{p}\right\},\end{aligned}$$

which further reveals that

$$\begin{aligned}
\theta_1 &= \sqrt{\sum_{l=1}^{n} \left\| \frac{1}{p} \sum_j \left(\delta_{l,j} - p\right) \boldsymbol{\Delta}_{l,\cdot}^t \boldsymbol{X}_{j,\cdot}^{\star\top} \boldsymbol{X}_{j,\cdot}^{\star} \right\|_2^2} \\
&\lesssim \left\|\boldsymbol{\Delta}^t\right\|_{\mathrm{F}} \left\{ \sqrt{\frac{\|\boldsymbol{X}^\star\|_{2,\infty}^2 \sigma_{\max} \log n}{p}} + \frac{\|\boldsymbol{X}^\star\|_{2,\infty}^2 \log n}{p} \right\} \\
&\overset{\text{(i)}}{\lesssim} \left\|\boldsymbol{\Delta}^t\right\| \left\{ \sqrt{\frac{\|\boldsymbol{X}^\star\|_{2,\infty}^2 r \sigma_{\max} \log n}{p}} + \frac{\sqrt{r}\|\boldsymbol{X}^\star\|_{2,\infty}^2 \log n}{p} \right\} \\
&\overset{\text{(ii)}}{\lesssim} \left\|\boldsymbol{\Delta}^t\right\| \left\{ \sqrt{\frac{\kappa\mu r^2 \log n}{np}} + \frac{\kappa\mu r^{3/2} \log n}{np} \right\} \sigma_{\max} \\
&\overset{\text{(iii)}}{\leq} \gamma \sigma_{\min} \left\|\boldsymbol{\Delta}^t\right\|,
\end{aligned}$$

for arbitrarily small $\gamma > 0$. Here, (i) follows from $\left\|\boldsymbol{\Delta}^t\right\|_{\mathrm{F}} \leq \sqrt{r}\left\|\boldsymbol{\Delta}^t\right\|$, (ii) holds owing to the incoherence condition (114), and (iii) follows as long as $n^2 p \gg \kappa^3 \mu r^2 n \log n$.

- For the second term θ_2 in (128), denote

$$\boldsymbol{A} = \mathcal{P}_\Omega\left(\boldsymbol{X}^\star \boldsymbol{\Delta}^{t\top}\right) \boldsymbol{X}^\star - p\left(\boldsymbol{X}^\star \boldsymbol{\Delta}^{t\top}\right) \boldsymbol{X}^\star,$$

whose lth row is given by

$$\boldsymbol{A}_{l,\cdot} = \boldsymbol{X}_{l,\cdot}^\star \sum_{j=1}^{n} \left(\delta_{l,j} - p\right) \boldsymbol{\Delta}_{j,\cdot}^{t\top} \boldsymbol{X}_{j,\cdot}^\star. \tag{129}$$

Recalling the induction hypotheses (28b) and (28c), we define

$$\left\|\boldsymbol{\Delta}^t\right\|_{2,\infty} \leq C_5 \rho^t \mu r \sqrt{\frac{\log n}{np}} \left\|\boldsymbol{X}^\star\right\|_{2,\infty} + C_8 \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n \log n}{p}} \left\|\boldsymbol{X}^\star\right\|_{2,\infty} := \xi \tag{130}$$

$$\left\|\boldsymbol{\Delta}^t\right\| \leq C_9 \rho^t \mu r \frac{1}{\sqrt{np}} \left\|\boldsymbol{X}^\star\right\| + C_{10} \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \left\|\boldsymbol{X}^\star\right\| := \psi. \tag{131}$$

With these two definitions in place, we now introduce a "truncation level"

$$\omega := 2p\xi\sigma_{\max} \tag{132}$$

that allows us to bound θ_2 in terms of the following two terms

$$\theta_2 = \frac{1}{p}\sqrt{\sum_{l=1}^{n}\left\|\boldsymbol{A}_{l,\cdot}\right\|_2^2} \le \frac{1}{p}\underbrace{\sqrt{\sum_{l=1}^{n}\left\|\boldsymbol{A}_{l,\cdot}\right\|_2^2 \mathbb{1}_{\{\|\boldsymbol{A}_{l,\cdot}\|_2\le\omega\}}}}_{:=\phi_1} + \frac{1}{p}\underbrace{\sqrt{\sum_{l=1}^{n}\left\|\boldsymbol{A}_{l,\cdot}\right\|_2^2 \mathbb{1}_{\{\|\boldsymbol{A}_{l,\cdot}\|_2\ge\omega\}}}}_{:=\phi_2}.$$

We will apply different strategies when upper bounding the terms ϕ_1 and ϕ_2, with their bounds given in the following two lemmas under the induction hypotheses (28b) and (28c).

Lemma 22 *Under the conditions in Lemma 9, there exist some constants $c, C > 0$ such that with probability exceeding $1 - c\exp(-Cnr\log n)$,*

$$\phi_1 \lesssim \xi\sqrt{p\sigma_{\max}\|\boldsymbol{X}^\star\|_{2,\infty}^2 nr\log^2 n} \tag{133}$$

holds simultaneously for all $\boldsymbol{\Delta}^t$ obeying (130) and (131). Here, ξ is defined in (130).

Lemma 23 *Under the conditions in Lemma 9, with probability at least $1 - O\left(n^{-10}\right)$,*

$$\phi_2 \lesssim \xi\sqrt{\kappa\mu r^2 p\log^2 n}\left\|\boldsymbol{X}^\star\right\|^2 \tag{134}$$

holds simultaneously for all $\boldsymbol{\Delta}^t$ obeying (130) and (131). Here, ξ is defined in (130).

Bounds (133) and (134) together with the incoherence condition (114) yield

$$\begin{aligned}\theta_2 &\lesssim \frac{1}{p}\xi\sqrt{p\sigma_{\max}\|\boldsymbol{X}^\star\|_{2,\infty}^2 nr\log^2 n} + \frac{1}{p}\xi\sqrt{\kappa\mu r^2 p\log^2 n}\left\|\boldsymbol{X}^\star\right\|^2 \\ &\lesssim \sqrt{\frac{\kappa\mu r^2\log^2 n}{p}}\xi\sigma_{\max}.\end{aligned}$$

- Next, we assert that the third term θ_3 in (128) has the same upper bound as θ_2. The proof follows by repeating the same argument used in bounding θ_2 and is hence omitted.

Take the previous three bounds on θ_1, θ_2, and θ_3 together to arrive at

$$\beta_2 \le \eta\left(|\theta_1| + |\theta_2| + |\theta_3|\right) \le \eta\gamma\sigma_{\min}\left\|\boldsymbol{\Delta}^t\right\| + \widetilde{C}\eta\sqrt{\frac{\kappa\mu r^2\log^2 n}{p}}\xi\sigma_{\max}$$

for some constant $\widetilde{C} > 0$.

(c) Substituting the preceding bounds on β_1 and β_2 into (123), we reach

$$\begin{aligned}
\alpha_2 &\overset{\text{(i)}}{\leq} \left(1-\eta\sigma_{\min}+\eta\gamma\sigma_{\min}+\eta\left\|\boldsymbol{\Delta}^t\right\|\left\|\boldsymbol{X}^\star\right\|\right)\left\|\boldsymbol{\Delta}^t\right\|+\eta\left\|\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{E}\right)\right\|\left\|\boldsymbol{X}^\star\right\| \\
&\quad+\widetilde{C}\eta\sqrt{\frac{\kappa\mu r^2\log^2 n}{p}}\sigma_{\max}\left(C_5\rho^t\mu r\sqrt{\frac{\log n}{np}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty}+C_8\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty}\right) \\
&\overset{\text{(ii)}}{\leq} \left(1-\frac{\sigma_{\min}}{2}\eta\right)\left\|\boldsymbol{\Delta}^t\right\|+\eta\left\|\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{E}\right)\right\|\left\|\boldsymbol{X}^\star\right\| \\
&\quad+\widetilde{C}\eta\sqrt{\frac{\kappa\mu r^2\log^2 n}{p}}\sigma_{\max}\left(C_5\rho^t\mu r\sqrt{\frac{\log n}{np}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty}+C_8\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty}\right) \\
&\overset{\text{(iii)}}{\leq} \left(1-\frac{\sigma_{\min}}{2}\eta\right)\left\|\boldsymbol{\Delta}^t\right\|+C\eta\sigma\sqrt{\frac{n}{p}}\left\|\boldsymbol{X}^\star\right\| \\
&\quad+\widetilde{C}\eta\sqrt{\frac{\kappa^2\mu^2 r^3\log^3 n}{np}}\sigma_{\max}\left(C_5\rho^t\mu r\sqrt{\frac{1}{np}}+C_8\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\right)\left\|\boldsymbol{X}^\star\right\| \qquad (135)
\end{aligned}$$

for some constant $C>0$. Here, (i) uses the definition of ξ (cf. (130)), (ii) holds if γ is small enough and $\left\|\boldsymbol{\Delta}^t\right\|\left\|\boldsymbol{X}^\star\right\|\ll\sigma_{\min}$, and (iii) follows from Lemma 40 as well as the incoherence condition (114). An immediate consequence of (135) is that under the sample size condition and the noise condition of this lemma, one has

$$\left\|\widetilde{\boldsymbol{X}}^{t+1}-\boldsymbol{X}^\star\right\|\left\|\boldsymbol{X}^\star\right\|\leq\sigma_{\min}/2 \qquad (136)$$

if $0<\eta\leq 1/\sigma_{\max}$.

2. We then move on to the first term α_1 in (121), which can be rewritten as

$$\alpha_1=\left\|\boldsymbol{X}^{t+1}\widehat{\boldsymbol{H}}^t\boldsymbol{R}_1-\widetilde{\boldsymbol{X}}^{t+1}\right\|,$$

with

$$\boldsymbol{R}_1=\left(\widehat{\boldsymbol{H}}^t\right)^{-1}\widehat{\boldsymbol{H}}^{t+1}:=\arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}}\left\|\boldsymbol{X}^{t+1}\widehat{\boldsymbol{H}}^t\boldsymbol{R}-\boldsymbol{X}^\star\right\|_{\mathrm{F}}. \qquad (137)$$

(a) First, we claim that $\widetilde{\boldsymbol{X}}^{t+1}$ satisfies

$$\boldsymbol{I}_r=\arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}}\left\|\widetilde{\boldsymbol{X}}^{t+1}\boldsymbol{R}-\boldsymbol{X}^\star\right\|_{\mathrm{F}}, \qquad (138)$$

meaning that $\widetilde{\boldsymbol{X}}^{t+1}$ is already rotated to the direction that is most "aligned" with $\boldsymbol{X}^\star$. This important property eases the analysis. In fact, in view of Lemma 35, (138) follows if one can show that $\boldsymbol{X}^{\star\top}\widetilde{\boldsymbol{X}}^{t+1}$ is symmetric and positive semidefinite. First of all, it follows from Lemma 35 that $\boldsymbol{X}^{\star\top}\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t$ is symmetric and, hence, by definition,

$$\boldsymbol{X}^{\star\top}\widetilde{\boldsymbol{X}}^{t+1}=\boldsymbol{X}^{\star\top}\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t-\frac{\eta}{p}\boldsymbol{X}^{\star\top}\mathcal{P}_\Omega\left[\boldsymbol{X}^t\boldsymbol{X}^{t\top}-\left(\boldsymbol{M}^\star+\boldsymbol{E}\right)\right]\boldsymbol{X}^\star$$

is also symmetric. Additionally,

$$\left\|\boldsymbol{X}^{\star\top}\widetilde{\boldsymbol{X}}^{t+1}-\boldsymbol{M}^{\star}\right\| \leq \left\|\widetilde{\boldsymbol{X}}^{t+1}-\boldsymbol{X}^{\star}\right\|\left\|\boldsymbol{X}^{\star}\right\| \leq \sigma_{\min}/2,$$

where the second inequality holds according to (136). Weyl's inequality guarantees that

$$\boldsymbol{X}^{\star\top}\widetilde{\boldsymbol{X}}^{t+1} \succeq \frac{1}{2}\sigma_{\min}\boldsymbol{I}_r,$$

thus justifying (138) via Lemma 35.

(b) With (137) and (138) in place, we resort to Lemma 37 to establish the bound. Specifically, take $\boldsymbol{X}_1 = \widetilde{\boldsymbol{X}}^{t+1}$ and $\boldsymbol{X}_2 = \boldsymbol{X}^{t+1}\widehat{\boldsymbol{H}}^{t}$, and it comes from (136) that

$$\left\|\boldsymbol{X}_1-\boldsymbol{X}^{\star}\right\|\left\|\boldsymbol{X}^{\star}\right\| \leq \sigma_{\min}/2.$$

Moreover, we have

$$\|\boldsymbol{X}_1-\boldsymbol{X}_2\|\left\|\boldsymbol{X}^{\star}\right\| = \left\|\boldsymbol{X}^{t+1}\widehat{\boldsymbol{H}}^{t}-\widetilde{\boldsymbol{X}}^{t+1}\right\|\left\|\boldsymbol{X}^{\star}\right\|,$$

in which

$$\begin{aligned}\boldsymbol{X}^{t+1}\widehat{\boldsymbol{H}}^{t}-\widetilde{\boldsymbol{X}}^{t+1} &= \left(\boldsymbol{X}^{t}-\eta\frac{1}{p}\mathcal{P}_{\Omega}\left[\boldsymbol{X}^{t}\boldsymbol{X}^{t\top}-\left(\boldsymbol{M}^{\star}+\boldsymbol{E}\right)\right]\boldsymbol{X}^{t}\right)\widehat{\boldsymbol{H}}^{t} \\ &\quad -\left[\boldsymbol{X}^{t}\widehat{\boldsymbol{H}}^{t}-\eta\frac{1}{p}\mathcal{P}_{\Omega}\left[\boldsymbol{X}^{t}\boldsymbol{X}^{t\top}-\left(\boldsymbol{M}^{\star}+\boldsymbol{E}\right)\right]\boldsymbol{X}^{\star}\right] \\ &= -\eta\frac{1}{p}\mathcal{P}_{\Omega}\left[\boldsymbol{X}^{t}\boldsymbol{X}^{t\top}-\left(\boldsymbol{M}^{\star}+\boldsymbol{E}\right)\right]\left(\boldsymbol{X}^{t}\widehat{\boldsymbol{H}}^{t}-\boldsymbol{X}^{\star}\right).\end{aligned}$$

This allows one to derive

$$\begin{aligned}&\left\|\boldsymbol{X}^{t+1}\widehat{\boldsymbol{H}}^{t}-\widetilde{\boldsymbol{X}}^{t+1}\right\| \\ &\quad\leq \eta\left\|\frac{1}{p}\mathcal{P}_{\Omega}\left[\boldsymbol{X}^{t}\boldsymbol{X}^{t\top}-\boldsymbol{M}^{\star}\right]\left(\boldsymbol{X}^{t}\widehat{\boldsymbol{H}}^{t}-\boldsymbol{X}^{\star}\right)\right\| \\ &\qquad+\eta\left\|\frac{1}{p}\mathcal{P}_{\Omega}(\boldsymbol{E})\left(\boldsymbol{X}^{t}\widehat{\boldsymbol{H}}^{t}-\boldsymbol{X}^{\star}\right)\right\| \\ &\quad\leq \eta\left(2n\left\|\boldsymbol{\Delta}^{t}\right\|_{2,\infty}^{2}+4\sqrt{n}\log n\left\|\boldsymbol{\Delta}^{t}\right\|_{2,\infty}\left\|\boldsymbol{X}^{\star}\right\|+C\sigma\sqrt{\frac{n}{p}}\right)\left\|\boldsymbol{\Delta}^{t}\right\| \quad (139)\end{aligned}$$

for some absolute constant $C > 0$. Here the last inequality follows from Lemmas 40 and 43. As a consequence,

$$\begin{aligned}&\|\boldsymbol{X}_1-\boldsymbol{X}_2\|\left\|\boldsymbol{X}^\star\right\|\\&\quad\le\eta\left(2n\left\|\boldsymbol{\Delta}^t\right\|_{2,\infty}^2+4\sqrt{n}\log n\left\|\boldsymbol{\Delta}^t\right\|_{2,\infty}\left\|\boldsymbol{X}^\star\right\|+C\sigma\sqrt{\frac{n}{p}}\right)\left\|\boldsymbol{\Delta}^t\right\|\left\|\boldsymbol{X}^\star\right\|.\end{aligned}$$

Under our sample size condition and the noise condition (27) and the induction hypotheses (28), one can show

$$\|\boldsymbol{X}_1-\boldsymbol{X}_2\|\left\|\boldsymbol{X}^\star\right\|\le\sigma_{\min}/4.$$

Apply Lemma 37 and (139) to reach

$$\begin{aligned}\alpha_1&\le 5\kappa\left\|\boldsymbol{X}^{t+1}\widehat{\boldsymbol{H}}^t-\widetilde{\boldsymbol{X}}^{t+1}\right\|\\&\le 5\kappa\eta\left(2n\left\|\boldsymbol{\Delta}^t\right\|_{2,\infty}^2+2\sqrt{n}\log n\left\|\boldsymbol{\Delta}^t\right\|_{2,\infty}\left\|\boldsymbol{X}^\star\right\|+C\sigma\sqrt{\frac{n}{p}}\right)\left\|\boldsymbol{\Delta}^t\right\|.\end{aligned}$$

3. Combining the above bounds on α_1 and α_2, we arrive at

$$\begin{aligned}&\left\|\boldsymbol{X}^{t+1}\widehat{\boldsymbol{H}}^{t+1}-\boldsymbol{X}^\star\right\|\\&\quad\le\left(1-\frac{\sigma_{\min}}{2}\eta\right)\left\|\boldsymbol{\Delta}^t\right\|+\eta C\sigma\sqrt{\frac{n}{p}}\left\|\boldsymbol{X}^\star\right\|\\&\qquad+\widetilde{C}\eta\sqrt{\frac{\kappa^2\mu^2r^3\log^3 n}{np}}\sigma_{\max}\left(C_5\rho^t\mu r\sqrt{\frac{1}{np}}+\frac{C_8}{\sigma_{\min}}\sigma\sqrt{\frac{n}{p}}\right)\left\|\boldsymbol{X}^\star\right\|\\&\qquad+5\eta\kappa\left(2n\left\|\boldsymbol{\Delta}^t\right\|_{2,\infty}^2+2\sqrt{n}\log n\left\|\boldsymbol{\Delta}^t\right\|_{2,\infty}\left\|\boldsymbol{X}^\star\right\|+C\sigma\sqrt{\frac{n}{p}}\right)\left\|\boldsymbol{\Delta}^t\right\|\\&\quad\le C_9\rho^{t+1}\mu r\frac{1}{\sqrt{np}}\left\|\boldsymbol{X}^\star\right\|+C_{10}\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\left\|\boldsymbol{X}^\star\right\|,\end{aligned}$$

with the proviso that $\rho\ge 1-(\sigma_{\min}/3)\cdot\eta$, κ is a constant, and $n^2p\gg\mu^3r^3n\log^3 n$.

B.3.1 Proof of Lemma 22

In what follows, we first assume that the $\delta_{j,k}$'s are independent and then use the standard decoupling trick to extend the result to symmetric sampling case (i.e., $\delta_{j,k}=\delta_{k,j}$).

To begin with, we justify the concentration bound for any $\boldsymbol{\Delta}^t$ independent of Ω, followed by the standard covering argument that extends the bound to all $\boldsymbol{\Delta}^t$. For any $\boldsymbol{\Delta}^t$ independent of Ω, one has

$$B:=\max_{1\le j\le n}\left\|\boldsymbol{X}^\star_{l,\cdot}\left(\delta_{l,j}-p\right)\boldsymbol{\Delta}^{t\top}_{j,\cdot}\boldsymbol{X}^\star_{j,\cdot}\right\|_2\le\left\|\boldsymbol{X}^\star\right\|_{2,\infty}^2\xi$$

and

$$V:=\left\|\mathbb{E}\left[\sum_{j=1}^n\left(\delta_{l,j}-p\right)^2\boldsymbol{X}^\star_{l,\cdot}\boldsymbol{\Delta}^{t\top}_{j,\cdot}\boldsymbol{X}^\star_{j,\cdot}\left(\boldsymbol{X}^\star_{l,\cdot}\boldsymbol{\Delta}^{t\top}_{j,\cdot}\boldsymbol{X}^\star_{j,\cdot}\right)^\top\right]\right\|$$

$$\begin{aligned}
&\leq p \left\| \boldsymbol{X}^{\star}_{l,\cdot} \right\|_2^2 \left\| \boldsymbol{X}^{\star} \right\|_{2,\infty}^2 \left\| \sum_{j=1}^{n} \boldsymbol{\Delta}^{t\top}_{j,\cdot} \boldsymbol{\Delta}^{t}_{j,\cdot} \right\| \\
&\leq p \left\| \boldsymbol{X}^{\star}_{l,\cdot} \right\|_2^2 \left\| \boldsymbol{X}^{\star} \right\|_{2,\infty}^2 \psi^2 \\
&\leq 2p \left\| \boldsymbol{X}^{\star} \right\|_{2,\infty}^2 \xi^2 \sigma_{\max},
\end{aligned}$$

where ξ and ψ are defined, respectively, in (130) and (131). Here, the last line makes use of the fact that

$$\left\| \boldsymbol{X}^{\star} \right\|_{2,\infty} \psi \ll \xi \left\| \boldsymbol{X}^{\star} \right\| = \xi \sqrt{\sigma_{\max}}, \tag{140}$$

as long as n is sufficiently large. Apply the matrix Bernstein inequality [114, Theorem 6.1.1] to get

$$\begin{aligned}
\mathbb{P}\left\{ \left\| \boldsymbol{A}_{l,\cdot} \right\|_2 \geq t \right\} &\leq 2r \exp\left(-\frac{ct^2}{2p\xi^2\sigma_{\max} \left\| \boldsymbol{X}^{\star} \right\|_{2,\infty}^2 + t \cdot \left\| \boldsymbol{X}^{\star} \right\|_{2,\infty}^2 \xi} \right) \\
&\leq 2r \exp\left(-\frac{ct^2}{4p\xi^2\sigma_{\max} \left\| \boldsymbol{X}^{\star} \right\|_{2,\infty}^2} \right)
\end{aligned}$$

for some constant $c > 0$, provided that

$$t \leq 2p\sigma_{\max}\xi.$$

This upper bound on t is exactly the truncation level ω we introduce in (132). With this in mind, we can easily verify that

$$\left\| \boldsymbol{A}_{l,\cdot} \right\|_2 \mathbb{1}_{\{\|\boldsymbol{A}_{l,\cdot}\|_2 \leq \omega\}}$$

is a sub-Gaussian random variable with variance proxy not exceeding $O\left(p\xi^2\sigma_{\max} \left\| \boldsymbol{X}^{\star} \right\|_{2,\infty}^2 \log r \right)$. Therefore, invoking the concentration bounds for quadratic functions [57, Theorem 2.1] yields that for some constants $C_0, C > 0$, with probability at least $1 - C_0 e^{-Cnr\log n}$,

$$\phi_1^2 = \sum_{l=1}^{n} \left\| \boldsymbol{A}_{l,\cdot} \right\|_2^2 \mathbb{1}_{\{\|\boldsymbol{A}_{l,\cdot}\|_2 \leq \omega\}} \lesssim p\xi^2\sigma_{\max} \|\boldsymbol{X}^{\star}\|_{2,\infty}^2 nr \log^2 n.$$

Now that we have established an upper bound on any fixed matrix $\boldsymbol{\Delta}^t$ (which holds with exponentially high probability), we can proceed to invoke the standard epsilon-net argument to establish a uniform bound over all feasible $\boldsymbol{\Delta}^t$. This argument is fairly standard and is thus omitted; see [111, Section 2.3.1] or the proof of Lemma 42. In conclusion, we have that with probability exceeding $1 - C_0 e^{-\frac{1}{2}Cnr\log n}$,

$$\phi_1 = \sqrt{\sum_{l=1}^{n} \left\|\boldsymbol{A}_{l,\cdot}\right\|_2^2 \mathbb{1}_{\{\|\boldsymbol{A}_{l,\cdot}\|_2 \le \omega\}}} \lesssim \sqrt{p\xi^2 \sigma_{\max} \|\boldsymbol{X}^\star\|_{2,\infty}^2 nr \log^2 n} \tag{141}$$

holds simultaneously for all $\boldsymbol{\Delta}^t \in \mathbb{R}^{n\times r}$ obeying the conditions of the lemma.

In the end, we comment on how to extend the bound to the symmetric sampling pattern where $\delta_{j,k} = \delta_{k,j}$. Recall from (129) that the diagonal element $\delta_{l,l}$ cannot change the ℓ_2 norm of $\boldsymbol{A}_{l,\cdot}$ by more than $\left\|\boldsymbol{X}^\star\right\|_{2,\infty}^2 \xi$. As a result, changing all the diagonals $\{\delta_{l,l}\}$ cannot change the quantity of interest (i.e., ϕ_1) by more than $\sqrt{n}\left\|\boldsymbol{X}^\star\right\|_{2,\infty}^2 \xi$. This is smaller than the right-hand side of (141) under our incoherence and sample size conditions. Hence, from now on, we ignore the effect of $\{\delta_{l,l}\}$ and focus on off-diagonal terms. The proof then follows from the same argument as in [48, Theorem D.2]. More specifically, we can employ the construction of Bernoulli random variables introduced therein to demonstrate that the upper bound in (141) still holds if the indicator $\delta_{i,j}$ is replaced by $(\tau_{i,j} + \tau'_{i,j})/2$, where $\tau_{i,j}$ and $\tau'_{i,j}$ are independent copies of the symmetric Bernoulli random variables. Recognizing that $\sup_{\boldsymbol{\Delta}^t} \phi_1$ is a norm of the Bernoulli random variables $\tau_{i,j}$, one can repeat the decoupling argument in [48, Claim D.3] to finish the proof. We omit the details here for brevity.

B.3.2 Proof of Lemma 23

Observe from (129) that

$$\begin{aligned}
\left\|\boldsymbol{A}_{l,\cdot}\right\|_2 &\le \left\|\boldsymbol{X}^\star\right\|_{2,\infty} \left\|\sum_{j=1}^{n} \left(\delta_{l,j} - p\right) \boldsymbol{\Delta}_{j,\cdot}^{t\top} \boldsymbol{X}_{j,\cdot}^\star\right\| \qquad (142)\\
&\le \left\|\boldsymbol{X}^\star\right\|_{2,\infty} \left(\left\|\sum_{j=1}^{n} \delta_{l,j} \boldsymbol{\Delta}_{j,\cdot}^{t\top} \boldsymbol{X}_{j,\cdot}^\star\right\| + p\left\|\boldsymbol{\Delta}^t\right\| \left\|\boldsymbol{X}^\star\right\|\right)\\
&\le \left\|\boldsymbol{X}^\star\right\|_{2,\infty} \left(\left\|\left[\delta_{l,1}\boldsymbol{\Delta}_{1,\cdot}^{t\top}, \cdots, \delta_{l,n}\boldsymbol{\Delta}_{n,\cdot}^{t\top}\right]\right\| \left\|\begin{bmatrix}\delta_{l,1}\boldsymbol{X}_{1,\cdot}^\star\\ \vdots \\ \delta_{l,n}\boldsymbol{X}_{n,\cdot}^\star\end{bmatrix}\right\| + p\psi\left\|\boldsymbol{X}^\star\right\|\right)\\
&\le \left\|\boldsymbol{X}^\star\right\|_{2,\infty} \left(\left\|\boldsymbol{G}_l\left(\boldsymbol{\Delta}^t\right)\right\| \cdot 1.2\sqrt{p}\left\|\boldsymbol{X}^\star\right\| + p\psi\left\|\boldsymbol{X}^\star\right\|\right), \qquad (143)
\end{aligned}$$

where ψ is as defined in (131) and $\boldsymbol{G}_l(\cdot)$ is as defined in Lemma 41. Here, the last inequality follows from Lemma 41; namely, for some constant $C > 0$, the following holds with probability at least $1 - O(n^{-10})$

$$\left\|\begin{bmatrix}\delta_{l,1}\boldsymbol{X}_{1,\cdot}^\star\\ \vdots \\ \delta_{l,n}\boldsymbol{X}_{n,\cdot}^\star\end{bmatrix}\right\| \le \left(p\left\|\boldsymbol{X}^\star\right\|^2 + C\sqrt{p\|\boldsymbol{X}^\star\|_{2,\infty}^2 \left\|\boldsymbol{X}^\star\right\|^2 \log n} + C\|\boldsymbol{X}^\star\|_{2,\infty}^2 \log n\right)^{\frac{1}{2}}$$

$$\leq \left(p + C\sqrt{p\frac{\kappa\mu r}{n}\log n} + C\frac{\kappa\mu r\log n}{n} \right)^{\frac{1}{2}} \left\| \boldsymbol{X}^\star \right\| \leq 1.2\sqrt{p}\left\| \boldsymbol{X}^\star \right\|, \tag{144}$$

where we also use the incoherence condition (114) and the sample complexity condition $n^2 p \gg \kappa\mu r n \log n$. Hence, the event

$$\left\| \boldsymbol{A}_{l,\cdot} \right\|_2 \geq \omega = 2p\sigma_{\max}\xi$$

together with (142) and (143) necessarily implies that

$$\left\| \sum_{j=1}^{n} \left(\delta_{l,j} - p\right) \boldsymbol{\Delta}_{j,\cdot}^{t\top} \boldsymbol{X}_{j,\cdot}^\star \right\| \geq 2p\sigma_{\max}\frac{\xi}{\left\| \boldsymbol{X}^\star \right\|_{2,\infty}} \quad \text{and}$$

$$\left\| \boldsymbol{G}_l\left(\boldsymbol{\Delta}^t\right) \right\| \geq \frac{\frac{2p\sigma_{\max}\xi}{\|\boldsymbol{X}^\star\|\|\boldsymbol{X}^\star\|_{2,\infty}} - p\psi}{1.2\sqrt{p}} \geq \frac{\frac{2\sqrt{p}\|\boldsymbol{X}^\star\|\xi}{\|\boldsymbol{X}^\star\|_{2,\infty}} - \sqrt{p}\psi}{1.2} \geq 1.5\sqrt{p}\frac{\xi}{\left\| \boldsymbol{X}^\star \right\|_{2,\infty}}\left\| \boldsymbol{X}^\star \right\|,$$

where the last inequality follows from bound (140). As a result, with probability at least $1 - O(n^{-10})$ (i.e., when (144) holds for all l's) we can upper bound ϕ_2 by

$$\phi_2 = \sqrt{\sum_{l=1}^{n} \left\| \boldsymbol{A}_{l,\cdot} \right\|_2^2 \mathbb{1}_{\{\|\boldsymbol{A}_{l,\cdot}\|_2 \geq \omega\}}} \leq \sqrt{\sum_{l=1}^{n} \left\| \boldsymbol{A}_{l,\cdot} \right\|_2^2 \mathbb{1}_{\left\{\|\boldsymbol{G}_l(\boldsymbol{\Delta}^t)\| \geq \frac{1.5\sqrt{p}\xi\sqrt{\sigma_{\max}}}{\|\boldsymbol{X}^\star\|_{2,\infty}}\right\}}},$$

where the indicator functions are now specified with respect to $\left\| \boldsymbol{G}_l\left(\boldsymbol{\Delta}^t\right) \right\|$.

Next, we divide into multiple cases based on the size of $\left\| \boldsymbol{G}_l\left(\boldsymbol{\Delta}^t\right) \right\|$. By Lemma 42, for some constants $c_1, c_2 > 0$, with probability at least $1 - c_1\exp\left(-c_2 nr\log n\right)$,

$$\sum_{l=1}^{n} \mathbb{1}_{\left\{\|\boldsymbol{G}_l(\boldsymbol{\Delta}^t)\| \geq 4\sqrt{p}\psi + \sqrt{2^k r}\xi\right\}} \leq \frac{\alpha n}{2^{k-3}} \tag{145}$$

for any $k \geq 0$ and any $\alpha \gtrsim \log n$. We claim that it suffices to consider the set of sufficiently large k obeying

$$\sqrt{2^k r}\xi \geq 4\sqrt{p}\psi \quad \text{or equivalently} \quad k \geq \log\frac{16p\psi^2}{r\xi^2}; \tag{146}$$

otherwise, we can use (140) to obtain

$$4\sqrt{p}\psi + \sqrt{2^k r}\xi \leq 8\sqrt{p}\psi \ll 1.5\sqrt{p}\frac{\xi}{\left\| \boldsymbol{X}^\star \right\|_{2,\infty}}\left\| \boldsymbol{X}^\star \right\|,$$

which contradicts the event $\left\|\boldsymbol{A}_{l,\cdot}\right\|_2 \geq \omega$. Consequently, we divide all indices into the following sets

$$S_k = \left\{1 \leq l \leq n : \left\|\boldsymbol{G}_l\left(\boldsymbol{\Delta}^t\right)\right\| \in \left(\sqrt{2^k r}\xi, \sqrt{2^{k+1} r}\xi\right]\right\} \tag{147}$$

defined for each integer k obeying (146). Under condition (146), it follows from (145) that

$$\sum_{l=1}^{n} \mathbb{1}_{\left\{\|\boldsymbol{G}_l(\boldsymbol{\Delta}^t)\| \geq \sqrt{2^{k+2} r}\xi\right\}} \leq \sum_{l=1}^{n} \mathbb{1}_{\left\{\|\boldsymbol{G}_l(\boldsymbol{\Delta}^t)\| \geq 4\sqrt{p}\psi + \sqrt{2^k r}\xi\right\}} \leq \frac{\alpha n}{2^{k-3}},$$

meaning that the cardinality of S_k satisfies

$$|S_{k+2}| \leq \frac{\alpha n}{2^{k-3}} \qquad \text{or} \qquad |S_k| \leq \frac{\alpha n}{2^{k-5}}$$

which decays exponentially fast as k increases. Therefore, when restricting attention to the set of indices within S_k, we can obtain

$$\begin{aligned}
\sqrt{\sum_{l \in S_k} \left\|\boldsymbol{A}_{l,\cdot}\right\|_2^2} &\overset{\text{(i)}}{\leq} \sqrt{|S_k| \cdot \left\|\boldsymbol{X}^\star\right\|_{2,\infty}^2 \left(1.2\sqrt{2^{k+1} r}\xi\sqrt{p}\left\|\boldsymbol{X}^\star\right\| + p\psi\left\|\boldsymbol{X}^\star\right\|\right)^2} \\
&\leq \sqrt{\frac{\alpha n}{2^{k-5}}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty}\left(2\sqrt{2^{k+1} r}\xi\sqrt{p}\left\|\boldsymbol{X}^\star\right\| + p\psi\left\|\boldsymbol{X}^\star\right\|\right) \\
&\overset{\text{(ii)}}{\leq} 4\sqrt{\frac{\alpha n}{2^{k-5}}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty}\sqrt{2^{k+1} r}\xi\sqrt{p}\left\|\boldsymbol{X}^\star\right\| \\
&\overset{\text{(iii)}}{\leq} 32\sqrt{\alpha\kappa\mu r^2 p}\xi\left\|\boldsymbol{X}^\star\right\|^2,
\end{aligned}$$

where (i) follows from bound (143) and constraint (147) in S_k, (ii) is a consequence of (146), and (iii) uses the incoherence condition (114).

Now that we have developed an upper bound with respect to each S_k, we can add them up to yield the final upper bound. Note that there are in total no more than $O(\log n)$ different sets, i.e., $S_k = \emptyset$ if $k \geq c_1 \log n$ for c_1 sufficiently large. This arises since

$$\|\boldsymbol{G}_l(\boldsymbol{\Delta}^t)\| \leq \|\boldsymbol{\Delta}^t\|_{\mathrm{F}} \leq \sqrt{n}\|\boldsymbol{\Delta}^t\|_{2,\infty} \leq \sqrt{n}\xi \leq \sqrt{n}\sqrt{r}\xi$$

and hence

$$\mathbb{1}_{\left\{\|\boldsymbol{G}_l(\boldsymbol{\Delta}^t)\| \geq 4\sqrt{p}\psi + \sqrt{2^k r}\xi\right\}} = 0 \qquad \text{and} \qquad S_k = \emptyset$$

if $k/\log n$ is sufficiently large. One can thus conclude that

$$\phi_2^2 \le \sum_{k=\log \frac{16p\psi^2}{r\xi^2}}^{c_1 \log n} \sum_{l \in S_k} \left\| \boldsymbol{A}_{l,\cdot} \right\|_2^2 \lesssim \left(\sqrt{\alpha\kappa\mu r^2 p \xi} \left\| \boldsymbol{X}^\star \right\|^2 \right)^2 \cdot \log n,$$

leading to $\phi_2 \lesssim \xi \sqrt{\alpha\kappa\mu r^2 p \log n} \left\| \boldsymbol{X}^\star \right\|^2$. The proof is finished by taking $\alpha = c \log n$ for some sufficiently large constant $c > 0$.

B.4 Proof of Lemma 10

1. To obtain (73a), we invoke Lemma 37. Setting $\boldsymbol{X}_1 = \boldsymbol{X}^t \widehat{\boldsymbol{H}}^t$ and $\boldsymbol{X}_2 = \boldsymbol{X}^{t,(l)} \boldsymbol{R}^{t,(l)}$, we get

$$\left\| \boldsymbol{X}_1 - \boldsymbol{X}^\star \right\| \left\| \boldsymbol{X}^\star \right\| \overset{\text{(i)}}{\le} C_9 \rho^t \mu r \frac{1}{\sqrt{np}} \sigma_{\max} + \frac{C_{10}}{\sigma_{\min}} \sigma \sqrt{\frac{n \log n}{p}} \sigma_{\max} \overset{\text{(ii)}}{\le} \frac{1}{2} \sigma_{\min},$$

where (i) follows from (70c) and (ii) holds as long as $n^2 p \gg \kappa^2 \mu^2 r^2 n$ and the noise satisfies (27). In addition,

$$\begin{aligned} \|\boldsymbol{X}_1 - \boldsymbol{X}_2\| \left\| \boldsymbol{X}^\star \right\| &\le \|\boldsymbol{X}_1 - \boldsymbol{X}_2\|_{\mathrm{F}} \left\| \boldsymbol{X}^\star \right\| \\ &\overset{\text{(i)}}{\le} \left(C_3 \rho^t \mu r \sqrt{\frac{\log n}{np}} \left\| \boldsymbol{X}^\star \right\|_{2,\infty} + \frac{C_7}{\sigma_{\min}} \sigma \sqrt{\frac{n \log n}{p}} \left\| \boldsymbol{X}^\star \right\|_{2,\infty} \right) \left\| \boldsymbol{X}^\star \right\| \\ &\overset{\text{(ii)}}{\le} C_3 \rho^t \mu r \sqrt{\frac{\log n}{np}} \sigma_{\max} + \frac{C_7}{\sigma_{\min}} \sigma \sqrt{\frac{n \log n}{p}} \sigma_{\max} \\ &\overset{\text{(iii)}}{\le} \frac{1}{2} \sigma_{\min}, \end{aligned}$$

where (i) utilizes (70d), (ii) follows since $\left\| \boldsymbol{X}^\star \right\|_{2,\infty} \le \left\| \boldsymbol{X}^\star \right\|$, and (iii) holds if $n^2 p \gg \kappa^2 \mu^2 r^2 n \log n$ and the noise satisfies (27). With these in place, Lemma 37 immediately yields (73a).

2. The first inequality in (73b) follows directly from the definition of $\widehat{\boldsymbol{H}}^{t,(l)}$. The second inequality is concerned with the estimation error of $\boldsymbol{X}^{t,(l)} \boldsymbol{R}^{t,(l)}$ with respect to the Frobenius norm. Combining (70a), (70d), and the triangle inequality yields

$$\begin{aligned} \left\| \boldsymbol{X}^{t,(l)} \boldsymbol{R}^{t,(l)} - \boldsymbol{X}^\star \right\|_{\mathrm{F}} &\le \left\| \boldsymbol{X}^t \widehat{\boldsymbol{H}}^t - \boldsymbol{X}^\star \right\|_{\mathrm{F}} + \left\| \boldsymbol{X}^t \widehat{\boldsymbol{H}}^t - \boldsymbol{X}^{t,(l)} \boldsymbol{R}^{t,(l)} \right\|_{\mathrm{F}} \\ &\le C_4 \rho^t \mu r \frac{1}{\sqrt{np}} \left\| \boldsymbol{X}^\star \right\|_{\mathrm{F}} + \frac{C_1 \sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \left\| \boldsymbol{X}^\star \right\|_{\mathrm{F}} + C_3 \rho^t \mu r \sqrt{\frac{\log n}{np}} \left\| \boldsymbol{X}^\star \right\|_{2,\infty} \\ &\quad + \frac{C_7 \sigma}{\sigma_{\min}} \sqrt{\frac{n \log n}{p}} \left\| \boldsymbol{X}^\star \right\|_{2,\infty} \end{aligned}$$

$$\leq C_4\rho^t\mu r\frac{1}{\sqrt{np}}\left\|\boldsymbol{X}^\star\right\|_{\mathrm{F}}+\frac{C_1\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\left\|\boldsymbol{X}^\star\right\|_{\mathrm{F}}+C_3\rho^t\mu r\sqrt{\frac{\log n}{np}}\sqrt{\frac{\kappa\mu}{n}}\left\|\boldsymbol{X}^\star\right\|_{\mathrm{F}}$$
$$+\frac{C_7\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\sqrt{\frac{\kappa\mu}{n}}\left\|\boldsymbol{X}^\star\right\|_{\mathrm{F}}$$
$$\leq 2C_4\rho^t\mu r\frac{1}{\sqrt{np}}\left\|\boldsymbol{X}^\star\right\|_{\mathrm{F}}+\frac{2C_1\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\left\|\boldsymbol{X}^\star\right\|_{\mathrm{F}}, \tag{148}$$

where the last step holds true as long as $n\gg\kappa\mu\log n$.

3. To obtain (73c), we use (70d) and (70b) to get

$$\left\|\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}-\boldsymbol{X}^\star\right\|_{2,\infty}\leq\left\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t-\boldsymbol{X}^\star\right\|_{2,\infty}+\left\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t-\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}\right\|_{\mathrm{F}}$$
$$\leq C_5\rho^t\mu r\sqrt{\frac{\log n}{np}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty}+\frac{C_8\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty}$$
$$+C_3\rho^t\mu r\sqrt{\frac{\log n}{np}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty}+\frac{C_7\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty}$$
$$\leq(C_3+C_5)\,\rho^t\mu r\sqrt{\frac{\log n}{np}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty}+\frac{C_8+C_7}{\sigma_{\min}}\sigma\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty}.$$

4. Finally, to obtain (73d), one can take the triangle inequality

$$\left\|\boldsymbol{X}^{t,(l)}\widehat{\boldsymbol{H}}^{t,(l)}-\boldsymbol{X}^\star\right\|\leq\left\|\boldsymbol{X}^{t,(l)}\widehat{\boldsymbol{H}}^{t,(l)}-\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t\right\|_{\mathrm{F}}+\left\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t-\boldsymbol{X}^\star\right\|$$
$$\leq 5\kappa\left\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t-\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}\right\|_{\mathrm{F}}+\left\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t-\boldsymbol{X}^\star\right\|,$$

where the second line follows from (73a). Combine (70d) and (70c) to yield

$$\left\|\boldsymbol{X}^{t,(l)}\widehat{\boldsymbol{H}}^{t,(l)}-\boldsymbol{X}^\star\right\|$$
$$\leq 5\kappa\left(C_3\rho^t\mu r\sqrt{\frac{\log n}{np}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty}+\frac{C_7}{\sigma_{\min}}\sigma\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty}\right)$$
$$+C_9\rho^t\mu r\frac{1}{\sqrt{np}}\left\|\boldsymbol{X}^\star\right\|+\frac{C_{10}}{\sigma_{\min}}\sigma\sqrt{\frac{n}{p}}\left\|\boldsymbol{X}^\star\right\|$$
$$\leq 5\kappa\sqrt{\frac{\kappa\mu r}{n}}\left\|\boldsymbol{X}^\star\right\|\left(C_3\rho^t\mu r\sqrt{\frac{\log n}{np}}+\frac{C_7}{\sigma_{\min}}\sigma\sqrt{\frac{n\log n}{p}}\right)$$
$$+C_9\rho^t\mu r\frac{1}{\sqrt{np}}\left\|\boldsymbol{X}^\star\right\|+\frac{C_{10}\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\left\|\boldsymbol{X}^\star\right\|$$
$$\leq 2C_9\rho^t\mu r\frac{1}{\sqrt{np}}\left\|\boldsymbol{X}^\star\right\|+\frac{2C_{10}\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\left\|\boldsymbol{X}^\star\right\|,$$

where the second inequality uses the incoherence of $\boldsymbol{X}^\star$ (cf. (114)) and the last inequality holds as long as $n \gg \kappa^3 \mu r \log n$.

B.5 Proof of Lemma 11

From the definition of $\boldsymbol{R}^{t+1,(l)}$ (see (72)), we must have

$$\left\| \boldsymbol{X}^{t+1} \widehat{\boldsymbol{H}}^{t+1} - \boldsymbol{X}^{t+1,(l)} \boldsymbol{R}^{t+1,(l)} \right\|_{\mathrm{F}} \le \left\| \boldsymbol{X}^{t+1} \widehat{\boldsymbol{H}}^{t} - \boldsymbol{X}^{t+1,(l)} \boldsymbol{R}^{t,(l)} \right\|_{\mathrm{F}}.$$

The gradient update rules in (24) and (69) allow one to express

$$\begin{aligned}
&\boldsymbol{X}^{t+1} \widehat{\boldsymbol{H}}^{t} - \boldsymbol{X}^{t+1,(l)} \boldsymbol{R}^{t,(l)} \\
&= \left[\boldsymbol{X}^{t} - \eta \nabla f\left(\boldsymbol{X}^{t}\right)\right] \widehat{\boldsymbol{H}}^{t} - \left[\boldsymbol{X}^{t,(l)} - \eta \nabla f^{(l)}\left(\boldsymbol{X}^{t,(l)}\right)\right] \boldsymbol{R}^{t,(l)} \\
&= \boldsymbol{X}^{t} \widehat{\boldsymbol{H}}^{t} - \eta \nabla f\left(\boldsymbol{X}^{t} \widehat{\boldsymbol{H}}^{t}\right) - \left[\boldsymbol{X}^{t,(l)} \boldsymbol{R}^{t,(l)} - \eta \nabla f^{(l)}\left(\boldsymbol{X}^{t,(l)} \boldsymbol{R}^{t,(l)}\right)\right] \\
&= \left(\boldsymbol{X}^{t} \widehat{\boldsymbol{H}}^{t} - \boldsymbol{X}^{t,(l)} \boldsymbol{R}^{t,(l)}\right) - \eta \left[\nabla f(\boldsymbol{X}^{t} \widehat{\boldsymbol{H}}^{t}) - \nabla f\left(\boldsymbol{X}^{t,(l)} \boldsymbol{R}^{t,(l)}\right)\right] \\
&\quad - \eta \left[\nabla f\left(\boldsymbol{X}^{t,(l)} \boldsymbol{R}^{t,(l)}\right) - \nabla f^{(l)}\left(\boldsymbol{X}^{t,(l)} \boldsymbol{R}^{t,(l)}\right)\right],
\end{aligned}$$

where we have again used the fact that $\nabla f\left(\boldsymbol{X}^{t}\right) \boldsymbol{R} = \nabla f(\boldsymbol{X}^{t} \boldsymbol{R})$ for any orthonormal matrix $\boldsymbol{R} \in \mathcal{O}^{r \times r}$ (similarly for $\nabla f^{(l)}\left(\boldsymbol{X}^{t,(l)}\right)$). Relate the right-hand side of the above equation with $\nabla f_{\mathrm{clean}}(\boldsymbol{X})$ to reach

$$\begin{aligned}
&\boldsymbol{X}^{t+1} \widehat{\boldsymbol{H}}^{t} - \boldsymbol{X}^{t+1,(l)} \boldsymbol{R}^{t,(l)} \\
&= \underbrace{\left(\boldsymbol{X}^{t} \widehat{\boldsymbol{H}}^{t} - \boldsymbol{X}^{t,(l)} \boldsymbol{R}^{t,(l)}\right) - \eta \left[\nabla f_{\mathrm{clean}}\left(\boldsymbol{X}^{t} \widehat{\boldsymbol{H}}^{t}\right) - \nabla f_{\mathrm{clean}}\left(\boldsymbol{X}^{t,(l)} \boldsymbol{R}^{t,(l)}\right)\right]}_{:=\boldsymbol{B}_1^{(l)}} \\
&\quad \underbrace{- \eta \left[\frac{1}{p} \mathcal{P}_{\Omega_l}\left(\boldsymbol{X}^{t,(l)} \boldsymbol{X}^{t,(l)\top} - \boldsymbol{M}^\star\right) - \mathcal{P}_l\left(\boldsymbol{X}^{t,(l)} \boldsymbol{X}^{t,(l)\top} - \boldsymbol{M}^\star\right)\right] \boldsymbol{X}^{t,(l)} \boldsymbol{R}^{t,(l)}}_{:=\boldsymbol{B}_2^{(l)}} \\
&\quad + \underbrace{\eta \frac{1}{p} \mathcal{P}_{\Omega}(\boldsymbol{E}) \left(\boldsymbol{X}^{t} \widehat{\boldsymbol{H}}^{t} - \boldsymbol{X}^{t,(l)} \boldsymbol{R}^{t,(l)}\right)}_{:=\boldsymbol{B}_3^{(l)}} + \underbrace{\eta \frac{1}{p} \mathcal{P}_{\Omega_l}(\boldsymbol{E}) \boldsymbol{X}^{t,(l)} \boldsymbol{R}^{t,(l)}}_{:=\boldsymbol{B}_4^{(l)}},
\end{aligned} \tag{149}$$

where we have used the following relationship between $\nabla f^{(l)}(\boldsymbol{X})$ and $\nabla f(\boldsymbol{X})$:

$$\nabla f^{(l)}(\boldsymbol{X}) = \nabla f(\boldsymbol{X}) - \frac{1}{p} \mathcal{P}_{\Omega_l}\left[\boldsymbol{X}\boldsymbol{X}^\top - \left(\boldsymbol{M}^\star + \boldsymbol{E}\right)\right] \boldsymbol{X} + \mathcal{P}_l\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M}^\star\right) \boldsymbol{X} \tag{150}$$

for all $\boldsymbol{X} \in \mathbb{R}^{n \times r}$ with $\mathcal{P}_{\Omega_l}$ and $\mathcal{P}_l$ defined, respectively, in (66) and (67). In the sequel, we control the four terms in reverse order.

1. The last term $\boldsymbol{B}_4^{(l)}$ is controlled via the following lemma.

Lemma 24 *Suppose that the sample size obeys $n^2 p > C\mu^2 r^2 n \log^2 n$ for some sufficiently large constant $C > 0$. Then with probability at least $1 - O\left(n^{-10}\right)$, the matrix $\boldsymbol{B}_4^{(l)}$ as defined in (149) satisfies*

$$\left\|\boldsymbol{B}_4^{(l)}\right\|_{\mathrm{F}} \lesssim \eta\sigma\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty}.$$

2. The third term $\boldsymbol{B}_3^{(l)}$ can be bounded as follows

$$\left\|\boldsymbol{B}_3^{(l)}\right\|_{\mathrm{F}} \le \eta\left\|\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{E}\right)\right\|\left\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t - \boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}\right\|_{\mathrm{F}} \lesssim \eta\sigma\sqrt{\frac{n}{p}}\left\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t - \boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}\right\|_{\mathrm{F}},$$

   where the second inequality comes from Lemma 40.
3. For the second term $\boldsymbol{B}_2^{(l)}$, we have the following lemma.

Lemma 25 *Suppose that the sample size obeys $n^2 p \gg \mu^2 r^2 n \log n$. Then with probability exceeding $1 - O\left(n^{-10}\right)$, the matrix $\boldsymbol{B}_2^{(l)}$ as defined in (149) satisfies*

$$\left\|\boldsymbol{B}_2^{(l)}\right\|_{\mathrm{F}} \lesssim \eta\sqrt{\frac{\kappa^2\mu^2 r^2\log n}{np}}\left\|\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)} - \boldsymbol{X}^\star\right\|_{2,\infty}\sigma_{\max}. \tag{151}$$

4. Regarding the first term $\boldsymbol{B}_1^{(l)}$, apply the fundamental theorem of calculus [70, Chapter XIII, Theorem 4.2] to get

$$\mathrm{vec}\left(\boldsymbol{B}_1^{(l)}\right) = \left(\boldsymbol{I}_{nr} - \eta\int_0^1 \nabla^2 f_{\mathrm{clean}}\left(\boldsymbol{X}(\tau)\right)\mathrm{d}\tau\right)\mathrm{vec}\left(\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t - \boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}\right), \tag{152}$$

   where we abuse the notation and denote $\boldsymbol{X}(\tau) := \boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)} + \tau\left(\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t - \boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}\right)$. Going through the same derivations as in the proof of Lemma 8 (see Appendix B.2), we get

$$\left\|\boldsymbol{B}_1^{(l)}\right\|_{\mathrm{F}} \le \left(1 - \frac{\sigma_{\min}}{4}\eta\right)\left\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t - \boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}\right\|_{\mathrm{F}} \tag{153}$$

   with the proviso that $0 < \eta \le (2\sigma_{\min})/(25\sigma_{\max}^2)$.

Applying the triangle inequality to (149) and invoking the preceding four bounds, we arrive at

$$\begin{aligned}&\left\|\boldsymbol{X}^{t+1}\widehat{\boldsymbol{H}}^{t+1} - \boldsymbol{X}^{t+1,(l)}\boldsymbol{R}^{t+1,(l)}\right\|_{\mathrm{F}}\\ &\le \left(1 - \frac{\sigma_{\min}}{4}\eta\right)\left\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t - \boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}\right\|_{\mathrm{F}}\end{aligned}$$

$$
\begin{aligned}
&+\widetilde{C}\eta\sqrt{\frac{\kappa^2\mu^2r^2\log n}{np}}\left\|\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}-\boldsymbol{X}^\star\right\|_{2,\infty}\sigma_{\max}\\
&+\widetilde{C}\eta\sigma\sqrt{\frac{n}{p}}\left\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t-\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}\right\|_{\mathrm{F}}+\widetilde{C}\eta\sigma\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty}\\
&=\left(1-\frac{\sigma_{\min}}{4}\eta+\widetilde{C}\eta\sigma\sqrt{\frac{n}{p}}\right)\left\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t-\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}\right\|_{\mathrm{F}}\\
&+\widetilde{C}\eta\sqrt{\frac{\kappa^2\mu^2r^2\log n}{np}}\left\|\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}-\boldsymbol{X}^\star\right\|_{2,\infty}\sigma_{\max}\\
&+\widetilde{C}\eta\sigma\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty}\\
&\le\left(1-\frac{2\sigma_{\min}}{9}\eta\right)\left\|\boldsymbol{X}^t\widehat{\boldsymbol{H}}^t-\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}\right\|_{\mathrm{F}}\\
&+\widetilde{C}\eta\sqrt{\frac{\kappa^2\mu^2r^2\log n}{np}}\left\|\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}-\boldsymbol{X}^\star\right\|_{2,\infty}\sigma_{\max}\\
&+\widetilde{C}\eta\sigma\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty}
\end{aligned}
$$

for some absolute constant $\widetilde{C}>0$. Here the last inequality holds as long as $\sigma\sqrt{n/p}\ll\sigma_{\min}$, which is satisfied under our noise condition (27). This taken collectively with hypotheses (70d) and (73c) leads to

$$
\begin{aligned}
&\left\|\boldsymbol{X}^{t+1}\widehat{\boldsymbol{H}}^{t+1}-\boldsymbol{X}^{t+1,(l)}\boldsymbol{R}^{t+1,(l)}\right\|_{\mathrm{F}}\\
&\le\left(1-\frac{2\sigma_{\min}}{9}\eta\right)\left(C_3\rho^t\mu r\sqrt{\frac{\log n}{np}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty}+C_7\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty}\right)\\
&\quad+\widetilde{C}\eta\sqrt{\frac{\kappa^2\mu^2r^2\log n}{np}}\left[(C_3+C_5)\,\rho^t\mu r\sqrt{\frac{\log n}{np}}+(C_8+C_7)\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\right]\left\|\boldsymbol{X}^\star\right\|_{2,\infty}\sigma_{\max}\\
&\quad+\widetilde{C}\eta\sigma\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty}\\
&\le\left(1-\frac{\sigma_{\min}}{5}\eta\right)C_3\rho^t\mu r\sqrt{\frac{\log n}{np}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty}+C_7\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty}
\end{aligned}
$$

as long as $C_7>0$ is sufficiently large, where we have used the sample complexity assumption $n^2p\gg\kappa^4\mu^2r^2n\log n$ and the step size $0<\eta\le1/(2\sigma_{\max})\le1/(2\sigma_{\min})$. This finishes the proof.

B.5.1 Proof of Lemma 24

By the unitary invariance of the Frobenius norm, one has

$$\left\|\boldsymbol{B}_4^{(l)}\right\|_{\mathrm{F}} = \frac{\eta}{p}\left\|\mathcal{P}_{\Omega_l}\left(\boldsymbol{E}\right)\boldsymbol{X}^{t,(l)}\right\|_{\mathrm{F}},$$

where all nonzero entries of the matrix $\mathcal{P}_{\Omega_l}(\boldsymbol{E})$ reside in the lth row/column. Decouple the effects of the lth row and the lth column of $\mathcal{P}_{\Omega_l}(\boldsymbol{E})$ to reach

$$\frac{p}{\eta}\left\|\boldsymbol{B}_4^{(l)}\right\|_{\mathrm{F}} \le \Big\|\sum_{j=1}^{n}\underbrace{\delta_{l,j}E_{l,j}\boldsymbol{X}_{j,\cdot}^{t,(l)}}_{:=\boldsymbol{u}_j}\Big\|_2 + \underbrace{\Big\|\sum_{j:j\neq l}\delta_{l,j}E_{l,j}\boldsymbol{X}_{l,\cdot}^{t,(l)}\Big\|_2}_{:=\alpha}, \tag{154}$$

where $\delta_{l,j} := \mathbb{1}_{\{(l,j)\in\Omega\}}$ indicates whether the (l,j)th entry is observed. Since $\boldsymbol{X}^{t,(l)}$ is independent of $\{\delta_{l,j}\}_{1\le j\le n}$ and $\{E_{l,j}\}_{1\le j\le n}$, we can treat the first term as a sum of independent vectors $\{\boldsymbol{u}_j\}$. It is easy to verify that

$$\left\|\|\boldsymbol{u}_j\|_2\right\|_{\psi_1} \le \left\|\boldsymbol{X}^{t,(l)}\right\|_{2,\infty}\left\|\delta_{l,j}E_{l,j}\right\|_{\psi_1} \lesssim \sigma\left\|\boldsymbol{X}^{t,(l)}\right\|_{2,\infty},$$

where $\|\cdot\|_{\psi_1}$ denotes the sub-exponential norm [66, Section A.1]. Further, one can calculate

$$\begin{aligned} V &:= \left\|\mathbb{E}\left[\sum_{j=1}^{n}\left(\delta_{l,j}E_{l,j}\right)^2\boldsymbol{X}_{j,\cdot}^{t,(l)}\boldsymbol{X}_{j,\cdot}^{t,(l)\top}\right]\right\| \lesssim p\sigma^2\left\|\mathbb{E}\left[\sum_{j=1}^{n}\boldsymbol{X}_{j,\cdot}^{t,(l)}\boldsymbol{X}_{j,\cdot}^{t,(l)\top}\right]\right\| \\ &= p\sigma^2\left\|\boldsymbol{X}^{t,(l)}\right\|_{\mathrm{F}}^2. \end{aligned}$$

Invoke the matrix Bernstein inequality [66, Theorem 2.7] to discover that with probability at least $1-O\left(n^{-10}\right)$,

$$\begin{aligned} \Big\|\sum\nolimits_{j=1}^{n}\boldsymbol{u}_j\Big\|_2 &\lesssim \sqrt{V\log n} + \left\|\|\boldsymbol{u}_j\|_2\right\|_{\psi_1}\log^2 n \\ &\lesssim \sqrt{p\sigma^2\left\|\boldsymbol{X}^{t,(l)}\right\|_{\mathrm{F}}^2\log n} + \sigma\left\|\boldsymbol{X}^{t,(l)}\right\|_{2,\infty}\log^2 n \\ &\lesssim \sigma\sqrt{np\log n}\left\|\boldsymbol{X}^{t,(l)}\right\|_{2,\infty} + \sigma\left\|\boldsymbol{X}^{t,(l)}\right\|_{2,\infty}\log^2 n \\ &\lesssim \sigma\sqrt{np\log n}\left\|\boldsymbol{X}^{t,(l)}\right\|_{2,\infty}, \end{aligned}$$

where the third inequality follows from $\left\|\boldsymbol{X}^{t,(l)}\right\|_{\mathrm{F}}^2 \le n\left\|\boldsymbol{X}^{t,(l)}\right\|_{2,\infty}^2$ and the last inequality holds as long as $np \gg \log^2 n$.

Additionally, the remaining term α in (154) can be controlled using the same argument, giving rise to

$$\alpha \lesssim \sigma\sqrt{np\log n}\left\|\boldsymbol{X}^{t,(l)}\right\|_{2,\infty}.$$

We then complete the proof by observing that

$$\left\|\boldsymbol{X}^{t,(l)}\right\|_{2,\infty}=\left\|\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}\right\|_{2,\infty}\leq\left\|\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}-\boldsymbol{X}^{\star}\right\|_{2,\infty}+\left\|\boldsymbol{X}^{\star}\right\|_{2,\infty}\leq 2\left\|\boldsymbol{X}^{\star}\right\|_{2,\infty}, \tag{155}$$

where the last inequality follows by combining (73c), the sample complexity condition $n^2p\gg\mu^2r^2n\log n$, and the noise condition (27).

B.5.2 Proof of Lemma 25

For notational simplicity, we denote

$$\boldsymbol{C}:=\boldsymbol{X}^{t,(l)}\boldsymbol{X}^{t,(l)\top}-\boldsymbol{M}^{\star}=\boldsymbol{X}^{t,(l)}\boldsymbol{X}^{t,(l)\top}-\boldsymbol{X}^{\star}\boldsymbol{X}^{\star\top}. \tag{156}$$

Since the Frobenius norm is unitarily invariant, we have

$$\left\|\boldsymbol{B}_2^{(l)}\right\|_{\mathrm{F}}=\eta\Big\|\underbrace{\Big[\frac{1}{p}\mathcal{P}_{\Omega_l}\left(\boldsymbol{C}\right)-\mathcal{P}_l\left(\boldsymbol{C}\right)\Big]}_{:=\boldsymbol{W}}\boldsymbol{X}^{t,(l)}\Big\|_{\mathrm{F}}.$$

Again, all nonzero entries of the matrix $\boldsymbol{W}$ reside in its lth row/column. We can deal with the lth row and the lth column of $\boldsymbol{W}$ separately as follows

$$\begin{aligned}\frac{p}{\eta}\left\|\boldsymbol{B}_2^{(l)}\right\|_{\mathrm{F}}&\leq\left\|\sum_{j=1}^{n}\left(\delta_{l,j}-p\right)C_{l,j}\boldsymbol{X}_{j,\cdot}^{t,(l)}\right\|_2+\sqrt{\sum_{j:j\neq l}\left(\delta_{l,j}-p\right)^2}\left\|\boldsymbol{C}\right\|_{\infty}\left\|\boldsymbol{X}_{l,\cdot}^{t,(l)}\right\|_2\\&\lesssim\left\|\sum_{j=1}^{n}\left(\delta_{l,j}-p\right)C_{l,j}\boldsymbol{X}_{j,\cdot}^{t,(l)}\right\|_2+\sqrt{np}\left\|\boldsymbol{C}\right\|_{\infty}\left\|\boldsymbol{X}_{l,\cdot}^{t,(l)}\right\|_2,\end{aligned}$$

where $\delta_{l,j}:=\mathbb{1}_{\{(l,j)\in\Omega\}}$ and the second line relies on the fact that $\sum_{j:j\neq l}\left(\delta_{l,j}-p\right)^2\asymp np$. It follows that

$$\begin{aligned}L&:=\max_{1\leq j\leq n}\left\|\left(\delta_{l,j}-p\right)C_{l,j}\boldsymbol{X}_{j,\cdot}^{t,(l)}\right\|_2\leq\left\|\boldsymbol{C}\right\|_{\infty}\left\|\boldsymbol{X}^{t,(l)}\right\|_{2,\infty}\overset{\text{(i)}}{\leq}2\left\|\boldsymbol{C}\right\|_{\infty}\left\|\boldsymbol{X}^{\star}\right\|_{2,\infty},\\V&:=\left\|\sum_{j=1}^{n}\mathbb{E}\big[\left(\delta_{l,j}-p\right)^2\big]C_{l,j}^2\boldsymbol{X}_{j,\cdot}^{t,(l)}\boldsymbol{X}_{j,\cdot}^{t,(l)\top}\right\|\leq p\|\boldsymbol{C}\|_{\infty}^2\left\|\sum_{j=1}^{n}\boldsymbol{X}_{j,\cdot}^{t,(l)}\boldsymbol{X}_{j,\cdot}^{t,(l)\top}\right\|\\&=p\left\|\boldsymbol{C}\right\|_{\infty}^2\left\|\boldsymbol{X}^{t,(l)}\right\|_{\mathrm{F}}^2\overset{\text{(ii)}}{\leq}4p\left\|\boldsymbol{C}\right\|_{\infty}^2\left\|\boldsymbol{X}^{\star}\right\|_{\mathrm{F}}^2.\end{aligned}$$

Here, (i) is a consequence of (155). In addition, (ii) follows from

$$\left\|\boldsymbol{X}^{t,(l)}\right\|_{\mathrm{F}}=\left\|\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}\right\|_{\mathrm{F}}\leq\left\|\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}-\boldsymbol{X}^{\star}\right\|_{\mathrm{F}}+\left\|\boldsymbol{X}^{\star}\right\|_{\mathrm{F}}\leq 2\left\|\boldsymbol{X}^{\star}\right\|_{\mathrm{F}},$$

where the last inequality comes from (73b), the sample complexity condition $n^2 p \gg \mu^2 r^2 n \log n$, and the noise condition (27). The matrix Bernstein inequality [114, Theorem 6.1.1] reveals that

$$\begin{aligned}\left\| \sum_{j=1}^{n} \left(\delta_{l,j} - p\right) C_{l,j} \boldsymbol{X}^{t,(l)}_{j,\cdot} \right\|_2 &\lesssim \sqrt{V \log n} + L \log n \\ &\lesssim \sqrt{p \left\|\boldsymbol{C}\right\|_\infty^2 \left\|\boldsymbol{X}^\star\right\|_{\mathrm{F}}^2 \log n} + \left\|\boldsymbol{C}\right\|_\infty \left\|\boldsymbol{X}^\star\right\|_{2,\infty} \log n\end{aligned}$$

with probability exceeding $1 - O\left(n^{-10}\right)$, and as a result,

$$\frac{p}{\eta} \left\|\boldsymbol{B}_2^{(l)}\right\|_{\mathrm{F}} \lesssim \sqrt{p \log n} \left\|\boldsymbol{C}\right\|_\infty \left\|\boldsymbol{X}^\star\right\|_{\mathrm{F}} + \sqrt{np} \left\|\boldsymbol{C}\right\|_\infty \left\|\boldsymbol{X}^\star\right\|_{2,\infty} \tag{157}$$

as soon as $np \gg \log n$.

To finish up, we make the observation that

$$\begin{aligned}\left\|\boldsymbol{C}\right\|_\infty &= \left\| \boldsymbol{X}^{t,(l)} \boldsymbol{R}^{t,(l)} \left(\boldsymbol{X}^{t,(l)} \boldsymbol{R}^{t,(l)}\right)^\top - \boldsymbol{X}^\star \boldsymbol{X}^{\star\top} \right\|_\infty \\ &\le \left\| \left(\boldsymbol{X}^{t,(l)} \boldsymbol{R}^{t,(l)} - \boldsymbol{X}^\star\right) \left(\boldsymbol{X}^{t,(l)} \boldsymbol{R}^{t,(l)}\right)^\top \right\|_\infty + \left\| \boldsymbol{X}^\star \left(\boldsymbol{X}^{t,(l)} \boldsymbol{R}^{t,(l)} - \boldsymbol{X}^\star\right)^\top - \boldsymbol{X}^\star \boldsymbol{X}^{\star\top} \right\|_\infty \\ &\le \left\| \boldsymbol{X}^{t,(l)} \boldsymbol{R}^{t,(l)} - \boldsymbol{X}^\star \right\|_{2,\infty} \left\| \boldsymbol{X}^{t,(l)} \boldsymbol{R}^{t,(l)} \right\|_{2,\infty} + \left\|\boldsymbol{X}^\star\right\|_{2,\infty} \left\| \boldsymbol{X}^{t,(l)} \boldsymbol{R}^{t,(l)} - \boldsymbol{X}^\star \right\|_{2,\infty} \\ &\le 3 \left\| \boldsymbol{X}^{t,(l)} \boldsymbol{R}^{t,(l)} - \boldsymbol{X}^\star \right\|_{2,\infty} \left\|\boldsymbol{X}^\star\right\|_{2,\infty}, \end{aligned} \tag{158}$$

where the last line arises from (155). This combined with (157) gives

$$\begin{aligned}\left\|\boldsymbol{B}_2^{(l)}\right\|_{\mathrm{F}} &\lesssim \eta \sqrt{\frac{\log n}{p}} \left\|\boldsymbol{C}\right\|_\infty \left\|\boldsymbol{X}^\star\right\|_{\mathrm{F}} + \eta \sqrt{\frac{n}{p}} \left\|\boldsymbol{C}\right\|_\infty \left\|\boldsymbol{X}^\star\right\|_{2,\infty} \\ &\overset{\text{(i)}}{\lesssim} \eta \sqrt{\frac{\log n}{p}} \left\| \boldsymbol{X}^{t,(l)} \boldsymbol{R}^{t,(l)} - \boldsymbol{X}^\star \right\|_{2,\infty} \left\|\boldsymbol{X}^\star\right\|_{2,\infty} \left\|\boldsymbol{X}^\star\right\|_{\mathrm{F}} \\ &\quad + \eta \sqrt{\frac{n}{p}} \left\| \boldsymbol{X}^{t,(l)} \boldsymbol{R}^{t,(l)} - \boldsymbol{X}^\star \right\|_{2,\infty} \left\|\boldsymbol{X}^\star\right\|_{2,\infty}^2 \\ &\overset{\text{(ii)}}{\lesssim} \eta \sqrt{\frac{\log n}{p}} \left\| \boldsymbol{X}^{t,(l)} \boldsymbol{R}^{t,(l)} - \boldsymbol{X}^\star \right\|_{2,\infty} \sqrt{\frac{\kappa \mu r^2}{n}} \sigma_{\max} \\ &\quad + \eta \sqrt{\frac{n}{p}} \left\| \boldsymbol{X}^{t,(l)} \boldsymbol{R}^{t,(l)} - \boldsymbol{X}^\star \right\|_{2,\infty} \frac{\kappa \mu r}{n} \sigma_{\max} \\ &\lesssim \eta \sqrt{\frac{\kappa^2 \mu^2 r^2 \log n}{np}} \left\| \boldsymbol{X}^{t,(l)} \boldsymbol{R}^{t,(l)} - \boldsymbol{X}^\star \right\|_{2,\infty} \sigma_{\max}, \end{aligned}$$

where (i) comes from (158) and (ii) makes use of the incoherence condition (114).

B.6 Proof of Lemma 12

We first introduce an auxiliary matrix

$$\widetilde{\boldsymbol{X}}^{t+1,(l)} := \boldsymbol{X}^{t,(l)}\widehat{\boldsymbol{H}}^{t,(l)} - \eta\left[\frac{1}{p}\mathcal{P}_{\Omega^{-l}}\left[\boldsymbol{X}^{t,(l)}\boldsymbol{X}^{t,(l)\top} - \left(\boldsymbol{M}^\star + \boldsymbol{E}\right)\right] + \mathcal{P}_l\left(\boldsymbol{X}^{t,(l)}\boldsymbol{X}^{t,(l)\top} - \boldsymbol{M}^\star\right)\right]\boldsymbol{X}^\star. \tag{159}$$

With this in place, we can use the triangle inequality to obtain

$$\left\|\left(\boldsymbol{X}^{t+1,(l)}\widehat{\boldsymbol{H}}^{t+1,(l)} - \boldsymbol{X}^\star\right)_{l,\cdot}\right\|_2 \le \underbrace{\left\|\left(\boldsymbol{X}^{t+1,(l)}\widehat{\boldsymbol{H}}^{t+1,(l)} - \widetilde{\boldsymbol{X}}^{t+1,(l)}\right)_{l,\cdot}\right\|_2}_{:=\alpha_1} + \underbrace{\left\|\left(\widetilde{\boldsymbol{X}}^{t+1,(l)} - \boldsymbol{X}^\star\right)_{l,\cdot}\right\|_2}_{:=\alpha_2}. \tag{160}$$

In what follows, we bound the two terms α_1 and α_2 separately.

1. Regarding the second term α_2 of (160), we see from the definition of $\widetilde{\boldsymbol{X}}^{t+1,(l)}$ (see (159)) that

$$\left(\widetilde{\boldsymbol{X}}^{t+1,(l)} - \boldsymbol{X}^\star\right)_{l,\cdot} = \left[\boldsymbol{X}^{t,(l)}\widehat{\boldsymbol{H}}^{t,(l)} - \eta\left(\boldsymbol{X}^{t,(l)}\boldsymbol{X}^{t,(l)\top} - \boldsymbol{X}^\star\boldsymbol{X}^{\star\top}\right)\boldsymbol{X}^\star - \boldsymbol{X}^\star\right]_{l,\cdot}, \tag{161}$$

where we also utilize the definitions of $\mathcal{P}_{\Omega^{-l}}$ and $\mathcal{P}_l$ in (67). For notational convenience, we denote

$$\boldsymbol{\Delta}^{t,(l)} := \boldsymbol{X}^{t,(l)}\widehat{\boldsymbol{H}}^{t,(l)} - \boldsymbol{X}^\star. \tag{162}$$

This allows us to rewrite (161) as

$$\begin{aligned}\left(\widetilde{\boldsymbol{X}}^{t+1,(l)} - \boldsymbol{X}^\star\right)_{l,\cdot} &= \boldsymbol{\Delta}^{t,(l)}_{l,\cdot} - \eta\left[\left(\boldsymbol{\Delta}^{t,(l)}\boldsymbol{X}^{\star\top} + \boldsymbol{X}^\star\boldsymbol{\Delta}^{t,(l)\top}\right)\boldsymbol{X}^\star\right]_{l,\cdot} \\ &\quad - \eta\left[\boldsymbol{\Delta}^{t,(l)}\boldsymbol{\Delta}^{t,(l)\top}\boldsymbol{X}^\star\right]_{l,\cdot} \\ &= \boldsymbol{\Delta}^{t,(l)}_{l,\cdot} - \eta\boldsymbol{\Delta}^{t,(l)}_{l,\cdot}\boldsymbol{\Sigma}^\star - \eta\boldsymbol{X}^\star_{l,\cdot}\boldsymbol{\Delta}^{t,(l)\top}\boldsymbol{X}^\star - \eta\boldsymbol{\Delta}^{t,(l)}_{l,\cdot}\boldsymbol{\Delta}^{t,(l)\top}\boldsymbol{X}^\star,\end{aligned}$$

which further implies that

$$\begin{aligned}\alpha_2 &\le \left\|\boldsymbol{\Delta}^{t,(l)}_{l,\cdot} - \eta\boldsymbol{\Delta}^{t,(l)}_{l,\cdot}\boldsymbol{\Sigma}^\star\right\|_2 + \eta\left\|\boldsymbol{X}^\star_{l,\cdot}\boldsymbol{\Delta}^{t,(l)\top}\boldsymbol{X}^\star\right\|_2 + \eta\left\|\boldsymbol{\Delta}^{t,(l)}_{l,\cdot}\boldsymbol{\Delta}^{t,(l)\top}\boldsymbol{X}^\star\right\|_2 \\ &\le \left\|\boldsymbol{\Delta}^{t,(l)}_{l,\cdot}\right\|_2\left\|\boldsymbol{I}_r - \eta\boldsymbol{\Sigma}^\star\right\| + \eta\left\|\boldsymbol{X}^\star\right\|_{2,\infty}\left\|\boldsymbol{\Delta}^{t,(l)}\right\|\left\|\boldsymbol{X}^\star\right\| + \eta\left\|\boldsymbol{\Delta}^{t,(l)}_{l,\cdot}\right\|_2\left\|\boldsymbol{\Delta}^{t,(l)}\right\|\left\|\boldsymbol{X}^\star\right\| \\ &\le \left\|\boldsymbol{\Delta}^{t,(l)}_{l,\cdot}\right\|_2\left\|\boldsymbol{I}_r - \eta\boldsymbol{\Sigma}^\star\right\| + 2\eta\left\|\boldsymbol{X}^\star\right\|_{2,\infty}\left\|\boldsymbol{\Delta}^{t,(l)}\right\|\left\|\boldsymbol{X}^\star\right\|.\end{aligned}$$

Here, the last line follows from the fact that $\left\|\boldsymbol{\Delta}_{l,\cdot}^{t,(l)}\right\|_2 \le \left\|\boldsymbol{X}^\star\right\|_{2,\infty}$. To see this, one can use the induction hypothesis (70e) to get

$$\left\|\boldsymbol{\Delta}_{l,\cdot}^{t,(l)}\right\|_2 \le C_2\rho^t\mu r\frac{1}{\sqrt{np}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty} + C_6\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty} \ll \left\|\boldsymbol{X}^\star\right\|_{2,\infty} \tag{163}$$

as long as $np \gg \mu^2 r^2$ and $\sigma\sqrt{(n\log n)/p} \ll \sigma_{\min}$. By taking $0 < \eta \le 1/\sigma_{\max}$, we have $\boldsymbol{0} \preceq \boldsymbol{I}_r - \eta\boldsymbol{\Sigma}^\star \preceq (1-\eta\sigma_{\min})\boldsymbol{I}_r$ and hence can obtain

$$\alpha_2 \le (1-\eta\sigma_{\min})\left\|\boldsymbol{\Delta}_{l,\cdot}^{t,(l)}\right\|_2 + 2\eta\left\|\boldsymbol{X}^\star\right\|_{2,\infty}\left\|\boldsymbol{\Delta}^{t,(l)}\right\|\left\|\boldsymbol{X}^\star\right\|. \tag{164}$$

An immediate consequence of the above two inequalities and (73d) is

$$\alpha_2 \le \|\boldsymbol{X}^\star\|_{2,\infty}. \tag{165}$$

2. The first term α_1 of (160) can be equivalently written as

$$\alpha_1 = \left\|\left(\boldsymbol{X}^{t+1,(l)}\widehat{\boldsymbol{H}}^{t,(l)}\boldsymbol{R}_1 - \widetilde{\boldsymbol{X}}^{t+1,(l)}\right)_{l,\cdot}\right\|_2,$$

where

$$\boldsymbol{R}_1 = \left(\widehat{\boldsymbol{H}}^{t,(l)}\right)^{-1}\widehat{\boldsymbol{H}}^{t+1,(l)} := \arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}}\left\|\boldsymbol{X}^{t+1,(l)}\widehat{\boldsymbol{H}}^{t,(l)}\boldsymbol{R} - \boldsymbol{X}^\star\right\|_{\mathrm{F}}.$$

Simple algebra yields

$$\begin{aligned}\alpha_1 &\le \left\|\left(\boldsymbol{X}^{t+1,(l)}\widehat{\boldsymbol{H}}^{t,(l)} - \widetilde{\boldsymbol{X}}^{t+1,(l)}\right)_{l,\cdot}\boldsymbol{R}_1\right\|_2 + \left\|\widetilde{\boldsymbol{X}}_{l,\cdot}^{t+1,(l)}\right\|_2\|\boldsymbol{R}_1 - \boldsymbol{I}_r\| \\ &\le \underbrace{\left\|\left(\boldsymbol{X}^{t+1,(l)}\widehat{\boldsymbol{H}}^{t,(l)} - \widetilde{\boldsymbol{X}}^{t+1,(l)}\right)_{l,\cdot}\right\|_2}_{:=\beta_1} + 2\left\|\boldsymbol{X}^\star\right\|_{2,\infty}\underbrace{\|\boldsymbol{R}_1 - \boldsymbol{I}_r\|}_{:=\beta_2}.\end{aligned}$$

Here, to bound the second term, we have used

$$\left\|\widetilde{\boldsymbol{X}}_{l,\cdot}^{t+1,(l)}\right\|_2 \le \left\|\widetilde{\boldsymbol{X}}_{l,\cdot}^{t+1,(l)} - \boldsymbol{X}_{l,\cdot}^\star\right\|_2 + \left\|\boldsymbol{X}_{l,\cdot}^\star\right\|_2 = \alpha_2 + \left\|\boldsymbol{X}_{l,\cdot}^\star\right\|_2 \le 2\left\|\boldsymbol{X}^\star\right\|_{2,\infty},$$

where the last inequality follows from (165). It remains to upper bound β_1 and β_2. For both β_1 and β_2, a central quantity to control is $\boldsymbol{X}^{t+1,(l)}\widehat{\boldsymbol{H}}^{t,(l)} - \widetilde{\boldsymbol{X}}^{t+1,(l)}$. By the definition of $\widetilde{\boldsymbol{X}}^{t+1,(l)}$ in (159) and the gradient update rule for $\boldsymbol{X}^{t+1,(l)}$ (see (69)), one has

$$\begin{aligned}&\boldsymbol{X}^{t+1,(l)}\widehat{\boldsymbol{H}}^{t,(l)} - \widetilde{\boldsymbol{X}}^{t+1,(l)} \\ &= \left\{\boldsymbol{X}^{t,(l)}\widehat{\boldsymbol{H}}^{t,(l)} - \eta\left[\frac{1}{p}\mathcal{P}_{\Omega^{-l}}\left[\boldsymbol{X}^{t,(l)}\boldsymbol{X}^{t,(l)\top} - (\boldsymbol{M}^\star + \boldsymbol{E})\right]\right.\right.\end{aligned}$$

$$
\begin{aligned}
&+\mathcal{P}_l\left(\boldsymbol{X}^{t,(l)}\boldsymbol{X}^{t,(l)\top}-\boldsymbol{M}^\star\right)\Big]\boldsymbol{X}^{t,(l)}\widehat{\boldsymbol{H}}^{t,(l)}\Big\}\\
&-\left\{\boldsymbol{X}^{t,(l)}\widehat{\boldsymbol{H}}^{t,(l)}-\eta\left[\frac{1}{p}\mathcal{P}_{\Omega^{-l}}\left[\boldsymbol{X}^{t,(l)}\boldsymbol{X}^{t,(l)\top}-(\boldsymbol{M}^\star+\boldsymbol{E})\right]+\mathcal{P}_l\left(\boldsymbol{X}^{t,(l)}\boldsymbol{X}^{t,(l)\top}-\boldsymbol{M}^\star\right)\right]\boldsymbol{X}^\star\right\}\\
&=-\eta\left[\frac{1}{p}\mathcal{P}_{\Omega^{-l}}\left(\boldsymbol{X}^{t,(l)}\boldsymbol{X}^{t,(l)\top}-\boldsymbol{X}^\star\boldsymbol{X}^{\star\top}\right)+\mathcal{P}_l\left(\boldsymbol{X}^{t,(l)}\boldsymbol{X}^{t,(l)\top}-\boldsymbol{X}^\star\boldsymbol{X}^{\star\top}\right)\right]\boldsymbol{\Delta}^{t,(l)}\\
&+\frac{\eta}{p}\mathcal{P}_{\Omega^{-l}}(\boldsymbol{E})\,\boldsymbol{\Delta}^{t,(l)}. \qquad (166)
\end{aligned}
$$

It is easy to verify that

$$
\frac{1}{p}\left\|\mathcal{P}_{\Omega^{-l}}(\boldsymbol{E})\right\|\overset{\text{(i)}}{\leq}\frac{1}{p}\left\|\mathcal{P}_{\Omega}(\boldsymbol{E})\right\|\overset{\text{(ii)}}{\lesssim}\sigma\sqrt{\frac{n}{p}}\overset{\text{(iii)}}{\leq}\frac{\delta}{2}\sigma_{\min}
$$

for $\delta > 0$ sufficiently small. Here, (i) uses the elementary fact that the spectral norm of a submatrix is no more than that of the matrix itself, (ii) arises from Lemma 40, and (iii) is a consequence of the noise condition (27). Therefore, in order to control (166), we need to upper bound the following quantity

$$
\gamma:=\left\|\frac{1}{p}\mathcal{P}_{\Omega^{-l}}\left(\boldsymbol{X}^{t,(l)}\boldsymbol{X}^{t,(l)\top}-\boldsymbol{X}^\star\boldsymbol{X}^{\star\top}\right)+\mathcal{P}_l\left(\boldsymbol{X}^{t,(l)}\boldsymbol{X}^{t,(l)\top}-\boldsymbol{X}^\star\boldsymbol{X}^{\star\top}\right)\right\|. \qquad (167)
$$

To this end, we make the observation that

$$
\begin{aligned}
\gamma\leq&\underbrace{\left\|\frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{X}^{t,(l)}\boldsymbol{X}^{t,(l)\top}-\boldsymbol{X}^\star\boldsymbol{X}^{\star\top}\right)\right\|}_{:=\gamma_1}\\
&+\underbrace{\left\|\frac{1}{p}\mathcal{P}_{\Omega_l}\left(\boldsymbol{X}^{t,(l)}\boldsymbol{X}^{t,(l)\top}-\boldsymbol{X}^\star\boldsymbol{X}^{\star\top}\right)-\mathcal{P}_l\left(\boldsymbol{X}^{t,(l)}\boldsymbol{X}^{t,(l)\top}-\boldsymbol{X}^\star\boldsymbol{X}^{\star\top}\right)\right\|}_{:=\gamma_2}, \qquad (168)
\end{aligned}
$$

where $\mathcal{P}_{\Omega_l}$ is defined in (66). An application of Lemma 43 reveals that

$$
\gamma_1\leq 2n\left\|\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}-\boldsymbol{X}^\star\right\|_{2,\infty}^2+4\sqrt{n}\log n\left\|\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}-\boldsymbol{X}^\star\right\|_{2,\infty}\left\|\boldsymbol{X}^\star\right\|,
$$

where $\boldsymbol{R}^{t,(l)}\in\mathcal{O}^{r\times r}$ is defined in (72). Let $\boldsymbol{C}=\boldsymbol{X}^{t,(l)}\boldsymbol{X}^{t,(l)\top}-\boldsymbol{X}^\star\boldsymbol{X}^{\star\top}$ as in (156), and one can bound the other term γ_2 by taking advantage of the triangle inequality and the symmetry property:

$$
\gamma_2\leq\frac{2}{p}\sqrt{\sum_{j=1}^{n}\left(\delta_{l,j}-p\right)^2C_{l,j}^2}\overset{\text{(i)}}{\lesssim}\sqrt{\frac{n}{p}}\left\|\boldsymbol{C}\right\|_\infty\overset{\text{(ii)}}{\lesssim}\sqrt{\frac{n}{p}}\left\|\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}-\boldsymbol{X}^\star\right\|_{2,\infty}\left\|\boldsymbol{X}^\star\right\|_{2,\infty},
$$

where (i) comes from the standard Chernoff bound $\sum_{j=1}^{n}\left(\delta_{l,j}-p\right)^2 \asymp np$, and in (ii) we utilize the bound established in (158). The previous two bounds taken collectively give

$$\begin{aligned}\gamma &\le 2n\left\|\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}-\boldsymbol{X}^{\star}\right\|_{2,\infty}^{2}+4\sqrt{n}\log n\left\|\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}-\boldsymbol{X}^{\star}\right\|_{2,\infty}\left\|\boldsymbol{X}^{\star}\right\| \\ &\quad+\widetilde{C}\sqrt{\frac{n}{p}}\left\|\boldsymbol{X}^{t,(l)}\boldsymbol{R}^{t,(l)}-\boldsymbol{X}^{\star}\right\|_{2,\infty}\left\|\boldsymbol{X}^{\star}\right\|_{2,\infty}\le\frac{\delta}{2}\sigma_{\min}\end{aligned} \tag{169}$$

for some constant $\widetilde{C}>0$ and $\delta>0$ sufficiently small. The last inequality follows from (73c), the incoherence condition (114), and our sample size condition. In summary, we obtain

$$\left\|\boldsymbol{X}^{t+1,(l)}\widehat{\boldsymbol{H}}^{t,(l)}-\widetilde{\boldsymbol{X}}^{t+1,(l)}\right\|\le\eta\left(\gamma+\left\|\frac{1}{p}\mathcal{P}_{\Omega^{-l}}(\boldsymbol{E})\right\|\right)\left\|\boldsymbol{\Delta}^{t,(l)}\right\|\le\eta\delta\sigma_{\min}\left\|\boldsymbol{\Delta}^{t,(l)}\right\|, \tag{170}$$

for $\delta>0$ sufficiently small. With the estimate (170) in place, we can continue our derivation on β_1 and β_2.

(a) With regard to β_1, in view of (166) we can obtain

$$\begin{aligned}\beta_1 &\overset{\text{(i)}}{=}\eta\left\|\left(\boldsymbol{X}^{t,(l)}\boldsymbol{X}^{t,(l)\top}-\boldsymbol{X}^{\star}\boldsymbol{X}^{\star\top}\right)_{l,\cdot}\boldsymbol{\Delta}^{t,(l)}\right\|_2 \\ &\le\eta\left\|\left(\boldsymbol{X}^{t,(l)}\boldsymbol{X}^{t,(l)\top}-\boldsymbol{X}^{\star}\boldsymbol{X}^{\star\top}\right)_{l,\cdot}\right\|_2\left\|\boldsymbol{\Delta}^{t,(l)}\right\| \\ &\overset{\text{(ii)}}{=}\eta\left\|\left[\boldsymbol{\Delta}^{t,(l)}\left(\boldsymbol{X}^{t,(l)}\widehat{\boldsymbol{H}}^{t,(l)}\right)^{\top}+\boldsymbol{X}^{\star}\boldsymbol{\Delta}^{t,(l)\top}\right]_{l,\cdot}\right\|_2\left\|\boldsymbol{\Delta}^{t,(l)}\right\| \\ &\le\eta\left(\left\|\boldsymbol{\Delta}_{l,\cdot}^{t,(l)}\right\|_2\left\|\boldsymbol{X}^{t,(l)}\right\|+\left\|\boldsymbol{X}_{l,\cdot}^{\star}\right\|_2\left\|\boldsymbol{\Delta}^{t,(l)}\right\|\right)\left\|\boldsymbol{\Delta}^{t,(l)}\right\| \\ &\le\eta\left\|\boldsymbol{\Delta}_{l,\cdot}^{t,(l)}\right\|_2\left\|\boldsymbol{X}^{t,(l)}\right\|\left\|\boldsymbol{\Delta}^{t,(l)}\right\|+\eta\left\|\boldsymbol{X}_{l,\cdot}^{\star}\right\|_2\left\|\boldsymbol{\Delta}^{t,(l)}\right\|^2,\end{aligned} \tag{171}$$

where (i) follows from the definitions of $\mathcal{P}_{\Omega^{-l}}$ and $\mathcal{P}_l$ (see (67) and note that all entries in the lth row of $\mathcal{P}_{\Omega^{-l}}(\cdot)$ are identically zero) and identity (ii) is due to the definition of $\boldsymbol{\Delta}^{t,(l)}$ in (162).

(b) For β_2, we first claim that

$$\boldsymbol{I}_r:=\arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}}\left\|\widetilde{\boldsymbol{X}}^{t+1,(l)}\boldsymbol{R}-\boldsymbol{X}^{\star}\right\|_{\mathrm{F}}, \tag{172}$$

whose justification follows similar reasonings as that of (138) and is therefore omitted. In particular, it gives rise to the facts that $\boldsymbol{X}^{\star\top}\widetilde{\boldsymbol{X}}^{t+1,(l)}$ is symmetric and

$$\left(\widetilde{\boldsymbol{X}}^{t+1,(l)}\right)^{\top}\boldsymbol{X}^{\star}\succeq\frac{1}{2}\sigma_{\min}\boldsymbol{I}_r. \tag{173}$$

We are now ready to invoke Lemma 36 to bound β_2. We abuse the notation and denote $\boldsymbol{C} := \left(\widetilde{\boldsymbol{X}}^{t+1,(l)}\right)^\top \boldsymbol{X}^\star$ and $\boldsymbol{E} := \left(\boldsymbol{X}^{t+1,(l)}\widehat{\boldsymbol{H}}^{t,(l)} - \widetilde{\boldsymbol{X}}^{t+1,(l)}\right)^\top \boldsymbol{X}^\star$. We have

$$\|\boldsymbol{E}\| \le \frac{1}{2}\sigma_{\min} \le \sigma_r(\boldsymbol{C}).$$

The first inequality arises from (170), namely

$$\begin{aligned}\|\boldsymbol{E}\| &\le \left\|\boldsymbol{X}^{t+1,(l)}\widehat{\boldsymbol{H}}^{t,(l)} - \widetilde{\boldsymbol{X}}^{t+1,(l)}\right\| \left\|\boldsymbol{X}^\star\right\| \le \eta\delta\sigma_{\min}\left\|\boldsymbol{\Delta}^{t,(l)}\right\| \left\|\boldsymbol{X}^\star\right\| \\ &\overset{\text{(i)}}{\le} \eta\delta\sigma_{\min}\left\|\boldsymbol{X}^\star\right\|^2 \overset{\text{(ii)}}{\le} \frac{1}{2}\sigma_{\min},\end{aligned}$$

where (i) holds since $\left\|\boldsymbol{\Delta}^{t,(l)}\right\| \le \left\|\boldsymbol{X}^\star\right\|$ and (ii) holds true for δ sufficiently small and $\eta \le 1/\sigma_{\max}$. Invoke Lemma 36 to obtain

$$\begin{aligned}\beta_2 = \|\boldsymbol{R}_1 - \boldsymbol{I}_r\| &\le \frac{2}{\sigma_{r-1}(\boldsymbol{C}) + \sigma_r(\boldsymbol{C})}\|\boldsymbol{E}\| \\ &\le \frac{2}{\sigma_{\min}}\left\|\boldsymbol{X}^{t+1,(l)}\widehat{\boldsymbol{H}}^{t,(l)} - \widetilde{\boldsymbol{X}}^{t+1,(l)}\right\| \left\|\boldsymbol{X}^\star\right\| \qquad (174) \\ &\le 2\delta\eta\left\|\boldsymbol{\Delta}^{t,(l)}\right\| \left\|\boldsymbol{X}^\star\right\|, \qquad (175)\end{aligned}$$

where (174) follows since $\sigma_{r-1}(\boldsymbol{C}) \ge \sigma_r(\boldsymbol{C}) \ge \sigma_{\min}/2$ from (173), and the last line comes from (170).

(c) Putting the previous bounds (171) and (175) together yields

$$\begin{aligned}\alpha_1 \le{}& \eta\left\|\boldsymbol{\Delta}^{t,(l)}_{l,\cdot}\right\|_2 \left\|\boldsymbol{X}^{t,(l)}\right\| \left\|\boldsymbol{\Delta}^{t,(l)}\right\| + \eta\left\|\boldsymbol{X}^\star_{l,\cdot}\right\|_2 \left\|\boldsymbol{\Delta}^{t,(l)}\right\|^2 \\ &+ 4\delta\eta\left\|\boldsymbol{X}^\star\right\|_{2,\infty}\left\|\boldsymbol{\Delta}^{t,(l)}\right\| \left\|\boldsymbol{X}^\star\right\|. \qquad (176)\end{aligned}$$

3. Combine (160), (164), and (176) to reach

$$\begin{aligned}&\left\|\left(\boldsymbol{X}^{t+1,(l)}\widehat{\boldsymbol{H}}^{t+1,(l)} - \boldsymbol{X}^\star\right)_{l,\cdot}\right\|_2 \le (1-\eta\sigma_{\min})\left\|\boldsymbol{\Delta}^{t,(l)}_{l,\cdot}\right\|_2 + 2\eta\left\|\boldsymbol{X}^\star\right\|_{2,\infty}\left\|\boldsymbol{\Delta}^{t,(l)}\right\| \left\|\boldsymbol{X}^\star\right\| \\ &\quad + \eta\left\|\boldsymbol{\Delta}^{t,(l)}_{l,\cdot}\right\|_2 \left\|\boldsymbol{X}^{t,(l)}\right\| \left\|\boldsymbol{\Delta}^{t,(l)}\right\| + \eta\left\|\boldsymbol{X}^\star_{l,\cdot}\right\|_2 \left\|\boldsymbol{\Delta}^{t,(l)}\right\|^2 + 4\delta\eta\left\|\boldsymbol{X}^\star\right\|_{2,\infty}\left\|\boldsymbol{\Delta}^{t,(l)}\right\| \left\|\boldsymbol{X}^\star\right\| \\ &\overset{\text{(i)}}{\le} \left(1 - \eta\sigma_{\min} + \eta\left\|\boldsymbol{X}^{t,(l)}\right\| \left\|\boldsymbol{\Delta}^{t,(l)}\right\|\right)\left\|\boldsymbol{\Delta}^{t,(l)}_{l,\cdot}\right\|_2 + 4\eta\left\|\boldsymbol{X}^\star\right\|_{2,\infty}\left\|\boldsymbol{\Delta}^{t,(l)}\right\| \left\|\boldsymbol{X}^\star\right\| \\ &\overset{\text{(ii)}}{\le} \left(1 - \frac{\sigma_{\min}}{2}\eta\right)\left(C_2\rho^t\mu r\frac{1}{\sqrt{np}} + \frac{C_6}{\sigma_{\min}}\sigma\sqrt{\frac{n\log n}{p}}\right)\left\|\boldsymbol{X}^\star\right\|_{2,\infty} \\ &\quad + 4\eta\left\|\boldsymbol{X}^\star\right\| \left\|\boldsymbol{X}^\star\right\|_{2,\infty}\left(2C_9\rho^t\mu r\frac{1}{\sqrt{np}}\left\|\boldsymbol{X}^\star\right\| + \frac{2C_{10}}{\sigma_{\min}}\sigma\sqrt{\frac{n}{p}}\left\|\boldsymbol{X}^\star\right\|\right) \\ &\overset{\text{(iii)}}{\le} C_2\rho^{t+1}\mu r\frac{1}{\sqrt{np}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty} + \frac{C_6}{\sigma_{\min}}\sigma\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty}.\end{aligned}$$

Here, (i) follows since $\left\|\boldsymbol{\Delta}^{t,(l)}\right\| \le \left\|\boldsymbol{X}^{\star}\right\|$ and δ is sufficiently small, (ii) invokes hypotheses (70e) and (73d) and recognizes that

$$\left\|\boldsymbol{X}^{t,(l)}\right\|\left\|\boldsymbol{\Delta}^{t,(l)}\right\| \le 2\left\|\boldsymbol{X}^{\star}\right\|\left(2C_9\mu r\frac{1}{\sqrt{np}}\left\|\boldsymbol{X}^{\star}\right\| + \frac{2C_{10}}{\sigma_{\min}}\sigma\sqrt{\frac{n\log n}{np}}\left\|\boldsymbol{X}^{\star}\right\|\right) \le \frac{\sigma_{\min}}{2}$$

holds under the sample size and noise condition, while (iii) is valid as long as $1-(\sigma_{\min}/3)\cdot\eta \le \rho < 1$, $C_2 \gg \kappa C_9$, and $C_6 \gg \kappa C_{10}/\sqrt{\log n}$.

B.7 Proof of Lemma 13

For notational convenience, we define the following two orthonormal matrices

$$\boldsymbol{Q} := \arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}}\left\|\boldsymbol{U}^0\boldsymbol{R}-\boldsymbol{U}^{\star}\right\|_{\mathrm{F}} \qquad \text{and} \qquad \boldsymbol{Q}^{(l)} := \arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}}\left\|\boldsymbol{U}^{0,(l)}\boldsymbol{R}-\boldsymbol{U}^{\star}\right\|_{\mathrm{F}}.$$

The problem of finding $\widehat{\boldsymbol{H}}^t$ (see (26)) is called the *orthogonal Procrustes problem* [112]. It is well known that the minimizer $\widehat{\boldsymbol{H}}^t$ always exists and is given by

$$\widehat{\boldsymbol{H}}^t = \operatorname{sgn}\left(\boldsymbol{X}^{t\top}\boldsymbol{X}^{\star}\right).$$

Here, the sign matrix $\operatorname{sgn}(\boldsymbol{B})$ is defined as

$$\operatorname{sgn}(\boldsymbol{B}) := \boldsymbol{U}\boldsymbol{V}^{\top} \tag{177}$$

for any matrix $\boldsymbol{B}$ with singular value decomposition $\boldsymbol{B}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\top}$, where the columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ are left and right singular vectors, respectively.

Before proceeding, we make note of the following perturbation bounds on $\boldsymbol{M}^0$ and $\boldsymbol{M}^{(l)}$ (as defined in Algorithms 2 and 5, respectively):

$$\begin{aligned}\left\|\boldsymbol{M}^0-\boldsymbol{M}^{\star}\right\| &\overset{\text{(i)}}{\le} \left\|\frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{M}^{\star}\right)-\boldsymbol{M}^{\star}\right\| + \left\|\frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{E}\right)\right\| \\ &\overset{\text{(ii)}}{\le} C\sqrt{\frac{n}{p}}\left\|\boldsymbol{M}^{\star}\right\|_{\infty} + C\sigma\sqrt{\frac{n}{p}} = C\sqrt{\frac{n}{p}}\left\|\boldsymbol{X}^{\star}\right\|_{2,\infty}^2 + C\frac{\sigma}{\sqrt{\sigma_{\min}}}\sqrt{\frac{n}{p}}\sqrt{\sigma_{\min}} \\ &\overset{\text{(iii)}}{\le} C\left\{\mu r\sqrt{\frac{1}{np}}\sqrt{\sigma_{\max}} + \frac{\sigma}{\sqrt{\sigma_{\min}}}\sqrt{\frac{n}{p}}\right\}\left\|\boldsymbol{X}^{\star}\right\| \overset{\text{(iv)}}{\ll} \sigma_{\min},\end{aligned} \tag{178}$$

for some universal constant $C>0$. Here, (i) arises from the triangle inequality, (ii) utilizes Lemmas 39 and 40, (iii) follows from the incoherence condition (114), and (iv) holds under our sample complexity assumption that $n^2p \gg \mu^2r^2n$ and the noise

condition (27). Similarly, we have

$$\left\|\boldsymbol{M}^{(l)}-\boldsymbol{M}^{\star}\right\| \lesssim \left\{\mu r\sqrt{\frac{1}{np}}\sqrt{\sigma_{\max}}+\frac{\sigma}{\sqrt{\sigma_{\min}}}\sqrt{\frac{n}{p}}\right\}\left\|\boldsymbol{X}^{\star}\right\| \ll \sigma_{\min}. \tag{179}$$

Combine Weyl's inequality, (178), and (179) to obtain

$$\left\|\boldsymbol{\Sigma}^{0}-\boldsymbol{\Sigma}^{\star}\right\| \leq \left\|\boldsymbol{M}^{0}-\boldsymbol{M}^{\star}\right\| \ll \sigma_{\min} \quad \text{and} \quad \left\|\boldsymbol{\Sigma}^{(l)}-\boldsymbol{\Sigma}^{\star}\right\| \leq \left\|\boldsymbol{M}^{(l)}-\boldsymbol{M}^{\star}\right\| \ll \sigma_{\min}, \tag{180}$$

which further implies

$$\frac{1}{2}\sigma_{\min} \leq \sigma_r\left(\boldsymbol{\Sigma}^{0}\right) \leq \sigma_1\left(\boldsymbol{\Sigma}^{0}\right) \leq 2\sigma_{\max} \quad \text{and} \quad \frac{1}{2}\sigma_{\min} \leq \sigma_r\left(\boldsymbol{\Sigma}^{(l)}\right) \leq \sigma_1\left(\boldsymbol{\Sigma}^{(l)}\right) \leq 2\sigma_{\max}. \tag{181}$$

We start by proving (70a), (70b), and (70c). The key decomposition we need is the following

$$\begin{aligned}\boldsymbol{X}^{0}\widehat{\boldsymbol{H}}^{0}-\boldsymbol{X}^{\star} = \boldsymbol{U}^{0}\left(\boldsymbol{\Sigma}^{0}\right)^{1/2}\left(\widehat{\boldsymbol{H}}^{0}-\boldsymbol{Q}\right)+\boldsymbol{U}^{0}\left[\left(\boldsymbol{\Sigma}^{0}\right)^{1/2}\boldsymbol{Q}-\boldsymbol{Q}\left(\boldsymbol{\Sigma}^{\star}\right)^{1/2}\right] \\ +\left(\boldsymbol{U}^{0}\boldsymbol{Q}-\boldsymbol{U}^{\star}\right)\left(\boldsymbol{\Sigma}^{\star}\right)^{1/2}.\end{aligned} \tag{182}$$

1. For the spectral norm error bound in (70c), the triangle inequality together with (182) yields

$$\begin{aligned}\left\|\boldsymbol{X}^{0}\widehat{\boldsymbol{H}}^{0}-\boldsymbol{X}^{\star}\right\| \leq \left\|\left(\boldsymbol{\Sigma}^{0}\right)^{1/2}\right\|\left\|\widehat{\boldsymbol{H}}^{0}-\boldsymbol{Q}\right\|+\left\|\left(\boldsymbol{\Sigma}^{0}\right)^{1/2}\boldsymbol{Q}-\boldsymbol{Q}\left(\boldsymbol{\Sigma}^{\star}\right)^{1/2}\right\| \\ +\sqrt{\sigma_{\max}}\left\|\boldsymbol{U}^{0}\boldsymbol{Q}-\boldsymbol{U}^{\star}\right\|,\end{aligned}$$

where we have also used the fact that $\|\boldsymbol{U}^{0}\|=1$. Recognizing that $\left\|\boldsymbol{M}^{0}-\boldsymbol{M}^{\star}\right\| \ll \sigma_{\min}$ (see (178)) and the assumption $\sigma_{\max}/\sigma_{\min} \lesssim 1$, we can apply Lemmas 47, 46, and 45 to obtain

$$\left\|\widehat{\boldsymbol{H}}^{0}-\boldsymbol{Q}\right\| \lesssim \frac{1}{\sigma_{\min}}\left\|\boldsymbol{M}^{0}-\boldsymbol{M}^{\star}\right\|, \tag{183a}$$

$$\left\|\left(\boldsymbol{\Sigma}^{0}\right)^{1/2}\boldsymbol{Q}-\boldsymbol{Q}\left(\boldsymbol{\Sigma}^{\star}\right)^{1/2}\right\| \lesssim \frac{1}{\sqrt{\sigma_{\min}}}\left\|\boldsymbol{M}^{0}-\boldsymbol{M}^{\star}\right\|, \tag{183b}$$

$$\left\|\boldsymbol{U}^{0}\boldsymbol{Q}-\boldsymbol{U}^{\star}\right\| \lesssim \frac{1}{\sigma_{\min}}\left\|\boldsymbol{M}^{0}-\boldsymbol{M}^{\star}\right\|. \tag{183c}$$

These taken collectively imply the advertised upper bound

$$\begin{aligned}\left\|\boldsymbol{X}^0\widehat{\boldsymbol{H}}^0-\boldsymbol{X}^\star\right\| &\lesssim \sqrt{\sigma_{\max}}\frac{1}{\sigma_{\min}}\left\|\boldsymbol{M}^0-\boldsymbol{M}^\star\right\|+\frac{1}{\sqrt{\sigma_{\min}}}\left\|\boldsymbol{M}^0-\boldsymbol{M}^\star\right\| \\ &\lesssim \frac{1}{\sqrt{\sigma_{\min}}}\left\|\boldsymbol{M}^0-\boldsymbol{M}^\star\right\| \\ &\lesssim \left\{\mu r\sqrt{\frac{1}{np}}\sqrt{\frac{\sigma_{\max}}{\sigma_{\min}}}+\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\right\}\left\|\boldsymbol{X}^\star\right\|,\end{aligned}$$

where we also utilize the fact that $\left\|\left(\boldsymbol{\Sigma}^0\right)^{1/2}\right\| \le \sqrt{2\sigma_{\max}}$ (see (181)) and the bounded condition number assumption, i.e., $\sigma_{\max}/\sigma_{\min} \lesssim 1$. This finishes the proof of (70c).

2. With regard to the Frobenius norm bound in (70a), one has

$$\begin{aligned}\left\|\boldsymbol{X}^0\widehat{\boldsymbol{H}}^0-\boldsymbol{X}^\star\right\|_{\mathrm{F}} &\le \sqrt{r}\left\|\boldsymbol{X}^0\widehat{\boldsymbol{H}}^0-\boldsymbol{X}^\star\right\| \\ &\overset{\text{(i)}}{\lesssim} \left\{\mu r\sqrt{\frac{1}{np}}+\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\right\}\sqrt{r}\left\|\boldsymbol{X}^\star\right\| \\ &= \left\{\mu r\sqrt{\frac{1}{np}}+\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\right\}\sqrt{r}\frac{\sqrt{\sigma_{\max}}}{\sqrt{\sigma_{\min}}}\sqrt{\sigma_{\min}} \\ &\overset{\text{(ii)}}{\lesssim} \left\{\mu r\sqrt{\frac{1}{np}}+\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\right\}\sqrt{r}\left\|\boldsymbol{X}^\star\right\|_{\mathrm{F}}.\end{aligned}$$

Here (i) arises from (70c) and (ii) holds true since $\sigma_{\max}/\sigma_{\min} \asymp 1$ and $\sqrt{r}\sqrt{\sigma_{\min}} \le \left\|\boldsymbol{X}^\star\right\|_{\mathrm{F}}$, thus completing the proof of (70a).

3. The proof of (70b) follows from similar arguments as used in proving (70c). Combine (182) and the triangle inequality to reach

$$\begin{aligned}&\left\|\boldsymbol{X}^0\widehat{\boldsymbol{H}}^0-\boldsymbol{X}^\star\right\|_{2,\infty} \\ &\le \left\|\boldsymbol{U}^0\right\|_{2,\infty}\left\{\left\|\left(\boldsymbol{\Sigma}^0\right)^{1/2}\right\|\left\|\widehat{\boldsymbol{H}}^0-\boldsymbol{Q}\right\|+\left\|\left(\boldsymbol{\Sigma}^0\right)^{1/2}\boldsymbol{Q}-\boldsymbol{Q}\left(\boldsymbol{\Sigma}^\star\right)^{1/2}\right\|\right\} \\ &\quad+\sqrt{\sigma_{\max}}\left\|\boldsymbol{U}^0\boldsymbol{Q}-\boldsymbol{U}^\star\right\|_{2,\infty}.\end{aligned}$$

Plugging in the estimates (178), (181), (183a), and (183b) results in

$$\begin{aligned}\left\|\boldsymbol{X}^0\widehat{\boldsymbol{H}}^0-\boldsymbol{X}^\star\right\|_{2,\infty} &\lesssim \left\{\mu r\sqrt{\frac{1}{np}}+\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\right\}\left\|\boldsymbol{X}^\star\right\|\left\|\boldsymbol{U}^0\right\|_{2,\infty} \\ &\quad+\sqrt{\sigma_{\max}}\left\|\boldsymbol{U}^0\boldsymbol{Q}-\boldsymbol{U}^\star\right\|_{2,\infty}.\end{aligned}$$

It remains to study the component-wise error of $\boldsymbol{U}^0$. To this end, it has already been shown in [1, Lemma 14] that

$$\left\|\boldsymbol{U}^0\boldsymbol{Q}-\boldsymbol{U}^\star\right\|_{2,\infty}\lesssim\left(\mu r\sqrt{\frac{1}{np}}+\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\right)\left\|\boldsymbol{U}^\star\right\|_{2,\infty}\quad\text{and}\quad\left\|\boldsymbol{U}^0\right\|_{2,\infty}\lesssim\left\|\boldsymbol{U}^\star\right\|_{2,\infty}\tag{184}$$

under our assumptions. These combined with the previous inequality give

$$\begin{aligned}\left\|\boldsymbol{X}^0\widehat{\boldsymbol{H}}^0-\boldsymbol{X}^\star\right\|_{2,\infty}&\lesssim\left\{\mu r\sqrt{\frac{1}{np}}+\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\right\}\sqrt{\sigma_{\max}}\left\|\boldsymbol{U}^\star\right\|_{2,\infty}\\&\lesssim\left\{\mu r\sqrt{\frac{1}{np}}+\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\right\}\left\|\boldsymbol{X}^\star\right\|_{2,\infty},\end{aligned}$$

where the last relation is due to the observation that

$$\sqrt{\sigma_{\max}}\left\|\boldsymbol{U}^\star\right\|_{2,\infty}\lesssim\sqrt{\sigma_{\min}}\left\|\boldsymbol{U}^\star\right\|_{2,\infty}\le\left\|\boldsymbol{X}^\star\right\|_{2,\infty}.$$

4. We now move on to proving (70e). Recall that $\boldsymbol{Q}^{(l)}=\arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}}\left\|\boldsymbol{U}^{0,(l)}\boldsymbol{R}-\boldsymbol{U}^\star\right\|_{\mathrm{F}}$. By the triangle inequality,

$$\begin{aligned}\left\|\left(\boldsymbol{X}^{0,(l)}\widehat{\boldsymbol{H}}^{0,(l)}-\boldsymbol{X}^\star\right)_{l,\cdot}\right\|_2&\le\left\|\boldsymbol{X}^{0,(l)}_{l,\cdot}\left(\widehat{\boldsymbol{H}}^{0,(l)}-\boldsymbol{Q}^{(l)}\right)\right\|_2+\left\|\left(\boldsymbol{X}^{0,(l)}\boldsymbol{Q}^{(l)}-\boldsymbol{X}^\star\right)_{l,\cdot}\right\|_2\\&\le\left\|\boldsymbol{X}^{0,(l)}_{l,\cdot}\right\|_2\left\|\widehat{\boldsymbol{H}}^{0,(l)}-\boldsymbol{Q}^{(l)}\right\|+\left\|\left(\boldsymbol{X}^{0,(l)}\boldsymbol{Q}^{(l)}-\boldsymbol{X}^\star\right)_{l,\cdot}\right\|_2.\end{aligned}\tag{185}$$

Note that $\boldsymbol{X}^\star_{l,\cdot}=\boldsymbol{M}^\star_{l,\cdot}\boldsymbol{U}^\star\left(\boldsymbol{\Sigma}^\star\right)^{-1/2}$ and, by construction of $\boldsymbol{M}^{(l)}$,

$$\boldsymbol{X}^{0,(l)}_{l,\cdot}=\boldsymbol{M}^{(l)}_{l,\cdot}\boldsymbol{U}^{0,(l)}\left(\boldsymbol{\Sigma}^{(l)}\right)^{-1/2}=\boldsymbol{M}^\star_{l,\cdot}\boldsymbol{U}^{0,(l)}\left(\boldsymbol{\Sigma}^{(l)}\right)^{-1/2}.$$

We can thus decompose

$$\begin{aligned}\left(\boldsymbol{X}^{0,(l)}\boldsymbol{Q}^{(l)}-\boldsymbol{X}^\star\right)_{l,\cdot}&=\boldsymbol{M}^\star_{l,\cdot}\left\{\boldsymbol{U}^{0,(l)}\left[\left(\boldsymbol{\Sigma}^{(l)}\right)^{-1/2}\boldsymbol{Q}^{(l)}-\boldsymbol{Q}^{(l)}\left(\boldsymbol{\Sigma}^\star\right)^{-1/2}\right]\right.\\&\quad\left.+\left(\boldsymbol{U}^{0,(l)}\boldsymbol{Q}^{(l)}-\boldsymbol{U}^\star\right)\left(\boldsymbol{\Sigma}^\star\right)^{-1/2}\right\},\end{aligned}$$

which further implies that

$$\begin{aligned}\left\|\left(\boldsymbol{X}^{0,(l)}\boldsymbol{Q}^{(l)}-\boldsymbol{X}^\star\right)_{l,\cdot}\right\|_2&\le\left\|\boldsymbol{M}^\star\right\|_{2,\infty}\left\{\left\|\left(\boldsymbol{\Sigma}^{(l)}\right)^{-1/2}\boldsymbol{Q}^{(l)}-\boldsymbol{Q}^{(l)}\left(\boldsymbol{\Sigma}^\star\right)^{-1/2}\right\|\right.\\&\quad\left.+\frac{1}{\sqrt{\sigma_{\min}}}\left\|\boldsymbol{U}^{0,(l)}\boldsymbol{Q}^{(l)}-\boldsymbol{U}^\star\right\|\right\}.\end{aligned}\tag{186}$$

In order to control this, we first see that

$$\begin{aligned}
&\left\| \left(\boldsymbol{\Sigma}^{(l)}\right)^{-1/2} \boldsymbol{Q}^{(l)} - \boldsymbol{Q}^{(l)} \left(\boldsymbol{\Sigma}^{\star}\right)^{-1/2} \right\| \\
&= \left\| \left(\boldsymbol{\Sigma}^{(l)}\right)^{-1/2} \left[\boldsymbol{Q}^{(l)} \left(\boldsymbol{\Sigma}^{\star}\right)^{1/2} - \left(\boldsymbol{\Sigma}^{(l)}\right)^{1/2} \boldsymbol{Q}^{(l)} \right] \left(\boldsymbol{\Sigma}^{\star}\right)^{-1/2} \right\| \\
&\lesssim \frac{1}{\sigma_{\min}} \left\| \boldsymbol{Q}^{(l)} \left(\boldsymbol{\Sigma}^{\star}\right)^{1/2} - \left(\boldsymbol{\Sigma}^{(l)}\right)^{-1/2} \boldsymbol{Q}^{(l)} \right\| \\
&\lesssim \frac{1}{\sigma_{\min}^{3/2}} \left\| \boldsymbol{M}^{(l)} - \boldsymbol{M}^{\star} \right\|,
\end{aligned}$$

where the penultimate inequality uses (181) and the last inequality arises from Lemma 46. Additionally, Lemma 45 gives

$$\left\| \boldsymbol{U}^{0,(l)} \boldsymbol{Q}^{(l)} - \boldsymbol{U}^{\star} \right\| \lesssim \frac{1}{\sigma_{\min}} \left\| \boldsymbol{M}^{(l)} - \boldsymbol{M}^{\star} \right\|.$$

Plugging the previous two bounds into (186), we reach

$$\begin{aligned}
\left\| \left(\boldsymbol{X}^{0,(l)} \boldsymbol{Q}^{(l)} - \boldsymbol{X}^{\star}\right)_{l,\cdot} \right\|_2 &\lesssim \frac{1}{\sigma_{\min}^{3/2}} \left\| \boldsymbol{M}^{(l)} - \boldsymbol{M}^{\star} \right\| \left\| \boldsymbol{M}^{\star} \right\|_{2,\infty} \\
&\lesssim \left\{ \mu r \sqrt{\frac{1}{np}} + \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \right\} \left\| \boldsymbol{X}^{\star} \right\|_{2,\infty}.
\end{aligned}$$

where the last relation follows from $\left\| \boldsymbol{M}^{\star} \right\|_{2,\infty} = \left\| \boldsymbol{X}^{\star} \boldsymbol{X}^{\star\top} \right\|_{2,\infty} \le \sqrt{\sigma_{\max}} \left\| \boldsymbol{X}^{\star} \right\|_{2,\infty}$ and estimate (179). Note that this also implies that $\left\| \boldsymbol{X}_{l,\cdot}^{0,(l)} \right\|_2 \le 2 \left\| \boldsymbol{X}^{\star} \right\|_{2,\infty}$. To see this, one has by the unitary invariance of $\left\| (\cdot)_{l,\cdot} \right\|_2$,

$$\left\| \boldsymbol{X}_{l,\cdot}^{0,(l)} \right\|_2 = \left\| \boldsymbol{X}_{l,\cdot}^{0,(l)} \boldsymbol{Q}^{(l)} \right\|_2 \le \left\| \left(\boldsymbol{X}^{0,(l)} \boldsymbol{Q}^{(l)} - \boldsymbol{X}^{\star}\right)_{l,\cdot} \right\|_2 + \left\| \boldsymbol{X}_{l,\cdot}^{\star} \right\|_2 \le 2 \left\| \boldsymbol{X}^{\star} \right\|_{2,\infty}.$$

Substituting the above bounds back to (185) yields in

$$\begin{aligned}
\left\| \left(\boldsymbol{X}^{0,(l)} \widehat{\boldsymbol{H}}^{0,(l)} - \boldsymbol{X}^{\star}\right)_{l,\cdot} \right\|_2 &\lesssim \left\| \boldsymbol{X}^{\star} \right\|_{2,\infty} \left\| \widehat{\boldsymbol{H}}^{0,(l)} - \boldsymbol{Q}^{(l)} \right\| \\
&\quad + \left\{ \mu r \sqrt{\frac{1}{np}} + \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \right\} \left\| \boldsymbol{X}^{\star} \right\|_{2,\infty} \\
&\lesssim \left\{ \mu r \sqrt{\frac{1}{np}} + \frac{\sigma}{\sigma_{\min}} \sqrt{\frac{n}{p}} \right\} \left\| \boldsymbol{X}^{\star} \right\|_{2,\infty},
\end{aligned}$$

where the second line relies on Lemma 47, bound (179), and condition $\sigma_{\max}/\sigma_{\min} \asymp 1$. This establishes (70e).

5. Our final step is to justify (70d). Define $\boldsymbol{B} := \arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \left\|\boldsymbol{U}^{0,(l)}\boldsymbol{R}-\boldsymbol{U}^{0}\right\|_{\mathrm{F}}$. From the definition of $\boldsymbol{R}^{0,(l)}$ (cf. (72)), one has

$$\left\|\boldsymbol{X}^{0}\widehat{\boldsymbol{H}}^{0}-\boldsymbol{X}^{0,(l)}\boldsymbol{R}^{0,(l)}\right\|_{\mathrm{F}} \le \left\|\boldsymbol{X}^{0,(l)}\boldsymbol{B}-\boldsymbol{X}^{0}\right\|_{\mathrm{F}}.$$

Recognizing that

$$\boldsymbol{X}^{0,(l)}\boldsymbol{B}-\boldsymbol{X}^{0}=\boldsymbol{U}^{0,(l)}\left[\left(\boldsymbol{\Sigma}^{(l)}\right)^{1/2}\boldsymbol{B}-\boldsymbol{B}\left(\boldsymbol{\Sigma}^{0}\right)^{1/2}\right]+\left(\boldsymbol{U}^{0,(l)}\boldsymbol{B}-\boldsymbol{U}^{0}\right)\left(\boldsymbol{\Sigma}^{0}\right)^{1/2},$$

we can use the triangle inequality to bound

$$\left\|\boldsymbol{X}^{0,(l)}\boldsymbol{B}-\boldsymbol{X}^{0}\right\|_{\mathrm{F}} \le \left\|\left(\boldsymbol{\Sigma}^{(l)}\right)^{1/2}\boldsymbol{B}-\boldsymbol{B}\left(\boldsymbol{\Sigma}^{0}\right)^{1/2}\right\|_{\mathrm{F}}+\left\|\boldsymbol{U}^{0,(l)}\boldsymbol{B}-\boldsymbol{U}^{0}\right\|_{\mathrm{F}}\left\|\left(\boldsymbol{\Sigma}^{0}\right)^{1/2}\right\|.$$

In view of Lemma 46 and bounds (178) and (179), one has

$$\left\|\left(\boldsymbol{\Sigma}^{(l)}\right)^{-1/2}\boldsymbol{B}-\boldsymbol{B}\boldsymbol{\Sigma}^{1/2}\right\|_{\mathrm{F}} \lesssim \frac{1}{\sqrt{\sigma_{\min}}}\left\|\left(\boldsymbol{M}^{0}-\boldsymbol{M}^{(l)}\right)\boldsymbol{U}^{0,(l)}\right\|_{\mathrm{F}}.$$

From Davis–Kahan's sinΘ theorem [39] we see that

$$\left\|\boldsymbol{U}^{0,(l)}\boldsymbol{B}-\boldsymbol{U}^{0}\right\|_{\mathrm{F}} \lesssim \frac{1}{\sigma_{\min}}\left\|\left(\boldsymbol{M}^{0}-\boldsymbol{M}^{(l)}\right)\boldsymbol{U}^{0,(l)}\right\|_{\mathrm{F}}.$$

These estimates taken together with (181) give

$$\left\|\boldsymbol{X}^{0,(l)}\boldsymbol{B}-\boldsymbol{X}^{0}\right\|_{\mathrm{F}} \lesssim \frac{1}{\sqrt{\sigma_{\min}}}\left\|\left(\boldsymbol{M}^{0}-\boldsymbol{M}^{(l)}\right)\boldsymbol{U}^{0,(l)}\right\|_{\mathrm{F}}.$$

It then boils down to controlling $\left\|\left(\boldsymbol{M}^{0}-\boldsymbol{M}^{(l)}\right)\boldsymbol{U}^{0,(l)}\right\|_{\mathrm{F}}$. Quantities of this type have shown up multiple times already, and hence, we omit the proof details for conciseness (see Appendix B.5). With probability at least $1-O\left(n^{-10}\right)$,

$$\left\|\left(\boldsymbol{M}^{0}-\boldsymbol{M}^{(l)}\right)\boldsymbol{U}^{0,(l)}\right\|_{\mathrm{F}} \lesssim \left\{\mu r\sqrt{\frac{\log n}{np}}\sigma_{\max}+\sigma\sqrt{\frac{n\log n}{p}}\right\}\left\|\boldsymbol{U}^{0,(l)}\right\|_{2,\infty}.$$

If one further has

$$\left\|\boldsymbol{U}^{0,(l)}\right\|_{2,\infty} \lesssim \left\|\boldsymbol{U}^{\star}\right\|_{2,\infty} \lesssim \frac{1}{\sqrt{\sigma_{\min}}}\left\|\boldsymbol{X}^{\star}\right\|_{2,\infty}, \tag{187}$$

then taking the previous bounds collectively establishes the desired bound

$$\left\|\boldsymbol{X}^{0}\widehat{\boldsymbol{H}}^{0}-\boldsymbol{X}^{0,(l)}\boldsymbol{R}^{0,(l)}\right\|_{\mathrm{F}} \lesssim \left\{\mu r\sqrt{\frac{\log n}{np}}+\frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\right\}\left\|\boldsymbol{X}^{\star}\right\|_{2,\infty}.$$

Proof of Claim (187) Denote by $\boldsymbol{M}^{(l),\mathrm{zero}}$ the matrix derived by zeroing out the lth row/column of $\boldsymbol{M}^{(l)}$, and $\boldsymbol{U}^{(l),\mathrm{zero}} \in \mathbb{R}^{n\times r}$ containing the leading r eigenvectors of $\boldsymbol{M}^{(l),\mathrm{zero}}$. On the one hand, [1, Lemma 4 and Lemma 14] demonstrate that

$$\max_{1\le l\le n} \|\boldsymbol{U}^{(l),\mathrm{zero}}\|_{2,\infty} \lesssim \|\boldsymbol{U}^{\star}\|_{2,\infty}.$$

On the other hand, by the Davis–Kahan sin Θ theorem [39] we obtain

$$\left\|\boldsymbol{U}^{0,(l)}\mathrm{sgn}\left(\boldsymbol{U}^{0,(l)\top}\boldsymbol{U}^{(l),\mathrm{zero}}\right)-\boldsymbol{U}^{(l),\mathrm{zero}}\right\|_{\mathrm{F}} \lesssim \frac{1}{\sigma_{\min}}\left\|\left(\boldsymbol{M}^{(l)}-\boldsymbol{M}^{(l),\mathrm{zero}}\right)\boldsymbol{U}^{(l),\mathrm{zero}}\right\|_{\mathrm{F}}, \tag{188}$$

where $\mathrm{sgn}(\boldsymbol{A})$ denotes the sign matrix of $\boldsymbol{A}$. For any $j \neq l$, one has

$$\left(\boldsymbol{M}^{(l)}-\boldsymbol{M}^{(l),\mathrm{zero}}\right)_{j,\cdot}\boldsymbol{U}^{(l),\mathrm{zero}} = \left(\boldsymbol{M}^{(l)}-\boldsymbol{M}^{(l),\mathrm{zero}}\right)_{j,l}\boldsymbol{U}^{(l),\mathrm{zero}}_{l,\cdot} = \boldsymbol{0}_{1\times r},$$

since the lth row of $\boldsymbol{U}^{(l),\mathrm{zero}}_{l,\cdot}$ is identically zero by construction. In addition,

$$\left\|\left(\boldsymbol{M}^{(l)}-\boldsymbol{M}^{(l),\mathrm{zero}}\right)_{l,\cdot}\boldsymbol{U}^{(l),\mathrm{zero}}\right\|_2 = \left\|\boldsymbol{M}^{\star}_{l,\cdot}\boldsymbol{U}^{(l),\mathrm{zero}}\right\|_2 \le \left\|\boldsymbol{M}^{\star}\right\|_{2,\infty} \le \sigma_{\max}\left\|\boldsymbol{U}^{\star}\right\|_{2,\infty}.$$

As a consequence, one has

$$\left\|\left(\boldsymbol{M}^{(l)}-\boldsymbol{M}^{(l),\mathrm{zero}}\right)\boldsymbol{U}^{(l),\mathrm{zero}}\right\|_{\mathrm{F}} = \left\|\left(\boldsymbol{M}^{(l)}-\boldsymbol{M}^{(l),\mathrm{zero}}\right)_{l,\cdot}\boldsymbol{U}^{(l),\mathrm{zero}}\right\|_2 \le \sigma_{\max}\left\|\boldsymbol{U}^{\star}\right\|_{2,\infty},$$

which combined with (188) and the assumption $\sigma_{\max}/\sigma_{\min} \asymp 1$ yields

$$\left\|\boldsymbol{U}^{0,(l)}\mathrm{sgn}\left(\boldsymbol{U}^{0,(l)\top}\boldsymbol{U}^{(l),\mathrm{zero}}\right)-\boldsymbol{U}^{(l),\mathrm{zero}}\right\|_{\mathrm{F}} \lesssim \left\|\boldsymbol{U}^{\star}\right\|_{2,\infty}$$

Claim (187) then follows by combining the above estimates:

$$\begin{aligned}\left\|\boldsymbol{U}^{0,(l)}\right\|_{2,\infty} &= \left\|\boldsymbol{U}^{0,(l)}\mathrm{sgn}\left(\boldsymbol{U}^{0,(l)\top}\boldsymbol{U}^{(l),\mathrm{zero}}\right)\right\|_{2,\infty}\\ &\le \|\boldsymbol{U}^{(l),\mathrm{zero}}\|_{2,\infty} + \left\|\boldsymbol{U}^{0,(l)}\mathrm{sgn}\left(\boldsymbol{U}^{0,(l)\top}\boldsymbol{U}^{(l),\mathrm{zero}}\right)-\boldsymbol{U}^{(l),\mathrm{zero}}\right\|_{\mathrm{F}} \lesssim \|\boldsymbol{U}^{\star}\|_{2,\infty},\end{aligned}$$

where we have utilized the unitary invariance of $\|\cdot\|_{2,\infty}$. □

C Proofs for Blind Deconvolution

Before proceeding to the proofs, we make note of the following concentration results. The standard Gaussian concentration inequality and the union bound give

$$\max_{1\le l\le m} \left|\boldsymbol{a}_l^{\mathsf{H}}\boldsymbol{x}^\star\right| \le 5\sqrt{\log m} \tag{189}$$

with probability at least $1 - O(m^{-10})$. In addition, with probability exceeding $1 - Cm\exp(-cK)$ for some constants $c, C > 0$,

$$\max_{1\le l\le m} \|\boldsymbol{a}_l\|_2 \le 3\sqrt{K}. \tag{190}$$

In addition, the population/expected Wirtinger Hessian at the truth $\boldsymbol{z}^\star$ is given by

$$\nabla^2 F(\boldsymbol{z}^\star) = \begin{bmatrix} \boldsymbol{I}_K & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{h}^\star\boldsymbol{x}^{\star\top} \\ \boldsymbol{0} & \boldsymbol{I}_K & \boldsymbol{x}^\star\boldsymbol{h}^{\star\top} & \boldsymbol{0} \\ \boldsymbol{0} & \left(\boldsymbol{x}^\star\boldsymbol{h}^{\star\top}\right)^{\mathsf{H}} & \boldsymbol{I}_K & \boldsymbol{0} \\ \left(\boldsymbol{h}^\star\boldsymbol{x}^{\star\top}\right)^{\mathsf{H}} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_K \end{bmatrix}. \tag{191}$$

C.1 Proof of Lemma 14

First, we find it convenient to decompose the Wirtinger Hessian (cf. (80)) into the expected Wirtinger Hessian at the truth (cf. (191)) and the perturbation part as follows:

$$\nabla^2 f(\boldsymbol{z}) = \nabla^2 F(\boldsymbol{z}^\star) + \left(\nabla^2 f(\boldsymbol{z}) - \nabla^2 F(\boldsymbol{z}^\star)\right). \tag{192}$$

The proof then proceeds by showing that (i) the population Hessian $\nabla^2 F(\boldsymbol{z}^\star)$ satisfies the restricted strong convexity and smoothness properties as advertised and (ii) the perturbation $\nabla^2 f(\boldsymbol{z}) - \nabla^2 F(\boldsymbol{z}^\star)$ is well controlled under our assumptions. We start by controlling the population Hessian in the following lemma.

Lemma 26 *Instate the notation and the conditions of Lemma 14. We have*

$$\left\|\nabla^2 F(\boldsymbol{z}^\star)\right\| = 2 \quad \text{and} \quad \boldsymbol{u}^{\mathsf{H}}\left[\boldsymbol{D}\nabla^2 F(\boldsymbol{z}^\star) + \nabla^2 F(\boldsymbol{z}^\star)\boldsymbol{D}\right]\boldsymbol{u} \ge \|\boldsymbol{u}\|_2^2.$$

The next step is to bound the perturbation. To this end, we define the set

$$\mathcal{S} := \{\boldsymbol{z} : \boldsymbol{z} \text{ satisfies (82)}\},$$

and derive the following lemma.

Lemma 27 *Suppose the sample complexity satisfies* $m \gg \mu^2 K \log^9 m$, $c > 0$ *is a sufficiently small constant, and* $\delta = c/\log^2 m$. *Then with probability at least* $1 - O\left(m^{-10} + e^{-K}\log m\right)$, *one has*

$$\sup_{z\in\mathcal{S}} \left\|\nabla^2 f(z) - \nabla^2 F\left(z^\star\right)\right\| \le 1/4.$$

Combining the two lemmas, we can easily see that for $z \in \mathcal{S}$,

$$\left\|\nabla^2 f(z)\right\| \le \left\|\nabla^2 F\left(z^\star\right)\right\| + \left\|\nabla^2 f(z) - \nabla^2 F\left(z^\star\right)\right\| \le 2 + 1/4 \le 3,$$

which verifies the smoothness upper bound. In addition,

$$\begin{aligned}
&\boldsymbol{u}^{\mathsf{H}}\left[\boldsymbol{D}\nabla^2 f(z) + \nabla^2 f(z)\,\boldsymbol{D}\right]\boldsymbol{u} \\
&= \boldsymbol{u}^{\mathsf{H}}\left[\boldsymbol{D}\nabla^2 F\left(z^\star\right) + \nabla^2 F\left(z^\star\right)\boldsymbol{D}\right]\boldsymbol{u} + \boldsymbol{u}^{\mathsf{H}}\boldsymbol{D}\left[\nabla^2 f(z) - \nabla^2 F\left(z^\star\right)\right]\boldsymbol{u} \\
&\quad + \boldsymbol{u}^{\mathsf{H}}\left[\nabla^2 f(z) - \nabla^2 F\left(z^\star\right)\right]\boldsymbol{D}\boldsymbol{u} \\
&\overset{\text{(i)}}{\ge} \boldsymbol{u}^{\mathsf{H}}\left[\boldsymbol{D}\nabla^2 F\left(z^\star\right) + \nabla^2 F\left(z^\star\right)\boldsymbol{D}\right]\boldsymbol{u} - 2\|\boldsymbol{D}\|\left\|\nabla^2 f(z) - \nabla^2 F\left(z^\star\right)\right\|\|\boldsymbol{u}\|_2^2 \\
&\overset{\text{(ii)}}{\ge} \|\boldsymbol{u}\|_2^2 - 2(1+\delta)\cdot\frac{1}{4}\|\boldsymbol{u}\|_2^2 \\
&\overset{\text{(iii)}}{\ge} \frac{1}{4}\|\boldsymbol{u}\|_2^2,
\end{aligned}$$

where (i) uses the triangle inequality, (ii) holds because of Lemma 27 and the fact that $\|\boldsymbol{D}\| \le 1+\delta$, and (iii) follows if $\delta \le 1/2$. This establishes the claim on the restricted strong convexity.

C.1.1 Proof of Lemma 26

We start by proving the identity $\left\|\nabla^2 F\left(z^\star\right)\right\| = 2$. Let

$$\boldsymbol{u}_1 = \frac{1}{\sqrt{2}}\begin{bmatrix}\boldsymbol{h}^\star \\ \boldsymbol{0} \\ \boldsymbol{0} \\ \overline{\boldsymbol{x}^\star}\end{bmatrix}, \qquad \boldsymbol{u}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix}\boldsymbol{0} \\ \boldsymbol{x}^\star \\ \overline{\boldsymbol{h}^\star} \\ \boldsymbol{0}\end{bmatrix}, \qquad \boldsymbol{u}_3 = \frac{1}{\sqrt{2}}\begin{bmatrix}\boldsymbol{h}^\star \\ \boldsymbol{0} \\ \boldsymbol{0} \\ -\overline{\boldsymbol{x}^\star}\end{bmatrix},$$

$$\boldsymbol{u}_4 = \frac{1}{\sqrt{2}}\begin{bmatrix}\boldsymbol{0} \\ \boldsymbol{x}^\star \\ -\overline{\boldsymbol{h}^\star} \\ \boldsymbol{0}\end{bmatrix}.$$

Recalling that $\|\boldsymbol{h}^\star\|_2 = \|\boldsymbol{x}^\star\|_2 = 1$, we can easily check that these four vectors form an orthonormal set. A little algebra reveals that

$$\nabla^2 F\left(\boldsymbol{z}^\star\right) = \boldsymbol{I}_{4K} + \boldsymbol{u}_1\boldsymbol{u}_1^{\mathsf{H}} + \boldsymbol{u}_2\boldsymbol{u}_2^{\mathsf{H}} - \boldsymbol{u}_3\boldsymbol{u}_3^{\mathsf{H}} - \boldsymbol{u}_4\boldsymbol{u}_4^{\mathsf{H}},$$

which immediately implies

$$\left\|\nabla^2 F\left(\boldsymbol{z}^\star\right)\right\| = 2.$$

We now turn attention to the restricted strong convexity. Since $\boldsymbol{u}^{\mathsf{H}}\boldsymbol{D}\nabla^2 F\left(\boldsymbol{z}^\star\right)\boldsymbol{u}$ is the complex conjugate of $\boldsymbol{u}^{\mathsf{H}}\nabla^2 F\left(\boldsymbol{z}^\star\right)\boldsymbol{D}\boldsymbol{u}$ as both $\nabla^2 F(\boldsymbol{z}^\star)$ and $\boldsymbol{D}$ are Hermitian, we will focus on the first term $\boldsymbol{u}^{\mathsf{H}}\boldsymbol{D}\nabla^2 F\left(\boldsymbol{z}^\star\right)\boldsymbol{u}$. This term can be rewritten as

$$\begin{aligned}
&\boldsymbol{u}^{\mathsf{H}}\boldsymbol{D}\nabla^2 F\left(\boldsymbol{z}^\star\right)\boldsymbol{u} \\
&\overset{\text{(i)}}{=} \left[(\boldsymbol{h}_1-\boldsymbol{h}_2)^{\mathsf{H}}, (\boldsymbol{x}_1-\boldsymbol{x}_2)^{\mathsf{H}}, \left(\overline{\boldsymbol{h}_1-\boldsymbol{h}_2}\right)^{\mathsf{H}}, \left(\overline{\boldsymbol{x}_1-\boldsymbol{x}_2}\right)^{\mathsf{H}}\right]\boldsymbol{D} \\
&\quad \begin{bmatrix} \boldsymbol{I}_K & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{h}^\star\boldsymbol{x}^{\star\top} \\ \boldsymbol{0} & \boldsymbol{I}_K & \boldsymbol{x}^\star\boldsymbol{h}^{\star\top} & \boldsymbol{0} \\ \boldsymbol{0} & \left(\boldsymbol{x}^\star\boldsymbol{h}^{\star\top}\right)^{\mathsf{H}} & \boldsymbol{I}_K & \boldsymbol{0} \\ \left(\boldsymbol{h}^\star\boldsymbol{x}^{\star\top}\right)^{\mathsf{H}} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_K \end{bmatrix} \begin{bmatrix} \boldsymbol{h}_1-\boldsymbol{h}_2 \\ \boldsymbol{x}_1-\boldsymbol{x}_2 \\ \overline{\boldsymbol{h}_1-\boldsymbol{h}_2} \\ \overline{\boldsymbol{x}_1-\boldsymbol{x}_2} \end{bmatrix} \\
&\overset{\text{(ii)}}{=} \left[\gamma_1(\boldsymbol{h}_1-\boldsymbol{h}_2)^{\mathsf{H}}, \gamma_2(\boldsymbol{x}_1-\boldsymbol{x}_2)^{\mathsf{H}}, \gamma_1\left(\overline{\boldsymbol{h}_1-\boldsymbol{h}_2}\right)^{\mathsf{H}}, \gamma_2\left(\overline{\boldsymbol{x}_1-\boldsymbol{x}_2}\right)^{\mathsf{H}}\right] \\
&\quad \begin{bmatrix} \boldsymbol{h}_1-\boldsymbol{h}_2+\boldsymbol{h}^\star\boldsymbol{x}^{\star\top}\overline{(\boldsymbol{x}_1-\boldsymbol{x}_2)} \\ \boldsymbol{x}_1-\boldsymbol{x}_2+\boldsymbol{x}^\star\boldsymbol{h}^{\star\top}\overline{(\boldsymbol{h}_1-\boldsymbol{h}_2)} \\ \left(\boldsymbol{x}^\star\boldsymbol{h}^{\star\top}\right)^{\mathsf{H}}(\boldsymbol{x}_1-\boldsymbol{x}_2)+\overline{(\boldsymbol{h}_1-\boldsymbol{h}_2)} \\ \left(\boldsymbol{h}^\star\boldsymbol{x}^{\star\top}\right)^{\mathsf{H}}(\boldsymbol{h}_1-\boldsymbol{h}_2)+\overline{(\boldsymbol{x}_1-\boldsymbol{x}_2)} \end{bmatrix} \\
&= \left[\gamma_1(\boldsymbol{h}_1-\boldsymbol{h}_2)^{\mathsf{H}}, \gamma_2(\boldsymbol{x}_1-\boldsymbol{x}_2)^{\mathsf{H}}, \gamma_1\left(\overline{\boldsymbol{h}_1-\boldsymbol{h}_2}\right)^{\mathsf{H}}, \gamma_2\left(\overline{\boldsymbol{x}_1-\boldsymbol{x}_2}\right)^{\mathsf{H}}\right] \\
&\quad \begin{bmatrix} \boldsymbol{h}_1-\boldsymbol{h}_2+\boldsymbol{h}^\star(\boldsymbol{x}_1-\boldsymbol{x}_2)^{\mathsf{H}}\boldsymbol{x}^\star \\ \boldsymbol{x}_1-\boldsymbol{x}_2+\boldsymbol{x}^\star(\boldsymbol{h}_1-\boldsymbol{h}_2)^{\mathsf{H}}\boldsymbol{h}^\star \\ \overline{\boldsymbol{h}_1-\boldsymbol{h}_2}+\overline{\boldsymbol{h}^\star(\boldsymbol{x}_1-\boldsymbol{x}_2)^{\mathsf{H}}\boldsymbol{x}^\star} \\ \overline{\boldsymbol{x}_1-\boldsymbol{x}_2}+\overline{\boldsymbol{x}^\star(\boldsymbol{h}_1-\boldsymbol{h}_2)^{\mathsf{H}}\boldsymbol{h}^\star} \end{bmatrix} \\
&= 2\gamma_1\|\boldsymbol{h}_1-\boldsymbol{h}_2\|_2^2 + 2\gamma_2\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_2^2 \\
&\quad + (\gamma_1+\gamma_2)\underbrace{(\boldsymbol{h}_1-\boldsymbol{h}_2)^{\mathsf{H}}\boldsymbol{h}^\star(\boldsymbol{x}_1-\boldsymbol{x}_2)^{\mathsf{H}}\boldsymbol{x}^\star}_{:=\beta} + (\gamma_1+\gamma_2)\underbrace{\overline{(\boldsymbol{h}_1-\boldsymbol{h}_2)^{\mathsf{H}}\boldsymbol{h}^\star(\boldsymbol{x}_1-\boldsymbol{x}_2)^{\mathsf{H}}\boldsymbol{x}^\star}}_{=\overline{\beta}},
\end{aligned} \tag{193}$$

where (i) uses the definitions of $\boldsymbol{u}$ and $\nabla^2 F\left(\boldsymbol{z}^\star\right)$, and (ii) follows from the definition of $\boldsymbol{D}$. In view of the assumption (84), we can obtain

$$\begin{aligned}
2\gamma_1\|\boldsymbol{h}_1-\boldsymbol{h}_2\|_2^2 + 2\gamma_2\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_2^2 &\geq 2\min\{\gamma_1,\gamma_2\}\left(\|\boldsymbol{h}_1-\boldsymbol{h}_2\|_2^2 + \|\boldsymbol{x}_1-\boldsymbol{x}_2\|_2^2\right) \\
&\geq (1-\delta)\|\boldsymbol{u}\|_2^2,
\end{aligned}$$

where the last inequality utilizes the identity

$$2\left\|\boldsymbol{h}_1-\boldsymbol{h}_2\right\|_2^2+2\left\|\boldsymbol{x}_1-\boldsymbol{x}_2\right\|_2^2=\|\boldsymbol{u}\|_2^2\,. \tag{194}$$

It then boils down to controlling β. Toward this goal, we decompose β into the following four terms

$$\begin{aligned}\beta=&\underbrace{\left(\boldsymbol{h}_1-\boldsymbol{h}_2\right)^{\mathsf{H}}\boldsymbol{h}_2\left(\boldsymbol{x}_1-\boldsymbol{x}_2\right)^{\mathsf{H}}\boldsymbol{x}_2}_{:=\beta_1}+\underbrace{\left(\boldsymbol{h}_1-\boldsymbol{h}_2\right)^{\mathsf{H}}\left(\boldsymbol{h}^\star-\boldsymbol{h}_2\right)\left(\boldsymbol{x}_1-\boldsymbol{x}_2\right)^{\mathsf{H}}\left(\boldsymbol{x}^\star-\boldsymbol{x}_2\right)}_{:=\beta_2}\\&+\underbrace{\left(\boldsymbol{h}_1-\boldsymbol{h}_2\right)^{\mathsf{H}}\left(\boldsymbol{h}^\star-\boldsymbol{h}_2\right)\left(\boldsymbol{x}_1-\boldsymbol{x}_2\right)^{\mathsf{H}}\boldsymbol{x}_2}_{:=\beta_3}+\underbrace{\left(\boldsymbol{h}_1-\boldsymbol{h}_2\right)^{\mathsf{H}}\boldsymbol{h}_2\left(\boldsymbol{x}_1-\boldsymbol{x}_2\right)^{\mathsf{H}}\left(\boldsymbol{x}^\star-\boldsymbol{x}_2\right)}_{:=\beta_4}.\end{aligned}$$

Since $\left\|\boldsymbol{h}_2-\boldsymbol{h}^\star\right\|_2$ and $\|\boldsymbol{x}_2-\boldsymbol{x}^\star\|_2$ are both small by (83), β_2, β_3, and β_4 are well bounded. Specifically, regarding β_2, we discover that

$$\begin{aligned}|\beta_2|&\le\left\|\boldsymbol{h}^\star-\boldsymbol{h}_2\right\|_2\left\|\boldsymbol{x}^\star-\boldsymbol{x}_2\right\|_2\left\|\boldsymbol{h}_1-\boldsymbol{h}_2\right\|_2\left\|\boldsymbol{x}_1-\boldsymbol{x}_2\right\|_2\\&\le\delta^2\left\|\boldsymbol{h}_1-\boldsymbol{h}_2\right\|_2\left\|\boldsymbol{x}_1-\boldsymbol{x}_2\right\|_2\le\delta\left\|\boldsymbol{h}_1-\boldsymbol{h}_2\right\|_2\left\|\boldsymbol{x}_1-\boldsymbol{x}_2\right\|_2,\end{aligned}$$

where the second inequality is due to (83) and the last one holds since $\delta<1$. Similarly, we can obtain

$$\begin{aligned}|\beta_3|&\le\delta\left\|\boldsymbol{x}_2\right\|_2\left\|\boldsymbol{h}_1-\boldsymbol{h}_2\right\|_2\left\|\boldsymbol{x}_1-\boldsymbol{x}_2\right\|_2\le2\delta\left\|\boldsymbol{h}_1-\boldsymbol{h}_2\right\|_2\left\|\boldsymbol{x}_1-\boldsymbol{x}_2\right\|_2,\text{ and}\\|\beta_4|&\le\delta\left\|\boldsymbol{h}_2\right\|_2\left\|\boldsymbol{h}_1-\boldsymbol{h}_2\right\|_2\left\|\boldsymbol{x}_1-\boldsymbol{x}_2\right\|_2\le2\delta\left\|\boldsymbol{h}_1-\boldsymbol{h}_2\right\|_2\left\|\boldsymbol{x}_1-\boldsymbol{x}_2\right\|_2,\end{aligned}$$

where both lines make use of the facts that

$$\begin{aligned}\left\|\boldsymbol{x}_2\right\|_2&\le\left\|\boldsymbol{x}_2-\boldsymbol{x}^\star\right\|_2+\left\|\boldsymbol{x}^\star\right\|_2\le1+\delta\le2\qquad\text{and}\\\left\|\boldsymbol{h}_2\right\|_2&\le\left\|\boldsymbol{h}_2-\boldsymbol{h}^\star\right\|_2+\left\|\boldsymbol{h}^\star\right\|_2\le1+\delta\le2.\end{aligned}\tag{195}$$

Combine the previous three bounds to reach

$$\begin{aligned}|\beta_2|+|\beta_3|+|\beta_4|&\le5\delta\left\|\boldsymbol{h}_1-\boldsymbol{h}_2\right\|_2\left\|\boldsymbol{x}_1-\boldsymbol{x}_2\right\|_2\\&\le5\delta\frac{\left\|\boldsymbol{h}_1-\boldsymbol{h}_2\right\|_2^2+\left\|\boldsymbol{x}_1-\boldsymbol{x}_2\right\|_2^2}{2}=\frac{5}{4}\delta\left\|\boldsymbol{u}\right\|_2^2,\end{aligned}$$

where we utilize the elementary inequality $ab\le(a^2+b^2)/2$ and identity (194).

The only remaining term is thus β_1. Recalling that $(\boldsymbol{h}_1,\boldsymbol{x}_1)$ and $(\boldsymbol{h}_2,\boldsymbol{x}_2)$ are aligned by our assumption, we can invoke Lemma 56 to obtain

$$\left(\boldsymbol{h}_1-\boldsymbol{h}_2\right)^{\mathsf{H}}\boldsymbol{h}_2=\left\|\boldsymbol{x}_1-\boldsymbol{x}_2\right\|_2^2+\boldsymbol{x}_2^{\mathsf{H}}\left(\boldsymbol{x}_1-\boldsymbol{x}_2\right)-\left\|\boldsymbol{h}_1-\boldsymbol{h}_2\right\|_2^2,$$

which allows one to rewrite β_1 as

$$\begin{aligned}\beta_1&=\left\{\left\|\boldsymbol{x}_1-\boldsymbol{x}_2\right\|_2^2+\boldsymbol{x}_2^{\mathsf{H}}\left(\boldsymbol{x}_1-\boldsymbol{x}_2\right)-\left\|\boldsymbol{h}_1-\boldsymbol{h}_2\right\|_2^2\right\}\cdot\left(\boldsymbol{x}_1-\boldsymbol{x}_2\right)^{\mathsf{H}}\boldsymbol{x}_2\\&=\left(\boldsymbol{x}_1-\boldsymbol{x}_2\right)^{\mathsf{H}}\boldsymbol{x}_2\left(\left\|\boldsymbol{x}_1-\boldsymbol{x}_2\right\|_2^2-\left\|\boldsymbol{h}_1-\boldsymbol{h}_2\right\|_2^2\right)+\left|\left(\boldsymbol{x}_1-\boldsymbol{x}_2\right)^{\mathsf{H}}\boldsymbol{x}_2\right|^2.\end{aligned}$$

Consequently,

$$\begin{aligned}\beta_1+\overline{\beta_1} &= 2\left|(\boldsymbol{x}_1-\boldsymbol{x}_2)^{\mathsf{H}}\boldsymbol{x}_2\right|_2^2+2\mathrm{Re}\left[(\boldsymbol{x}_1-\boldsymbol{x}_2)^{\mathsf{H}}\boldsymbol{x}_2\right]\left(\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_2^2-\|\boldsymbol{h}_1-\boldsymbol{h}_2\|_2^2\right)\\ &\geq 2\mathrm{Re}\left[(\boldsymbol{x}_1-\boldsymbol{x}_2)^{\mathsf{H}}\boldsymbol{x}_2\right]\left(\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_2^2-\|\boldsymbol{h}_1-\boldsymbol{h}_2\|_2^2\right)\\ &\overset{\text{(i)}}{\geq} -\left|(\boldsymbol{x}_1-\boldsymbol{x}_2)^{\mathsf{H}}\boldsymbol{x}_2\right|\|\boldsymbol{u}\|_2^2\\ &\overset{\text{(ii)}}{\geq} -4\delta\|\boldsymbol{u}\|_2^2.\end{aligned}$$

Here, (i) arises from the triangle inequality that

$$\left|\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_2^2-\|\boldsymbol{h}_1-\boldsymbol{h}_2\|_2^2\right|\leq\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_2^2+\|\boldsymbol{h}_1-\boldsymbol{h}_2\|_2^2=\frac{1}{2}\|\boldsymbol{u}\|_2^2,$$

and (ii) occurs since $\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_2\leq\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|_2+\|\boldsymbol{x}_2-\boldsymbol{x}^\star\|_2\leq 2\delta$ and $\|\boldsymbol{x}_2\|_2\leq 2$ (see (195)).

To finish up, note that $\gamma_1+\gamma_2\leq 2(1+\delta)\leq 3$ for $\delta<1/2$. Substitute these bounds into (193) to obtain

$$\begin{aligned}\boldsymbol{u}^{\mathsf{H}}\boldsymbol{D}\nabla^2F\left(\boldsymbol{z}^\star\right)\boldsymbol{u} &\geq (1-\delta)\|\boldsymbol{u}\|_2^2+(\gamma_1+\gamma_2)\left(\beta+\overline{\beta}\right)\\ &\geq (1-\delta)\|\boldsymbol{u}\|_2^2+(\gamma_1+\gamma_2)\left(\beta_1+\overline{\beta_1}\right)-2(\gamma_1+\gamma_2)\left(|\beta_2|+|\beta_3|+|\beta_4|\right)\\ &\geq (1-\delta)\|\boldsymbol{u}\|_2^2-12\delta\|\boldsymbol{u}\|_2^2-6\cdot\frac{5}{4}\delta\|\boldsymbol{u}\|_2^2\\ &\geq (1-20.5\delta)\|\boldsymbol{u}\|_2^2\\ &\geq \frac{1}{2}\|\boldsymbol{u}\|_2^2\end{aligned}$$

as long as δ is small enough.

C.1.2 Proof of Lemma 27

In view of the expressions of $\nabla^2 f(\boldsymbol{z})$ and $\nabla^2 F(\boldsymbol{z}^\star)$ (cf. (80) and (191)) and the triangle inequality, we get

$$\left\|\nabla^2 f(\boldsymbol{z})-\nabla^2 F\left(\boldsymbol{z}^\star\right)\right\|\leq 2\alpha_1+2\alpha_2+4\alpha_3+4\alpha_4, \tag{196}$$

where the four terms on the right-hand side are defined as follows

$$\alpha_1=\left\|\sum_{j=1}^m\left|\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}\right|^2\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}-\boldsymbol{I}_K\right\|,\qquad \alpha_2=\left\|\sum_{j=1}^m\left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}\right|^2\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}-\boldsymbol{I}_K\right\|,$$

$$\alpha_3=\left\|\sum_{j=1}^m\left(\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}\boldsymbol{a}_j-y_j\right)\boldsymbol{b}_j\boldsymbol{a}_j^{\mathsf{H}}\right\|,\qquad \alpha_4=\left\|\sum_{j=1}^m\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}\left(\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}\right)^\top-\boldsymbol{h}^\star\boldsymbol{x}^{\star\top}\right\|.$$

In what follows, we shall control $\sup_{z\in\mathcal{S}} \alpha_j$ for $j = 1, 2, 3, 4$ separately.

1. Regarding the first term α_1, the triangle inequality gives

$$\alpha_1 \le \underbrace{\left\| \sum_{j=1}^{m} \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x} \right|^2 \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} - \sum_{j=1}^{m} \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \right\|}_{:=\beta_1} + \underbrace{\left\| \sum_{j=1}^{m} \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} - \boldsymbol{I}_K \right\|}_{:=\beta_2}.$$

- To control β_1, the key observation is that $\boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}$ and $\boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star$ are extremely close. We can rewrite β_1 as

$$\beta_1 = \left\| \sum_{j=1}^{m} \left(\left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x} \right|^2 - \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 \right) \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \right\| \le \left\| \sum_{j=1}^{m} \left| \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x} \right|^2 - \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 \right| \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \right\|, \tag{197}$$

where

$$\begin{aligned}
&\left| \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x} \right|^2 - \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 \right| \\
&\overset{\text{(i)}}{=} \left| \left[\boldsymbol{a}_j^{\mathsf{H}} \left(\boldsymbol{x} - \boldsymbol{x}^\star \right) \right]^{\mathsf{H}} \boldsymbol{a}_j^{\mathsf{H}} \left(\boldsymbol{x} - \boldsymbol{x}^\star \right) + \left[\boldsymbol{a}_j^{\mathsf{H}} \left(\boldsymbol{x} - \boldsymbol{x}^\star \right) \right]^{\mathsf{H}} \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star + \left(\boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right)^{\mathsf{H}} \boldsymbol{a}_j^{\mathsf{H}} \left(\boldsymbol{x} - \boldsymbol{x}^\star \right) \right| \\
&\overset{\text{(ii)}}{\le} \left| \boldsymbol{a}_j^{\mathsf{H}} \left(\boldsymbol{x} - \boldsymbol{x}^\star \right) \right|^2 + 2 \left| \boldsymbol{a}_j^{\mathsf{H}} \left(\boldsymbol{x} - \boldsymbol{x}^\star \right) \right| \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right| \\
&\overset{\text{(iii)}}{\le} 4C_3^2 \frac{1}{\log^3 m} + 4C_3 \frac{1}{\log^{3/2} m} \cdot 5\sqrt{\log m} \\
&\lesssim C_3 \frac{1}{\log m}.
\end{aligned}$$

Here, the first line (i) uses the identity for $u, v \in \mathbb{C}$,

$$|u|^2 - |v|^2 = u^{\mathsf{H}} u - v^{\mathsf{H}} v = (u - v)^{\mathsf{H}} (u - v) + (u - v)^{\mathsf{H}} v + v^{\mathsf{H}} (u - v),$$

the second relation (ii) comes from the triangle inequality, and the third line (iii) follows from (189) and assumption (82b). Substitution into (197) gives

$$\beta_1 \le \max_{1\le j\le m} \left| \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x} \right|^2 - \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 \right| \left\| \sum_{j=1}^{m} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \right\| \lesssim C_3 \frac{1}{\log m},$$

where the last inequality comes from the fact that $\sum_{j=1}^{m} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} = \boldsymbol{I}_K$.

- The other term β_2 can be bounded through Lemma 59, which reveals that with probability $1 - O\left(m^{-10}\right)$,

$$\beta_2 \lesssim \sqrt{\frac{K}{m} \log m}.$$

Taken collectively, the preceding two bounds give

$$\sup_{z\in\mathcal{S}} \alpha_1 \lesssim \sqrt{\frac{K}{m}\log m} + C_3\frac{1}{\log m}.$$

Hence, $\mathbb{P}(\sup_{z\in\mathcal{S}} \alpha_1 \le 1/32) = 1 - O(m^{-10})$.

2. We are going to prove that $\mathbb{P}(\sup_{z\in\mathcal{S}} \alpha_2 \le 1/32) = 1 - O(m^{-10})$. The triangle inequality allows us to bound α_2 as

$$\alpha_2 \le \underbrace{\left\| \sum_{j=1}^{m} \left| \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h} \right|^2 \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} - \|\boldsymbol{h}\|_2^2 \boldsymbol{I}_K \right\|}_{:=\theta_1(\boldsymbol{h})} + \underbrace{\left\| \|\boldsymbol{h}\|_2^2 \boldsymbol{I}_K - \boldsymbol{I}_K \right\|}_{:=\theta_2(\boldsymbol{h})}.$$

The second term $\theta_2(\boldsymbol{h})$ is easy to control. To see this, we have

$$\theta_2(\boldsymbol{h}) = \left| \|\boldsymbol{h}\|_2^2 - 1 \right| = \left| \|\boldsymbol{h}\|_2 - 1 \right| (\|\boldsymbol{h}\|_2 + 1) \le 3\delta < 1/64,$$

where the penultimate relation uses the assumption that $\left\| \boldsymbol{h} - \boldsymbol{h}^\star \right\|_2 \le \delta$ and hence

$$\left| \|\boldsymbol{h}\|_2 - 1 \right| \le \delta, \qquad \|\boldsymbol{h}\|_2 \le 1 + \delta \le 2.$$

For the first term $\theta_1(\boldsymbol{h})$, we define a new set

$$\mathcal{H} := \left\{ \boldsymbol{h} \in \mathbb{C}^K : \ \|\boldsymbol{h} - \boldsymbol{h}^\star\|_2 \le \delta \quad \text{and} \quad \max_{1\le j\le m} \left| \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h} \right| \le \frac{2C_4 \mu \log^2 m}{\sqrt{m}} \right\}.$$

It is easily seen that $\sup_{z\in\mathcal{S}} \theta_1 \le \sup_{\boldsymbol{h}\in\mathcal{H}} \theta_1$. We plan to use the standard covering argument to show that

$$\mathbb{P}\left(\sup_{\boldsymbol{h}\in\mathcal{H}} \theta_1(\boldsymbol{h}) \le 1/64 \right) = 1 - O(m^{-10}). \tag{198}$$

To this end, we define $c_j(\boldsymbol{h}) = |\boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}|^2$ for every $1 \le j \le m$. It is straightforward to check that

$$\theta_1(\boldsymbol{h}) = \left\| \sum_{j=1}^{m} c_j(\boldsymbol{h}) \left(\boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} - \boldsymbol{I}_K \right) \right\|, \qquad \max_{1\le j\le m} |c_j| \le \left(\frac{2C_4 \mu \log^2 m}{\sqrt{m}} \right)^2, \tag{199}$$

$$\sum_{j=1}^{m} c_j^2 = \sum_{j=1}^{m} |\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}|^4 \le \left\{\max_{1\le j\le m} |\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}|^2\right\} \sum_{j=1}^{m} |\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}|^2 = \left\{\max_{1\le j\le m} |\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}|^2\right\} \|\boldsymbol{h}\|_2^2$$
$$\le 4\left(\frac{2C_4\mu \log^2 m}{\sqrt{m}}\right)^2 \tag{200}$$

for $\boldsymbol{h} \in \mathcal{H}$. In the above argument, we have used the facts that $\sum_{j=1}^{m} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} = \boldsymbol{I}_K$ and

$$\sum_{j=1}^{m} |\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}|^2 = \boldsymbol{h}^{\mathsf{H}} \left(\sum_{j=1}^{m} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}}\right) \boldsymbol{h} = \|\boldsymbol{h}\|_2^2 \le (1+\delta)^2 \le 4,$$

together with the definition of $\mathcal{H}$. Lemma 57 combined with (199) and (200) readily yields that for any fixed $\boldsymbol{h} \in \mathcal{H}$ and any $t \ge 0$,

$$\mathbb{P}(\theta_1(\boldsymbol{h}) \ge t) \le 2\exp\left(\widetilde{C}_1 K - \widetilde{C}_2 \min\left\{\frac{t}{\max_{1\le j\le m} |c_j|}, \frac{t^2}{\sum_{j=1}^{m} c_j^2}\right\}\right)$$
$$\le 2\exp\left(\widetilde{C}_1 K - \widetilde{C}_2 \frac{mt \min\{1, t/4\}}{4C_4^2\mu^2 \log^4 m}\right), \tag{201}$$

where $\widetilde{C}_1, \widetilde{C}_2 > 0$ are some universal constants.
Now we are in a position to strengthen this bound to obtain uniform control of θ_1 over $\mathcal{H}$. Note that for any $\boldsymbol{h}_1, \boldsymbol{h}_2 \in \mathcal{H}$,

$$|\theta_1(\boldsymbol{h}_1) - \theta_1(\boldsymbol{h}_2)| \le \left\|\sum_{j=1}^{m} \left(|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}_1|^2 - |\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}_2|^2\right) \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}}\right\| + \left|\|\boldsymbol{h}_1\|_2^2 - \|\boldsymbol{h}_2\|_2^2\right|$$
$$= \max_{1\le j\le m} \left||\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}_1|^2 - |\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}_2|^2\right| \left\|\sum_{j=1}^{m} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}}\right\| + \left|\|\boldsymbol{h}_1\|_2^2 - \|\boldsymbol{h}_2\|_2^2\right|,$$

where

$$\left||\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}_2|^2 - |\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}_1|^2\right| = \left|(\boldsymbol{h}_2 - \boldsymbol{h}_1)^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}_2 + \boldsymbol{h}_1^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} (\boldsymbol{h}_2 - \boldsymbol{h}_1)\right|$$
$$\le 2\max\{\|\boldsymbol{h}_1\|_2, \|\boldsymbol{h}_2\|_2\} \|\boldsymbol{h}_2 - \boldsymbol{h}_1\|_2 \|\boldsymbol{b}_j\|_2^2$$
$$\le 4\|\boldsymbol{h}_2 - \boldsymbol{h}_1\|_2 \|\boldsymbol{b}_j\|_2^2 \le \frac{4K}{m} \|\boldsymbol{h}_2 - \boldsymbol{h}_1\|_2$$

and

$$\left|\|\boldsymbol{h}_1\|_2^2 - \|\boldsymbol{h}_2\|_2^2\right| = \left|\boldsymbol{h}_1^{\mathsf{H}}(\boldsymbol{h}_1 - \boldsymbol{h}_2) - (\boldsymbol{h}_1 - \boldsymbol{h}_2)^{\mathsf{H}} \boldsymbol{h}_2\right|$$
$$\le 2\max\{\|\boldsymbol{h}_1\|_2, \|\boldsymbol{h}_2\|_2\} \|\boldsymbol{h}_2 - \boldsymbol{h}_1\|_2 \le 4\|\boldsymbol{h}_1 - \boldsymbol{h}_2\|_2.$$

Define an event $\mathcal{E}_0 = \left\{ \left\| \sum_{j=1}^m \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \right\| \le 2m \right\}$. When $\mathcal{E}_0$ happens, the previous estimates give

$$|\theta_1(\boldsymbol{h}_1) - \theta_1(\boldsymbol{h}_2)| \le (8K+4)\|\boldsymbol{h}_1 - \boldsymbol{h}_2\|_2 \le 10K\|\boldsymbol{h}_1 - \boldsymbol{h}_2\|_2, \qquad \forall \boldsymbol{h}_1, \boldsymbol{h}_2 \in \mathcal{H}.$$

Let $\varepsilon = 1/(1280K)$, and $\widetilde{\mathcal{H}}$ be an ε-net covering $\mathcal{H}$ (see [116, Definition 5.1]). We have

$$\left(\left\{ \sup_{\boldsymbol{h} \in \widetilde{\mathcal{H}}} \theta_1(\boldsymbol{h}) \le \frac{1}{128} \right\} \cap \mathcal{E}_0 \right) \subseteq \left\{ \sup_{\boldsymbol{h} \in \mathcal{H}} \theta_1 \le \frac{1}{64} \right\}$$

and, as a result,

$$\begin{aligned} \mathbb{P}\left(\sup_{\boldsymbol{h} \in \mathcal{H}} \theta_1(\boldsymbol{h}) \ge \frac{1}{64} \right) &\le \mathbb{P}\left(\sup_{\boldsymbol{h} \in \widetilde{\mathcal{H}}} \theta_1(\boldsymbol{h}) \ge \frac{1}{128} \right) + \mathbb{P}(\mathcal{E}_0^c) \\ &\le |\widetilde{\mathcal{H}}| \cdot \max_{\boldsymbol{h} \in \widetilde{\mathcal{H}}} \mathbb{P}\left(\theta_1(\boldsymbol{h}) \ge \frac{1}{128} \right) + \mathbb{P}(\mathcal{E}_0^c). \end{aligned}$$

Lemma 57 forces that $\mathbb{P}(\mathcal{E}_0^c) = O(m^{-10})$. Additionally, we have $\log|\widetilde{\mathcal{H}}| \le \widetilde{C}_3 K \log K$ for some absolute constant $\widetilde{C}_3 > 0$ according to [116, Lemma 5.2]. Hence, (201) leads to

$$\begin{aligned} &|\widetilde{\mathcal{H}}| \cdot \max_{\boldsymbol{h} \in \widetilde{\mathcal{H}}} \mathbb{P}\left(\theta_1(\boldsymbol{h}) \ge \frac{1}{128} \right) \\ &\le 2\exp\left(\widetilde{C}_3 K \log K + \widetilde{C}_1 K - \widetilde{C}_2 \frac{m(1/128)\min\{1, (1/128)/4\}}{4C_4^2 \mu^2 \log^4 m} \right) \\ &\le 2\exp\left(2\widetilde{C}_3 K \log m - \frac{\widetilde{C}_4 m}{\mu^2 \log^4 m} \right) \end{aligned}$$

for some constant $\widetilde{C}_4 > 0$. Under the sample complexity $m \gg \mu^2 K \log^5 m$, the right-hand side of the above display is at most $O\left(m^{-10}\right)$. Combine the estimates above to establish the desired high-probability bound for $\sup_{z \in \mathcal{S}} \alpha_2$.

3. Next, we will demonstrate that

$$\mathbb{P}(\sup_{z \in \mathcal{S}} \alpha_3 \le 1/96) = 1 - O\left(m^{-10} + e^{-K}\log m\right).$$

To this end, we let

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{a}_1^{\mathsf{H}} \\ \vdots \\ \boldsymbol{a}_m^{\mathsf{H}} \end{bmatrix} \in \mathbb{C}^{m \times K}, \quad \boldsymbol{B} = \begin{bmatrix} \boldsymbol{b}_1^{\mathsf{H}} \\ \vdots \\ \boldsymbol{b}_m^{\mathsf{H}} \end{bmatrix} \in \mathbb{C}^{m \times K},$$

$$\boldsymbol{C} = \begin{bmatrix} c_1(\boldsymbol{z}) & & & \\ & c_2(\boldsymbol{z}) & & \\ & & \cdots & \\ & & & c_m(\boldsymbol{z}) \end{bmatrix} \in \mathbb{C}^{m\times m},$$

where for each $1 \le j \le m$,

$$c_j(\boldsymbol{z}) := \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} \boldsymbol{a}_j - y_j = \boldsymbol{b}_j^{\mathsf{H}} (\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} - \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}}) \boldsymbol{a}_j.$$

As a consequence, we can write $\alpha_3 = \|\boldsymbol{B}^{\mathsf{H}} \boldsymbol{C} \boldsymbol{A}\|$.
The key observation is that both the ℓ_∞ norm and the Frobenius norm of $\boldsymbol{C}$ are well controlled. Specifically, we claim for the moment that with probability at least $1 - O\left(m^{-10}\right)$,

$$\|\boldsymbol{C}\|_\infty = \max_{1\le j\le m} |c_j| \le C \frac{\mu \log^{5/2} m}{\sqrt{m}}; \tag{202a}$$

$$\|\boldsymbol{C}\|_{\mathrm{F}}^2 = \sum_{j=1}^{m} |c_j|^2 \le 12\delta^2, \tag{202b}$$

where $C > 0$ is some absolute constant. This motivates us to divide the entries in $\boldsymbol{C}$ into multiple groups based on their magnitudes.
To be precise, introduce $R := 1 + \lceil \log_2(C\mu \log^{7/2} m) \rceil$ sets $\{\mathcal{I}_r\}_{1\le r\le R}$, where

$$\mathcal{I}_r = \left\{ j \in [m] : \frac{C\mu \log^{5/2} m}{2^r \sqrt{m}} < |c_j| \le \frac{C\mu \log^{5/2} m}{2^{r-1}\sqrt{m}} \right\}, \qquad 1 \le r \le R-1$$

and $\mathcal{I}_R = \{1, \cdots, m\} \setminus \left(\bigcup_{r=1}^{R-1} \mathcal{I}_r \right)$. An immediate consequence of the definition of $\mathcal{I}_r$ and the norm constraints in (202) is the following cardinality bound

$$|\mathcal{I}_r| \le \frac{\|\boldsymbol{C}\|_{\mathrm{F}}^2}{\min_{j\in\mathcal{I}_r} |c_j|^2} \le \frac{12\delta^2}{\left(\frac{C\mu \log^{5/2} m}{2^r\sqrt{m}}\right)^2} = \underbrace{\frac{12\delta^2 4^r}{C^2\mu^2 \log^5 m}}_{\delta_r} m \tag{203}$$

for $1 \le r \le R-1$. Since $\{\mathcal{I}_r\}_{1\le r\le R}$ form a partition of the index set $\{1, \cdots, m\}$, it is easy to see that

$$\boldsymbol{B}^{\mathsf{H}} \boldsymbol{C} \boldsymbol{A} = \sum_{r=1}^{R} (\boldsymbol{B}_{\mathcal{I}_r,\cdot})^{\mathsf{H}} \boldsymbol{C}_{\mathcal{I}_r,\mathcal{I}_r} \boldsymbol{A}_{\mathcal{I}_r,\cdot},$$

where $\boldsymbol{D}_{\mathcal{I},\mathcal{J}}$ denotes the submatrix of $\boldsymbol{D}$ induced by the rows and columns of $\boldsymbol{D}$ having indices from $\mathcal{I}$ and $\mathcal{J}$, respectively, and $\boldsymbol{D}_{\mathcal{I},\cdot}$ refers to the submatrix

formed by the rows from the index set $\mathcal{I}$. As a result, one can invoke the triangle inequality to derive

$$\alpha_3 \le \sum_{r=1}^{R-1} \left\|\boldsymbol{B}_{\mathcal{I}_r,\cdot}\right\| \cdot \left\|\boldsymbol{C}_{\mathcal{I}_r,\mathcal{I}_r}\right\| \cdot \left\|\boldsymbol{A}_{\mathcal{I}_r,\cdot}\right\| + \left\|\boldsymbol{B}_{\mathcal{I}_R,\cdot}\right\| \cdot \left\|\boldsymbol{C}_{\mathcal{I}_R,\mathcal{I}_R}\right\| \cdot \left\|\boldsymbol{A}_{\mathcal{I}_R,\cdot}\right\|. \tag{204}$$

Recognizing that $\boldsymbol{B}^{\mathsf{H}}\boldsymbol{B} = \boldsymbol{I}_K$, we obtain

$$\left\|\boldsymbol{B}_{\mathcal{I}_r,\cdot}\right\| \le \|\boldsymbol{B}\| = 1$$

for every $1 \le r \le R$. In addition, by construction of $\mathcal{I}_r$, we have

$$\left\|\boldsymbol{C}_{\mathcal{I}_r,\mathcal{I}_r}\right\| = \max_{j\in\mathcal{I}_r} |c_j| \le \frac{C\mu \log^{5/2} m}{2^{r-1}\sqrt{m}}$$

for $1 \le r \le R$, and specifically for R, one has

$$\left\|\boldsymbol{C}_{\mathcal{I}_R,\mathcal{I}_R}\right\| = \max_{j\in\mathcal{I}_R} |c_j| \le \frac{C\mu \log^{5/2} m}{2^{R-1}\sqrt{m}} \le \frac{1}{\sqrt{m}\log m},$$

which follows from the definition of R, i.e., $R = 1+\lceil \log_2(C\mu \log^{7/2} m)\rceil$. Regarding $\left\|\boldsymbol{A}_{\mathcal{I}_r,\cdot}\right\|$, we discover that $\left\|\boldsymbol{A}_{I_R,\cdot}\right\| \le \|\boldsymbol{A}\|$ and, in view of (203),

$$\left\|\boldsymbol{A}_{\mathcal{I}_r,\cdot}\right\| \le \sup_{\mathcal{I}:|\mathcal{I}|\le\delta_r m} \left\|\boldsymbol{A}_{\mathcal{I},\cdot}\right\|, \qquad 1 \le r \le R-1.$$

Substitute the above estimates into (204) to get

$$\alpha_3 \le \sum_{r=1}^{R-1} \frac{C\mu \log^{5/2} m}{2^{r-1}\sqrt{m}} \sup_{\mathcal{I}:|\mathcal{I}|\le\delta_r m} \left\|\boldsymbol{A}_{\mathcal{I},\cdot}\right\| + \frac{\|\boldsymbol{A}\|}{\sqrt{m}\log m}. \tag{205}$$

It remains to upper bound $\|\boldsymbol{A}\|$ and $\sup_{\mathcal{I}:|\mathcal{I}|\le\delta_r m} \left\|\boldsymbol{A}_{\mathcal{I},\cdot}\right\|$. Lemma 57 tells us that $\|\boldsymbol{A}\| \le 2\sqrt{m}$ with probability at least $1 - O\left(m^{-10}\right)$. Furthermore, we can invoke Lemma 58 to bound $\sup_{\mathcal{I}:|\mathcal{I}|\le\delta_r m} \left\|\boldsymbol{A}_{\mathcal{I},\cdot}\right\|$ for each $1 \le r \le R-1$. It is easily seen from our assumptions $m \gg \mu^2 K \log^9 m$ and $\delta = c/\log^2 m$ that $\delta_r \gg K/m$. In addition,

$$\delta_r \le \frac{12\delta^2 4^{R-1}}{C^2\mu^2 \log^5 m} \le \frac{12\delta^2 4^{1+\log_2(C\mu \log^{7/2} m)}}{C^2\mu^2 \log^5 m} = 48\delta^2 \log^2 m = \frac{48c}{\log^2 m} \ll 1.$$

By Lemma 58, we obtain that for some constants $\widetilde{C}_2, \widetilde{C}_3 > 0$

$$\mathbb{P}\left(\sup_{\mathcal{I}:|\mathcal{I}|\le\delta_r m}\left\|\boldsymbol{A}_{\mathcal{I},\cdot}\right\| \ge \sqrt{4\widetilde{C}_3\delta_r m\log(e/\delta_r)}\right) \le 2\exp\left(-\frac{\widetilde{C}_2\widetilde{C}_3}{3}\delta_r m\log(e/\delta_r)\right)$$
$$\le 2\exp\left(-\frac{\widetilde{C}_2\widetilde{C}_3}{3}\delta_r m\right) \le 2e^{-K}.$$

Taking the union bound and substituting the estimates above into (205), we see that with probability at least $1 - O\left(m^{-10}\right) - O\left((R-1)e^{-K}\right)$,

$$\begin{aligned}\alpha_3 &\le \sum_{r=1}^{R-1}\frac{C\mu\log^{5/2} m}{2^{r-1}\sqrt{m}}\cdot\sqrt{4\widetilde{C}_3\delta_r m\log(e/\delta_r)} + \frac{2\sqrt{m}}{\sqrt{m}\log m}\\ &\le \sum_{r=1}^{R-1} 4\delta\sqrt{12\widetilde{C}_3\log(e/\delta_r)} + \frac{2}{\log m}\\ &\lesssim (R-1)\delta\sqrt{\log(e/\delta_1)} + \frac{1}{\log m}.\end{aligned}$$

Note that $\mu \le \sqrt{m}$, $R - 1 = \lceil \log_2(C\mu\log^{7/2} m)\rceil \lesssim \log m$, and

$$\sqrt{\log\frac{e}{\delta_1}} = \sqrt{\log\left(\frac{eC^2\mu^2\log^5 m}{48\delta^2}\right)} \lesssim \log m.$$

Therefore, with probability exceeding $1 - O\left(m^{-10}\right) - O\left(e^{-K}\log m\right)$,

$$\sup_{z\in\mathcal{S}}\alpha_3 \lesssim \delta\log^2 m + \frac{1}{\log m}.$$

By taking c to be small enough in $\delta = c/\log^2 m$, we get

$$\mathbb{P}\left(\sup_{z\in\mathcal{S}}\alpha_3 \ge 1/96\right) \le O\left(m^{-10}\right) + O\left(e^{-K}\log m\right)$$

as claimed.

Finally, it remains to justify (202). For all $\boldsymbol{z} \in \mathcal{S}$, the triangle inequality tells us that

$$\begin{aligned}|c_j| &\le \left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}(\boldsymbol{x}-\boldsymbol{x}^\star)^{\mathsf{H}}\boldsymbol{a}_j\right| + \left|\boldsymbol{b}_j^{\mathsf{H}}(\boldsymbol{h}-\boldsymbol{h}^\star)\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\right|\\ &\le \left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}\right|\cdot\left|\boldsymbol{a}_j^{\mathsf{H}}(\boldsymbol{x}-\boldsymbol{x}^\star)\right| + \left(\left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}\right| + \left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\right|\right)\cdot\left|\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}^\star\right|\\ &\le \frac{2C_4\mu\log^2 m}{\sqrt{m}}\cdot\frac{2C_3}{\log^{3/2} m} + \left(\frac{2C_4\mu\log^2 m}{\sqrt{m}} + \frac{\mu}{\sqrt{m}}\right)5\sqrt{\log m}\end{aligned}$$

$$\leq C \frac{\mu \log^{5/2} m}{\sqrt{m}},$$

for some large constant $C > 0$, where we have used the definition of $\mathcal{S}$ and fact (189). Claim (202b) follows directly from [76, Lemma 5.14]. To avoid confusion, we use μ_1 to refer to the parameter μ therein. Let $L = m$, $N = K$, $d_0 = 1$, $\mu_1 = C_4 \mu \log^2 m/2$, and $\varepsilon = 1/15$. Then

$$\mathcal{S} \subseteq \mathcal{N}_{d_0} \cap \mathcal{N}_{\mu_1} \cap \mathcal{N}_{\varepsilon},$$

and the sample complexity condition $L \gg \mu_1^2 (K+N) \log^2 L$ is satisfied because we have assumed $m \gg \mu^2 K \log^6 m$. Therefore, with probability exceeding $1 - O\left(m^{-10} + e^{-K}\right)$, we obtain that for all $\boldsymbol{z} \in \mathcal{S}$,

$$\|\boldsymbol{C}\|_{\mathrm{F}}^2 \leq \frac{5}{4} \left\| \boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} - \boldsymbol{h}^\star \boldsymbol{x}^{\star \mathsf{H}} \right\|_{\mathrm{F}}^2.$$

Claim (202b) can then be justified by observing that

$$\begin{aligned} \left\| \boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} - \boldsymbol{h}^\star \boldsymbol{x}^{\star \mathsf{H}} \right\|_{\mathrm{F}} &= \left\| \boldsymbol{h} \left(\boldsymbol{x} - \boldsymbol{x}^\star\right)^{\mathsf{H}} + \left(\boldsymbol{h} - \boldsymbol{h}^\star\right) \boldsymbol{x}^{\star \mathsf{H}} \right\|_{\mathrm{F}} \\ &\leq \|\boldsymbol{h}\|_2 \left\|\boldsymbol{x} - \boldsymbol{x}^\star\right\|_2 + \left\|\boldsymbol{h} - \boldsymbol{h}^\star\right\|_2 \left\|\boldsymbol{x}^\star\right\|_2 \leq 3\delta. \end{aligned}$$

4. It remains to control α_4, for which we make note of the following inequality

$$\alpha_4 \leq \underbrace{\left\| \sum_{j=1}^m \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} (\boldsymbol{h}\boldsymbol{x}^\top - \boldsymbol{h}^\star \boldsymbol{x}^{\star \top}) \overline{\boldsymbol{a}_j}\, \overline{\boldsymbol{a}_j}^{\mathsf{H}} \right\|}_{\theta_3} + \underbrace{\left\| \sum_{j=1}^m \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{x}^{\star \top} (\overline{\boldsymbol{a}_j}\, \overline{\boldsymbol{a}_j}^{\mathsf{H}} - \boldsymbol{I}_K) \right\|}_{\theta_4}$$

with $\overline{\boldsymbol{a}_j}$ denoting the entrywise conjugate of $\boldsymbol{a}_j$. Since $\{\overline{\boldsymbol{a}_j}\}$ has the same joint distribution as $\{\boldsymbol{a}_j\}$, by the same argument used for bounding α_3 we obtain control of the first term, namely

$$\mathbb{P}\left(\sup_{\boldsymbol{z} \in \mathcal{S}} \theta_3 \geq 1/96 \right) = O(m^{-10} + e^{-K} \log m).$$

Note that $m \gg \mu^2 K \log m/\delta^2$ and $\delta \ll 1$. According to [76, Lemma 5.20],

$$\mathbb{P}\left(\sup_{\boldsymbol{z} \in \mathcal{S}} \theta_4 \geq 1/96 \right) \leq \mathbb{P}\left(\sup_{\boldsymbol{z} \in \mathcal{S}} \theta_4 \geq \delta \right) = O(m^{-10}).$$

Putting together the above bounds, we reach $\mathbb{P}(\sup_{\boldsymbol{z} \in \mathcal{S}} \alpha_4 \leq 1/48) = 1 - O(m^{-10} + e^{-K} \log m)$.

5. Combining all the previous bounds for $\sup_{z\in\mathcal{S}} \alpha_j$ and (196), we deduce that with probability $1 - O(m^{-10} + e^{-K}\log m)$,

$$\left\|\nabla^2 f(z) - \nabla^2 F\left(z^\star\right)\right\| \le 2\cdot\frac{1}{32} + 2\cdot\frac{1}{32} + 4\cdot\frac{1}{96} + 4\cdot\frac{1}{48} = \frac{1}{4}.$$

C.2 Proofs of Lemmas 15 and 16

Proof of Lemma 15 In view of the definition of α^{t+1} (see (38)), one has

$$\begin{aligned}\operatorname{dist}\left(z^{t+1}, z^\star\right)^2 &= \left\|\frac{1}{\alpha^{t+1}}\boldsymbol{h}^{t+1} - \boldsymbol{h}^\star\right\|_2^2 + \left\|\alpha^{t+1}\boldsymbol{x}^{t+1} - \boldsymbol{x}^\star\right\|_2^2 \\ &\le \left\|\frac{1}{\alpha^{t}}\boldsymbol{h}^{t+1} - \boldsymbol{h}^\star\right\|_2^2 + \left\|\alpha^{t}\boldsymbol{x}^{t+1} - \boldsymbol{x}^\star\right\|_2^2.\end{aligned}$$

The gradient update rules (79) imply that

$$\begin{aligned}\frac{1}{\alpha^t}\boldsymbol{h}^{t+1} &= \frac{1}{\alpha^t}\left(\boldsymbol{h}^t - \frac{\eta}{\|\boldsymbol{x}^t\|_2^2}\nabla_{\boldsymbol{h}} f\left(z^t\right)\right) = \widetilde{\boldsymbol{h}}^t - \frac{\eta}{\|\widetilde{\boldsymbol{x}}^t\|_2^2}\nabla_{\boldsymbol{h}} f\left(\widetilde{z}^t\right), \\ \alpha^t\boldsymbol{x}^{t+1} &= \alpha^t\left(\boldsymbol{x}^t - \frac{\eta}{\|\boldsymbol{h}^t\|_2^2}\nabla_{\boldsymbol{x}} f\left(z^t\right)\right) = \widetilde{\boldsymbol{x}}^t - \frac{\eta}{\|\widetilde{\boldsymbol{h}}^t\|_2^2}\nabla_{\boldsymbol{x}} f\left(\widetilde{z}^t\right),\end{aligned}$$

where we denote $\widetilde{\boldsymbol{h}}^t = \frac{1}{\alpha^t}\boldsymbol{h}^t$ and $\widetilde{\boldsymbol{x}}^t = \alpha^t\boldsymbol{x}^t$ as in (81). Let $\widehat{\boldsymbol{h}}^{t+1} = \frac{1}{\alpha^t}\boldsymbol{h}^{t+1}$ and $\widehat{\boldsymbol{x}}^{t+1} = \alpha^t\boldsymbol{x}^{t+1}$. We further get

$$\begin{bmatrix}\widehat{\boldsymbol{h}}^{t+1} - \boldsymbol{h}^\star \\ \widehat{\boldsymbol{x}}^{t+1} - \boldsymbol{x}^\star \\ \hline \widehat{\boldsymbol{h}}^{t+1} - \boldsymbol{h}^\star \\ \widehat{\boldsymbol{x}}^{t+1} - \boldsymbol{x}^\star\end{bmatrix} = \begin{bmatrix}\widetilde{\boldsymbol{h}}^{t} - \boldsymbol{h}^\star \\ \widetilde{\boldsymbol{x}}^{t} - \boldsymbol{x}^\star \\ \hline \widetilde{\boldsymbol{h}}^{t} - \boldsymbol{h}^\star \\ \widetilde{\boldsymbol{x}}^{t} - \boldsymbol{x}^\star\end{bmatrix} - \eta\underbrace{\begin{bmatrix}\|\widetilde{\boldsymbol{x}}^t\|_2^{-2}\boldsymbol{I}_K & & & \\ & \|\widetilde{\boldsymbol{h}}^t\|_2^{-2}\boldsymbol{I}_K & & \\ & & \|\widetilde{\boldsymbol{x}}^t\|_2^{-2}\boldsymbol{I}_K & \\ & & & \|\widetilde{\boldsymbol{h}}^t\|_2^{-2}\boldsymbol{I}_K\end{bmatrix}}_{:=\boldsymbol{D}}\begin{bmatrix}\nabla_{\boldsymbol{h}} f\left(\widetilde{z}^t\right) \\ \nabla_{\boldsymbol{x}} f\left(\widetilde{z}^t\right) \\ \hline \nabla_{\boldsymbol{h}} f\left(\widetilde{z}^t\right) \\ \nabla_{\boldsymbol{x}} f\left(\widetilde{z}^t\right)\end{bmatrix}. \tag{206}$$

The fundamental theorem of calculus (see Appendix D.3.1) together with the fact that $\nabla f(z^\star) = \mathbf{0}$ tells us

$$\begin{bmatrix} \nabla_{\boldsymbol{h}} f\left(\widetilde{z}^t\right) \\ \nabla_{\boldsymbol{x}} f\left(\widetilde{z}^t\right) \\ \hline \nabla_{\boldsymbol{h}} f\left(\widetilde{z}^t\right) \\ \nabla_{\boldsymbol{x}} f\left(\widetilde{z}^t\right) \end{bmatrix} = \begin{bmatrix} \nabla_{\boldsymbol{h}} f\left(\widetilde{z}^t\right) - \nabla_{\boldsymbol{h}} f\left(z^\star\right) \\ \nabla_{\boldsymbol{x}} f\left(\widetilde{z}^t\right) - \nabla_{\boldsymbol{x}} f\left(z^\star\right) \\ \hline \nabla_{\boldsymbol{h}} f\left(\widetilde{z}^t\right) - \nabla_{\boldsymbol{h}} f\left(z^\star\right) \\ \nabla_{\boldsymbol{x}} f\left(\widetilde{z}^t\right) - \nabla_{\boldsymbol{x}} f\left(z^\star\right) \end{bmatrix} = \underbrace{\int_0^1 \nabla^2 f\left(z\left(\tau\right)\right) \mathrm{d}\tau}_{:=\boldsymbol{A}} \begin{bmatrix} \widetilde{\boldsymbol{h}}^t - \boldsymbol{h}^\star \\ \widetilde{\boldsymbol{x}}^t - \boldsymbol{x}^\star \\ \hline \widetilde{\boldsymbol{h}}^t - \boldsymbol{h}^\star \\ \widetilde{\boldsymbol{x}}^t - \boldsymbol{x}^\star \end{bmatrix}, \tag{207}$$

where we denote $z(\tau) := z^\star + \tau\left(\widetilde{z}^t - z^\star\right)$ and $\nabla^2 f$ is the Wirtinger Hessian. To further simplify notation, denote $\widehat{z}^{t+1} = \begin{bmatrix} \widehat{\boldsymbol{h}}^{t+1} \\ \widehat{\boldsymbol{x}}^{t+1} \end{bmatrix}$. Identity (207) allows us to rewrite (206) as

$$\begin{bmatrix} \widehat{z}^{t+1} - z^\star \\ \hline \widetilde{z}^{t+1} - z^\star \end{bmatrix} = (\boldsymbol{I} - \eta \boldsymbol{D}\boldsymbol{A}) \begin{bmatrix} \widetilde{z}^t - z^\star \\ \hline \widetilde{z}^t - z^\star \end{bmatrix}. \tag{208}$$

Take the squared Euclidean norm of both sides of (208) to reach

$$\begin{aligned} \left\|\widehat{z}^{t+1} - z^\star\right\|_2^2 &= \frac{1}{2}\begin{bmatrix} \widetilde{z}^t - z^\star \\ \hline \widetilde{z}^t - z^\star \end{bmatrix}^{\mathsf{H}} (\boldsymbol{I} - \eta \boldsymbol{D}\boldsymbol{A})^{\mathsf{H}} (\boldsymbol{I} - \eta \boldsymbol{D}\boldsymbol{A}) \begin{bmatrix} \widetilde{z}^t - z^\star \\ \hline \widetilde{z}^t - z^\star \end{bmatrix} \\ &= \frac{1}{2}\begin{bmatrix} \widetilde{z}^t - z^\star \\ \hline \widetilde{z}^t - z^\star \end{bmatrix}^{\mathsf{H}} \left(\boldsymbol{I} + \eta^2 \boldsymbol{A}\boldsymbol{D}^2\boldsymbol{A} - \eta\left(\boldsymbol{D}\boldsymbol{A} + \boldsymbol{A}\boldsymbol{D}\right)\right) \begin{bmatrix} \widetilde{z}^t - z^\star \\ \hline \widetilde{z}^t - z^\star \end{bmatrix} \\ &\le \left(1 + \eta^2 \|\boldsymbol{A}\|^2 \|\boldsymbol{D}\|^2\right) \left\|\widetilde{z}^t - z^\star\right\|_2^2 - \frac{\eta}{2}\begin{bmatrix} \widetilde{z}^t - z^\star \\ \hline \widetilde{z}^t - z^\star \end{bmatrix}^{\mathsf{H}} \left(\boldsymbol{D}\boldsymbol{A} + \boldsymbol{A}\boldsymbol{D}\right) \begin{bmatrix} \widetilde{z}^t - z^\star \\ \hline \widetilde{z}^t - z^\star \end{bmatrix}. \end{aligned} \tag{209}$$

Since $z(\tau)$ lies between $\widetilde{z}^t$ and $z^\star$, we conclude from assumptions (85) that for all $0 \le \tau \le 1$,

$$\begin{aligned} \max\left\{\left\|\boldsymbol{h}(\tau) - \boldsymbol{h}^\star\right\|_2, \left\|\boldsymbol{x}(\tau) - \boldsymbol{x}^\star\right\|_2\right\} &\le \operatorname{dist}\left(z^t, z^\star\right) \le \xi \le \delta; \\ \max_{1\le j\le m} \left|\boldsymbol{a}_j^{\mathsf{H}}\left(\boldsymbol{x}(\tau) - \boldsymbol{x}^\star\right)\right| &\le C_3 \frac{1}{\log^{3/2} m}; \\ \max_{1\le j\le m} \left|\boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}(\tau)\right| &\le C_4 \frac{\mu}{\sqrt{m}} \log^2 m \end{aligned}$$

for $\xi > 0$ sufficiently small. Moreover, it is straightforward to see that

$$\gamma_1 := \left\|\widetilde{\boldsymbol{x}}^t\right\|_2^{-2} \qquad \text{and} \qquad \gamma_2 := \left\|\widetilde{\boldsymbol{h}}^t\right\|_2^{-2}$$

satisfy

$$\max\left\{\left|\gamma_1 - 1\right|, \left|\gamma_2 - 1\right|\right\} \lesssim \max\left\{\left\|\widetilde{\boldsymbol{h}}^t - \boldsymbol{h}^\star\right\|_2, \left\|\widetilde{\boldsymbol{x}}^t - \boldsymbol{x}^\star\right\|_2\right\} \le \delta$$

as long as $\xi > 0$ is sufficiently small. We can now readily invoke Lemma 14 to arrive at

$$\|\boldsymbol{A}\|\,\|\boldsymbol{D}\| \le 3(1+\delta) \le 4 \qquad \text{and}$$
$$\begin{bmatrix} \widetilde{z}^t - z^\star \\ \widetilde{z}^t - z^\star \end{bmatrix}^{\mathsf{H}} (\boldsymbol{DA} + \boldsymbol{AD}) \begin{bmatrix} \widetilde{z}^t - z^\star \\ \widetilde{z}^t - z^\star \end{bmatrix} \ge \frac{1}{4} \left\| \begin{bmatrix} \widetilde{z}^t - z^\star \\ \widetilde{z}^t - z^\star \end{bmatrix} \right\|_2^2 = \frac{1}{2} \left\| \widetilde{z}^t - z^\star \right\|_2^2.$$

Substitution into (209) indicates that

$$\left\| \widehat{z}^{t+1} - z^\star \right\|_2^2 \le \left(1 + 16\eta^2 - \eta/4\right) \left\| \widetilde{z}^t - z^\star \right\|_2^2.$$

When $0 < \eta \le 1/128$, this implies that

$$\left\| \widehat{z}^t - z^\star \right\|_2^2 \le (1 - \eta/8) \left\| \widetilde{z}^t - z^\star \right\|_2^2,$$

and hence

$$\left\| \widetilde{z}^{t+1} - z^\star \right\|_2 \le \left\| \widehat{z}^{t+1} - z^\star \right\|_2 \le (1-\eta/8)^{1/2} \left\| \widetilde{z}^t - z^\star \right\|_2 \le (1 - \eta/16)\,\mathrm{dist}(z^t, z^\star). \tag{210}$$

This completes the proof of Lemma 15. □

Proof of Lemma 16 Reuse the notation in this subsection, namely $\widehat{z}^{t+1} = \begin{bmatrix} \widehat{\boldsymbol{h}}^{t+1} \\ \widehat{\boldsymbol{x}}^{t+1} \end{bmatrix}$ with $\widehat{\boldsymbol{h}}^{t+1} = \frac{1}{\overline{\alpha}^t} \boldsymbol{h}^{t+1}$ and $\widehat{\boldsymbol{x}}^{t+1} = \alpha^t \boldsymbol{x}^{t+1}$. From (210), one can tell that

$$\left\| \widetilde{z}^{t+1} - z^\star \right\|_2 \le \left\| \widehat{z}^{t+1} - z^\star \right\|_2 \le \mathrm{dist}(z^t, z^\star).$$

Invoke Lemma 52 with $\beta = \alpha^t$ to get

$$\left| \alpha^{t+1} - \alpha^t \right| \lesssim \left\| \widehat{z}^{t+1} - z^\star \right\|_2 \le \mathrm{dist}(z^t, z^\star).$$

This combined with the assumption $||\alpha^t| - 1| \le 1/2$ implies that

$$\left|\alpha^t\right| \ge \frac{1}{2} \qquad \text{and} \qquad \left| \frac{\alpha^{t+1}}{\alpha^t} - 1 \right| = \left| \frac{\alpha^{t+1} - \alpha^t}{\alpha^t} \right| \lesssim \mathrm{dist}(z^t, z^\star) \lesssim C_1 \frac{1}{\log^2 m}.$$

This finishes the proof of the first claim.

The second claim can be proved by induction. Suppose that $\left||\alpha^s| - 1\right| \le 1/2$ and $\mathrm{dist}(z^s, z^\star) \le C_1(1-\eta/16)^s/\log^2 m$ hold for all $0 \le s \le \tau \le t$, then using our result in the first part gives

$$\begin{aligned}\left||\alpha^{\tau+1}|-1\right| &\le \left||\alpha^{0}|-1\right| + \sum_{s=0}^{\tau}\left|\alpha^{s+1}-\alpha^{s}\right| \le \frac{1}{4} + c\sum_{s=0}^{\tau}\mathrm{dist}(z^{s}, z^{\star}) \\ &\le \frac{1}{4} + \frac{cC_1}{\frac{\eta}{16}\log^2 m} \le \frac{1}{2}\end{aligned}$$

for m sufficiently large. The proof is then complete by induction. □

C.3 Proof of Lemma 17

Define the alignment parameter between $z^{t,(l)}$ and $\widetilde{z}^{t}$ as

$$\alpha_{\text{mutual}}^{t,(l)} := \operatorname*{argmin}_{\alpha\in\mathbb{C}} \left\| \frac{1}{\overline{\alpha}}\boldsymbol{h}^{t,(l)} - \frac{1}{\alpha^{t}}\boldsymbol{h}^{t} \right\|_2^2 + \left\| \alpha \boldsymbol{x}^{t,(l)} - \alpha^{t}\boldsymbol{x}^{t} \right\|_2^2.$$

Further denote, for simplicity of presentation, $\widehat{z}^{t,(l)} = \begin{bmatrix} \widehat{\boldsymbol{h}}^{t,(l)} \\ \widehat{\boldsymbol{x}}^{t,(l)} \end{bmatrix}$ with

$$\widehat{\boldsymbol{h}}^{t,(l)} := \frac{1}{\overline{\alpha_{\text{mutual}}^{t,(l)}}}\boldsymbol{h}^{t,(l)} \qquad \text{and} \qquad \widehat{\boldsymbol{x}}^{t,(l)} := \alpha_{\text{mutual}}^{t,(l)}\boldsymbol{x}^{t,(l)}.$$

Clearly, $\widehat{z}^{t,(l)}$ is aligned with $\widetilde{z}^{t}$.

Armed with the above notation, we have

$$\begin{aligned}
\mathrm{dist}\big(z^{t+1,(l)}, \widetilde{z}^{t+1}\big) &= \min_{\alpha} \sqrt{\left\| \frac{1}{\overline{\alpha}}\boldsymbol{h}^{t+1,(l)} - \frac{1}{\overline{\alpha^{t+1}}}\boldsymbol{h}^{t+1} \right\|_2^2 + \left\| \alpha\boldsymbol{x}^{t+1,(l)} - \alpha^{t+1}\boldsymbol{x}^{t+1} \right\|_2^2} \\
&= \min_{\alpha} \sqrt{\left\| \left(\frac{\overline{\alpha^{t}}}{\overline{\alpha^{t+1}}}\right)\left(\frac{1}{\overline{\alpha}}\frac{\overline{\alpha^{t+1}}}{\overline{\alpha^{t}}}\boldsymbol{h}^{t+1,(l)} - \frac{1}{\overline{\alpha^{t}}}\boldsymbol{h}^{t+1}\right) \right\|_2^2 + \left\| \left(\frac{\alpha^{t+1}}{\alpha^{t}}\right)\left(\alpha\frac{\alpha^{t}}{\alpha^{t+1}}\boldsymbol{x}^{t+1,(l)} - \alpha^{t}\boldsymbol{x}^{t+1}\right) \right\|_2^2} \\
&\le \sqrt{\left\| \left(\frac{\overline{\alpha^{t}}}{\overline{\alpha^{t+1}}}\right)\left(\frac{1}{\overline{\alpha_{\text{mutual}}^{t,(l)}}}\boldsymbol{h}^{t+1,(l)} - \frac{1}{\overline{\alpha^{t}}}\boldsymbol{h}^{t+1}\right) \right\|_2^2 + \left\| \left(\frac{\alpha^{t+1}}{\alpha^{t}}\right)\left(\alpha_{\text{mutual}}^{t,(l)}\boldsymbol{x}^{t+1,(l)} - \alpha^{t}\boldsymbol{x}^{t+1}\right) \right\|_2^2} \qquad (211) \\
&\le \max\left\{ \left|\frac{\alpha^{t+1}}{\alpha^{t}}\right|, \left|\frac{\alpha^{t}}{\alpha^{t+1}}\right| \right\} \left\| \begin{bmatrix} \frac{1}{\overline{\alpha_{\text{mutual}}^{t,(l)}}}\boldsymbol{h}^{t+1,(l)} - \frac{1}{\overline{\alpha^{t}}}\boldsymbol{h}^{t+1} \\ \alpha_{\text{mutual}}^{t,(l)}\boldsymbol{x}^{t+1,(l)} - \alpha^{t}\boldsymbol{x}^{t+1} \end{bmatrix} \right\|_2, \qquad (212)
\end{aligned}$$

where (211) follows by taking $\alpha = \frac{\alpha^{t+1}}{\alpha^{t}}\alpha_{\text{mutual}}^{t,(l)}$. The latter bound is more convenient to work with when controlling the gap between $z^{t,(l)}$ and z^{t}.

We can then apply the gradient update rules (79) and (89) to get

$$\begin{bmatrix} \frac{1}{\overline{\alpha_{\text{mutual}}^{t,(l)}}}\boldsymbol{h}^{t+1,(l)} - \frac{1}{\overline{\alpha^{t}}}\boldsymbol{h}^{t+1} \\ \alpha_{\text{mutual}}^{t,(l)}\boldsymbol{x}^{t+1,(l)} - \alpha^{t}\boldsymbol{x}^{t+1} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{\alpha_{\text{mutual}}^{t,(l)}} \left(\boldsymbol{h}^{t,(l)} - \frac{\eta}{\|\boldsymbol{x}^{t,(l)}\|_2^2} \nabla_{\boldsymbol{h}} f^{(l)} \left(\boldsymbol{h}^{t,(l)}, \boldsymbol{x}^{t,(l)} \right) \right) - \frac{1}{\alpha^t} \left(\boldsymbol{h}^t - \frac{\eta}{\|\boldsymbol{x}^t\|_2^2} \nabla_{\boldsymbol{h}} f \left(\boldsymbol{h}^t, \boldsymbol{x}^t \right) \right) \\ \alpha_{\text{mutual}}^{t,(l)} \left(\boldsymbol{x}^{t,(l)} - \frac{\eta}{\left\|\boldsymbol{h}^{t,(l)}\right\|_2^2} \nabla_{\boldsymbol{x}} f^{(l)} \left(\boldsymbol{h}^{t,(l)}, \boldsymbol{x}^{t,(l)} \right) \right) - \alpha^t \left(\boldsymbol{x}^t - \frac{\eta}{\|\boldsymbol{h}^t\|_2^2} \nabla_{\boldsymbol{x}} f \left(\boldsymbol{h}^t, \boldsymbol{x}^t \right) \right) \end{bmatrix}$$

$$= \begin{bmatrix} \widehat{\boldsymbol{h}}^{t,(l)} - \frac{\eta}{\|\widehat{\boldsymbol{x}}^{t,(l)}\|_2^2} \nabla_{\boldsymbol{h}} f^{(l)} \left(\widehat{\boldsymbol{h}}^{t,(l)}, \widehat{\boldsymbol{x}}^{t,(l)} \right) - \left(\widetilde{\boldsymbol{h}}^t - \frac{\eta}{\|\widetilde{\boldsymbol{x}}^t\|_2^2} \nabla_{\boldsymbol{h}} f \left(\widetilde{\boldsymbol{h}}^t, \widetilde{\boldsymbol{x}}^t \right) \right) \\ \widehat{\boldsymbol{x}}^{t,(l)} - \frac{\eta}{\|\widehat{\boldsymbol{h}}^{t,(l)}\|_2^2} \nabla_{\boldsymbol{x}} f^{(l)} \left(\widehat{\boldsymbol{h}}^{t,(l)}, \widehat{\boldsymbol{x}}^{t,(l)} \right) - \left(\widetilde{\boldsymbol{x}}^t - \frac{\eta}{\|\widetilde{\boldsymbol{h}}^t\|_2^2} \nabla_{\boldsymbol{x}} f \left(\widetilde{\boldsymbol{h}}^t, \widetilde{\boldsymbol{x}}^t \right) \right) \end{bmatrix}.$$

By construction, we can write the leave-one-out gradients as

$$\nabla_{\boldsymbol{h}} f^{(l)} (\boldsymbol{h}, \boldsymbol{x}) = \nabla_{\boldsymbol{h}} f (\boldsymbol{h}, \boldsymbol{x}) - \left(\boldsymbol{b}_l^{\mathsf{H}} \boldsymbol{h} \boldsymbol{x}^{\mathsf{H}} \boldsymbol{a}_l - y_l \right) \boldsymbol{b}_l \boldsymbol{a}_l^{\mathsf{H}} \boldsymbol{x} \qquad \text{and}$$
$$\nabla_{\boldsymbol{x}} f^{(l)} (\boldsymbol{h}, \boldsymbol{x}) = \nabla_{\boldsymbol{h}} f (\boldsymbol{h}, \boldsymbol{x}) - \overline{(\boldsymbol{b}_l^{\mathsf{H}} \boldsymbol{h} \boldsymbol{x}^{\mathsf{H}} \boldsymbol{a}_l - y_l)} \boldsymbol{a}_l \boldsymbol{b}_l^{\mathsf{H}} \boldsymbol{h},$$

which allow us to continue the derivation and obtain

$$\begin{bmatrix} \frac{1}{\alpha_{\text{mutual}}^{t,(l)}} \boldsymbol{h}^{t+1,(l)} - \frac{1}{\alpha^t} \boldsymbol{h}^{t+1} \\ \alpha_{\text{mutual}}^{t,(l)} \boldsymbol{x}^{t+1,(l)} - \alpha^t \boldsymbol{x}^{t+1} \end{bmatrix}$$
$$= \begin{bmatrix} \widehat{\boldsymbol{h}}^{t,(l)} - \frac{\eta}{\|\widehat{\boldsymbol{x}}^{t,(l)}\|_2^2} \nabla_{\boldsymbol{h}} f \left(\widehat{\boldsymbol{h}}^{t,(l)}, \widehat{\boldsymbol{x}}^{t,(l)} \right) - \left(\widetilde{\boldsymbol{h}}^t - \frac{\eta}{\|\widetilde{\boldsymbol{x}}^t\|_2^2} \nabla_{\boldsymbol{h}} f \left(\widetilde{\boldsymbol{h}}^t, \widetilde{\boldsymbol{x}}^t \right) \right) \\ \widehat{\boldsymbol{x}}^{t,(l)} - \frac{\eta}{\|\widehat{\boldsymbol{h}}^{t,(l)}\|_2^2} \nabla_{\boldsymbol{x}} f \left(\widehat{\boldsymbol{h}}^{t,(l)}, \widehat{\boldsymbol{x}}^{t,(l)} \right) - \left(\widetilde{\boldsymbol{x}}^t - \frac{\eta}{\|\widetilde{\boldsymbol{h}}^t\|_2^2} \nabla_{\boldsymbol{x}} f \left(\widetilde{\boldsymbol{h}}^t, \widetilde{\boldsymbol{x}}^t \right) \right) \end{bmatrix}$$
$$- \eta \underbrace{\begin{bmatrix} \frac{1}{\|\widehat{\boldsymbol{x}}^{t,(l)}\|_2^2} \left(\boldsymbol{b}_l^{\mathsf{H}} \widehat{\boldsymbol{h}}^{t,(l)} \widehat{\boldsymbol{x}}^{t,(l)\mathsf{H}} \boldsymbol{a}_l - y_l \right) \boldsymbol{b}_l \boldsymbol{a}_l^{\mathsf{H}} \widehat{\boldsymbol{x}}^{t,(l)} \\ \frac{1}{\|\widehat{\boldsymbol{h}}^{t,(l)}\|_2^2} \overline{\left(\boldsymbol{b}_l^{\mathsf{H}} \widehat{\boldsymbol{h}}^{t,(l)} \widehat{\boldsymbol{x}}^{t,(l)\mathsf{H}} \boldsymbol{a}_l - y_l \right)} \boldsymbol{a}_l \boldsymbol{b}_l^{\mathsf{H}} \widehat{\boldsymbol{h}}^{t,(l)} \end{bmatrix}}_{:=\boldsymbol{J}_3}.$$

This further gives

$$\begin{bmatrix} \frac{1}{\alpha_{\text{mutual}}^{t,(l)}} \boldsymbol{h}^{t+1,(l)} - \frac{1}{\alpha^t} \boldsymbol{h}^{t+1} \\ \alpha_{\text{mutual}}^{t,(l)} \boldsymbol{x}^{t+1,(l)} - \alpha^t \boldsymbol{x}^{t+1} \end{bmatrix}$$
$$= \underbrace{\begin{bmatrix} \widehat{\boldsymbol{h}}^{t,(l)} - \frac{\eta}{\|\widehat{\boldsymbol{x}}^{t,(l)}\|_2^2} \nabla_{\boldsymbol{h}} f \left(\widehat{\boldsymbol{h}}^{t,(l)}, \widehat{\boldsymbol{x}}^{t,(l)} \right) - \left(\widetilde{\boldsymbol{h}}^t - \frac{\eta}{\|\widehat{\boldsymbol{x}}^{t,(l)}\|_2^2} \nabla_{\boldsymbol{h}} f \left(\widetilde{\boldsymbol{h}}^t, \widetilde{\boldsymbol{x}}^t \right) \right) \\ \widehat{\boldsymbol{x}}^{t,(l)} - \frac{\eta}{\left\|\widehat{\boldsymbol{h}}^{t,(l)}\right\|_2^2} \nabla_{\boldsymbol{x}} f \left(\widehat{\boldsymbol{h}}^{t,(l)}, \widehat{\boldsymbol{x}}^{t,(l)} \right) - \left(\widetilde{\boldsymbol{x}}^t - \frac{\eta}{\left\|\widehat{\boldsymbol{h}}^{t,(l)}\right\|_2^2} \nabla_{\boldsymbol{x}} f \left(\widetilde{\boldsymbol{h}}^t, \widetilde{\boldsymbol{x}}^t \right) \right) \end{bmatrix}}_{:=\boldsymbol{v}_1}$$

$$+\eta \underbrace{\begin{bmatrix} \left(\frac{1}{\|\widetilde{\boldsymbol{x}}^t\|_2^2}-\frac{1}{\|\widehat{\boldsymbol{x}}^{t,(l)}\|_2^2}\right)\nabla_{\boldsymbol{h}} f\left(\widetilde{\boldsymbol{h}}^t,\widetilde{\boldsymbol{x}}^t\right) \\ \left(\frac{1}{\|\widetilde{\boldsymbol{h}}^t\|_2^2}-\frac{1}{\|\widehat{\boldsymbol{h}}^{t,(l)}\|_2^2}\right)\nabla_{\boldsymbol{x}} f\left(\widetilde{\boldsymbol{h}}^t,\widetilde{\boldsymbol{x}}^t\right) \end{bmatrix}}_{:=\boldsymbol{v}_2} -\eta\boldsymbol{v}_3. \tag{213}$$

In what follows, we bound the three terms $\boldsymbol{v}_1$, $\boldsymbol{v}_2$, and $\boldsymbol{v}_3$ separately.

1. Regarding the first term $\boldsymbol{v}_1$, one can adopt the same strategy as in Appendix C.2. Specifically, write

$$\begin{bmatrix} \widehat{\boldsymbol{h}}^{t,(l)}-\frac{\eta}{\|\widehat{\boldsymbol{x}}^{t,(l)}\|_2^2}\nabla_{\boldsymbol{h}} f\left(\widehat{\boldsymbol{z}}^{t,(l)}\right)-\left(\widetilde{\boldsymbol{h}}^t-\frac{\eta}{\|\widetilde{\boldsymbol{x}}^{t,(l)}\|_2^2}\nabla_{\boldsymbol{h}} f\left(\widetilde{\boldsymbol{z}}^t\right)\right) \\ \widehat{\boldsymbol{x}}^{t,(l)}-\frac{\eta}{\|\widehat{\boldsymbol{h}}^{t,(l)}\|_2^2}\nabla_{\boldsymbol{x}} f\left(\widehat{\boldsymbol{z}}^{t,(l)}\right)-\left(\widetilde{\boldsymbol{x}}^t-\frac{\eta}{\|\widetilde{\boldsymbol{h}}^{t,(l)}\|_2^2}\nabla_{\boldsymbol{x}} f\left(\widetilde{\boldsymbol{z}}^t\right)\right) \\ \widehat{\boldsymbol{h}}^{t,(l)}-\frac{\eta}{\|\widehat{\boldsymbol{x}}^{t,(l)}\|_2^2}\nabla_{\boldsymbol{h}} f\left(\widehat{\boldsymbol{z}}^{t,(l)}\right)-\left(\widetilde{\boldsymbol{h}}^t-\frac{\eta}{\|\widetilde{\boldsymbol{x}}^{t,(l)}\|_2^2}\nabla_{\boldsymbol{h}} f\left(\widetilde{\boldsymbol{z}}^t\right)\right) \\ \widehat{\boldsymbol{x}}^{t,(l)}-\frac{\eta}{\|\widehat{\boldsymbol{h}}^{t,(l)}\|_2^2}\nabla_{\boldsymbol{x}} f\left(\widehat{\boldsymbol{z}}^{t,(l)}\right)-\left(\widetilde{\boldsymbol{x}}^t-\frac{\eta}{\|\widehat{\boldsymbol{h}}^{t,(l)}\|_2^2}\nabla_{\boldsymbol{x}} f\left(\widetilde{\boldsymbol{z}}^t\right)\right) \end{bmatrix} = \begin{bmatrix} \widehat{\boldsymbol{h}}^{t,(l)}-\widetilde{\boldsymbol{h}}^t \\ \widehat{\boldsymbol{x}}^{t,(l)}-\widetilde{\boldsymbol{x}}^t \\ \widehat{\boldsymbol{h}}^{t,(l)}-\widetilde{\boldsymbol{h}}^t \\ \widehat{\boldsymbol{x}}^{t,(l)}-\widetilde{\boldsymbol{x}}^t \end{bmatrix}$$

$$-\eta \underbrace{\begin{bmatrix} \left\|\widehat{\boldsymbol{x}}^{t,(l)}\right\|_2^{-2}\boldsymbol{I}_K & & & \\ & \left\|\widehat{\boldsymbol{h}}^{t,(l)}\right\|_2^{-2}\boldsymbol{I}_K & & \\ & & \left\|\widehat{\boldsymbol{x}}^{t,(l)}\right\|_2^{-2}\boldsymbol{I}_K & \\ & & & \left\|\widehat{\boldsymbol{h}}^{t,(l)}\right\|_2^{-2}\boldsymbol{I}_K \end{bmatrix}}_{:=\boldsymbol{D}}$$

$$\begin{bmatrix} \nabla_{\boldsymbol{h}} f\left(\widehat{\boldsymbol{z}}^{t,(l)}\right)-\nabla_{\boldsymbol{h}} f\left(\widetilde{\boldsymbol{z}}^t\right) \\ \nabla_{\boldsymbol{x}} f\left(\widehat{\boldsymbol{z}}^{t,(l)}\right)-\nabla_{\boldsymbol{x}} f\left(\widetilde{\boldsymbol{z}}^t\right) \\ \nabla_{\boldsymbol{h}} f\left(\widehat{\boldsymbol{z}}^{t,(l)}\right)-\nabla_{\boldsymbol{h}} f\left(\widetilde{\boldsymbol{z}}^t\right) \\ \nabla_{\boldsymbol{x}} f\left(\widehat{\boldsymbol{z}}^{t,(l)}\right)-\nabla_{\boldsymbol{x}} f\left(\widetilde{\boldsymbol{z}}^t\right) \end{bmatrix}.$$

The fundamental theorem of calculus (see Appendix D.3.1) reveals that

$$\begin{bmatrix} \nabla_{\boldsymbol{h}} f\left(\widehat{\boldsymbol{z}}^{t,(l)}\right)-\nabla_{\boldsymbol{h}} f\left(\widetilde{\boldsymbol{z}}^t\right) \\ \nabla_{\boldsymbol{x}} f\left(\widehat{\boldsymbol{z}}^{t,(l)}\right)-\nabla_{\boldsymbol{x}} f\left(\widetilde{\boldsymbol{z}}^t\right) \\ \nabla_{\boldsymbol{h}} f\left(\widehat{\boldsymbol{z}}^{t,(l)}\right)-\nabla_{\boldsymbol{h}} f\left(\widetilde{\boldsymbol{z}}^t\right) \\ \nabla_{\boldsymbol{x}} f\left(\widehat{\boldsymbol{z}}^{t,(l)}\right)-\nabla_{\boldsymbol{x}} f\left(\widetilde{\boldsymbol{z}}^t\right) \end{bmatrix} = \underbrace{\int_0^1 \nabla^2 f\left(\boldsymbol{z}\left(\tau\right)\right)\mathrm{d}\tau}_{:=\boldsymbol{A}} \begin{bmatrix} \widehat{\boldsymbol{h}}^{t,(l)}-\widetilde{\boldsymbol{h}}^t \\ \widehat{\boldsymbol{x}}^{t,(l)}-\widetilde{\boldsymbol{x}}^t \\ \widehat{\boldsymbol{h}}^{t,(l)}-\widetilde{\boldsymbol{h}}^t \\ \widehat{\boldsymbol{x}}^{t,(l)}-\widetilde{\boldsymbol{x}}^t \end{bmatrix},$$

where we abuse the notation and denote $\boldsymbol{z}\left(\tau\right)=\widetilde{\boldsymbol{z}}^t+\tau\left(\widehat{\boldsymbol{z}}^{t,(l)}-\widetilde{\boldsymbol{z}}^t\right)$. In order to invoke Lemma 14, we need to verify the conditions required therein. Recall the induction hypothesis (90b) that

$$\mathrm{dist}\left(\boldsymbol{z}^{t,(l)},\widetilde{\boldsymbol{z}}^t\right)=\left\|\widehat{\boldsymbol{z}}^{t,(l)}-\widetilde{\boldsymbol{z}}^t\right\|_2 \le C_2\frac{\mu}{\sqrt{m}}\sqrt{\frac{\mu^2 K\log^9 m}{m}},$$

and the fact that $z(\tau)$ lies between $\widehat{z}^{t,(l)}$ and $\widetilde{z}^t$. For all $0 \le \tau \le 1$:

(a) If $m \gg \mu^2 \sqrt{K} \log^{13/2} m$, then

$$\begin{aligned}\left\|z(\tau) - z^\star\right\|_2 &\le \max\left\{\left\|\widehat{z}^{t,(l)} - z^\star\right\|_2, \left\|\widetilde{z}^t - z^\star\right\|_2\right\} \le \left\|\widetilde{z}^t - z^\star\right\|_2 + \left\|\widehat{z}^{t,(l)} - \widetilde{z}^t\right\|_2 \\ &\le C_1 \frac{1}{\log^2 m} + C_2 \frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^9 m}{m}} \le 2C_1 \frac{1}{\log^2 m},\end{aligned}$$

where we have used the induction hypotheses (90a) and (90b);

(b) If $m \gg \mu^2 K \log^6 m$, then

$$\begin{aligned}&\max_{1\le j\le m} \left|\boldsymbol{a}_j^{\mathsf{H}}\left(\boldsymbol{x}(\tau) - \boldsymbol{x}^\star\right)\right| \\ &= \max_{1\le j\le m} \left|\tau \boldsymbol{a}_j^{\mathsf{H}}\left(\widehat{\boldsymbol{x}}^{t,(l)} - \widetilde{\boldsymbol{x}}^t\right) + \boldsymbol{a}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^t - \boldsymbol{x}^\star\right)\right| \\ &\le \max_{1\le j\le m} \left|\boldsymbol{a}_j^{\mathsf{H}}\left(\widehat{\boldsymbol{x}}^{t,(l)} - \widetilde{\boldsymbol{x}}^t\right)\right| + \max_{1\le j\le m} \left|\boldsymbol{a}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^t - \boldsymbol{x}^\star\right)\right| \\ &\le \max_{1\le j\le m} \left\|\boldsymbol{a}_j\right\|_2 \left\|\widehat{\boldsymbol{z}}^{t,(l)} - \widetilde{\boldsymbol{z}}^t\right\|_2 + C_3 \frac{1}{\log^{3/2} m} \\ &\le 3\sqrt{K} \cdot C_2 \frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^9 m}{m}} + C_3 \frac{1}{\log^{3/2} m} \le 2C_3 \frac{1}{\log^{3/2} m}, \end{aligned} \tag{214}$$

which follows from bound (190) and the induction hypotheses (90b) and (90c);

(c) If $m \gg \mu K \log^{5/2} m$, then

$$\begin{aligned}\max_{1\le j\le m} \left|\boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}(\tau)\right| &= \max_{1\le j\le m} \left|\tau \boldsymbol{b}_j^{\mathsf{H}}\left(\widehat{\boldsymbol{h}}^{t,(l)} - \widetilde{\boldsymbol{h}}^t\right) + \boldsymbol{b}_j^{\mathsf{H}} \widetilde{\boldsymbol{h}}^t\right| \\ &\le \max_{1\le j\le m} \left|\boldsymbol{b}_j^{\mathsf{H}}\left(\widehat{\boldsymbol{h}}^{t,(l)} - \widetilde{\boldsymbol{h}}^t\right)\right| + \max_{1\le j\le m} \left|\boldsymbol{b}_j^{\mathsf{H}} \widetilde{\boldsymbol{h}}^t\right| \\ &\le \max_{1\le j\le m} \left\|\boldsymbol{b}_j\right\|_2 \left\|\widehat{\boldsymbol{h}}^{t,(l)} - \widetilde{\boldsymbol{h}}^t\right\|_2 + \max_{1\le j\le m} \left|\boldsymbol{b}_j^{\mathsf{H}} \widetilde{\boldsymbol{h}}^t\right| \\ &\le \sqrt{\frac{K}{m}} \cdot C_2 \frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^9 m}{m}} + C_4 \frac{\mu}{\sqrt{m}} \log^2 m \le 2C_4 \frac{\mu}{\sqrt{m}} \log^2 m, \end{aligned} \tag{215}$$

which makes use of the fact $\|\boldsymbol{b}_j\|_2 = \sqrt{K/m}$ as well as the induction hypotheses (90b) and (90d).

These properties satisfy condition (82) required in Lemma 14. The other two conditions (83) and (84) are also straightforward to check, and hence, we omit it. Thus, we can repeat the argument used in Appendix C.2 to obtain

$$\|\boldsymbol{v}_1\|_2 \le (1 - \eta/16) \cdot \left\|\widehat{\boldsymbol{z}}^{t,(l)} - \widetilde{\boldsymbol{z}}^t\right\|_2.$$

2. In terms of the second term $\boldsymbol{v}_2$, it is easily seen that

$$\|\boldsymbol{v}_2\|_2 \le \max\left\{\left|\frac{1}{\|\widetilde{\boldsymbol{x}}^t\|_2^2}-\frac{1}{\|\widehat{\boldsymbol{x}}^{t,(l)}\|_2^2}\right|,\left|\frac{1}{\|\widetilde{\boldsymbol{h}}^t\|_2^2}-\frac{1}{\|\widehat{\boldsymbol{h}}^{t,(l)}\|_2^2}\right|\right\}\left\|\begin{bmatrix}\nabla_{\boldsymbol{h}} f\left(\widetilde{\boldsymbol{z}}^t\right)\\ \nabla_{\boldsymbol{x}} f\left(\widetilde{\boldsymbol{z}}^t\right)\end{bmatrix}\right\|_2.$$

We first note that the upper bound on $\|\nabla^2 f(\cdot)\|$ (which essentially provides a Lipschitz constant on the gradient) in Lemma 14 forces

$$\left\|\begin{bmatrix}\nabla_{\boldsymbol{h}} f\left(\widetilde{\boldsymbol{z}}^t\right)\\ \nabla_{\boldsymbol{x}} f\left(\widetilde{\boldsymbol{z}}^t\right)\end{bmatrix}\right\|_2 = \left\|\begin{bmatrix}\nabla_{\boldsymbol{h}} f\left(\widetilde{\boldsymbol{z}}^t\right)-\nabla_{\boldsymbol{h}} f\left(\boldsymbol{z}^\star\right)\\ \nabla_{\boldsymbol{x}} f\left(\widetilde{\boldsymbol{z}}^t\right)-\nabla_{\boldsymbol{x}} f\left(\boldsymbol{z}^\star\right)\end{bmatrix}\right\|_2 \lesssim \left\|\widetilde{\boldsymbol{z}}^t-\boldsymbol{z}^\star\right\|_2 \lesssim C_1\frac{1}{\log^2 m},$$

where the first identity follows since $\nabla_{\boldsymbol{h}} f\left(\boldsymbol{z}^\star\right)=\boldsymbol{0}$, and the last inequality comes from the induction hypothesis (90a). Additionally, recognizing that $\left\|\widetilde{\boldsymbol{x}}^t\right\|_2 \asymp \left\|\widehat{\boldsymbol{x}}^{t,(l)}\right\|_2 \asymp 1$, one can easily verify that

$$\left|\frac{1}{\|\widetilde{\boldsymbol{x}}^t\|_2^2}-\frac{1}{\|\widehat{\boldsymbol{x}}^{t,(l)}\|_2^2}\right| = \left|\frac{\|\widehat{\boldsymbol{x}}^{t,(l)}\|_2^2-\|\widetilde{\boldsymbol{x}}^t\|_2^2}{\|\widetilde{\boldsymbol{x}}^t\|_2^2\cdot\|\widehat{\boldsymbol{x}}^{t,(l)}\|_2^2}\right| \lesssim \left|\left\|\widehat{\boldsymbol{x}}^{t,(l)}\right\|_2-\left\|\widetilde{\boldsymbol{x}}^t\right\|_2\right| \lesssim \left\|\widehat{\boldsymbol{x}}^{t,(l)}-\widetilde{\boldsymbol{x}}^t\right\|_2.$$

A similar bound holds for the other term involving $\boldsymbol{h}$. Combining the estimates above thus yields

$$\|\boldsymbol{v}_2\|_2 \lesssim C_1\frac{1}{\log^2 m}\left\|\widehat{\boldsymbol{z}}^{t,(l)}-\widetilde{\boldsymbol{z}}^t\right\|_2.$$

3. When it comes to the last term $\boldsymbol{v}_3$, one first sees that

$$\left\|\left(\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\widehat{\boldsymbol{x}}^{t,(l)\mathsf{H}}\boldsymbol{a}_l-y_l\right)\boldsymbol{b}_l\boldsymbol{a}_l^{\mathsf{H}}\widehat{\boldsymbol{x}}^{t,(l)}\right\|_2 \le \left|\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\widehat{\boldsymbol{x}}^{t,(l)\mathsf{H}}\boldsymbol{a}_l-y_l\right|\|\boldsymbol{b}_l\|_2\left|\boldsymbol{a}_l^{\mathsf{H}}\widehat{\boldsymbol{x}}^{t,(l)}\right|. \tag{216}$$

Bounds (189) and (214) taken collectively yield

$$\left|\boldsymbol{a}_l^{\mathsf{H}}\widehat{\boldsymbol{x}}^{t,(l)}\right| \le \left|\boldsymbol{a}_l^{\mathsf{H}}\boldsymbol{x}^\star\right|+\left|\boldsymbol{a}_l^{\mathsf{H}}\left(\widehat{\boldsymbol{x}}^{t,(l)}-\boldsymbol{x}^\star\right)\right| \lesssim \sqrt{\log m}+C_3\frac{1}{\log^{3/2} m} \asymp \sqrt{\log m}.$$

In addition, the same argument as in obtaining (215) tells us that

$$\left|\boldsymbol{b}_l^{\mathsf{H}}(\widehat{\boldsymbol{h}}^{t,(l)}-\boldsymbol{h}^\star)\right| \lesssim C_4\frac{\mu}{\sqrt{m}}\log^2 m.$$

Combine the previous two bounds to obtain

$$\begin{aligned}\left|\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\widehat{\boldsymbol{x}}^{t,(l)\mathsf{H}}\boldsymbol{a}_l-y_l\right| &\le \left|\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}(\widehat{\boldsymbol{x}}^{t,(l)}-\boldsymbol{x}^\star)^{\mathsf{H}}\boldsymbol{a}_l\right|+\left|\boldsymbol{b}_l^{\mathsf{H}}(\widehat{\boldsymbol{h}}^{t,(l)}-\boldsymbol{h}^\star)\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_l\right|\\ &\le \left|\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\right|\cdot\left|\boldsymbol{a}_l^{\mathsf{H}}(\widehat{\boldsymbol{x}}^{t,(l)}-\boldsymbol{x}^\star)\right|+\left|\boldsymbol{b}_l^{\mathsf{H}}(\widehat{\boldsymbol{h}}^{t,(l)}-\boldsymbol{h}^\star)\right|\cdot\left|\boldsymbol{a}_l^{\mathsf{H}}\boldsymbol{x}^\star\right|\end{aligned}$$

$$\leq \left(\left|\boldsymbol{b}_l^{\mathsf{H}}(\widehat{\boldsymbol{h}}^{t,(l)}-\boldsymbol{h}^\star)\right|+\left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^\star\right|\right)\cdot\left|\boldsymbol{a}_l^{\mathsf{H}}(\widehat{\boldsymbol{x}}^{t,(l)}-\boldsymbol{x}^\star)\right|+\left|\boldsymbol{b}_l^{\mathsf{H}}(\widehat{\boldsymbol{h}}^{t,(l)}-\boldsymbol{h}^\star)\right|\cdot\left|\boldsymbol{a}_l^{\mathsf{H}}\boldsymbol{x}^\star\right|$$
$$\lesssim \left(C_4\mu\frac{\log^2 m}{\sqrt{m}}+\frac{\mu}{\sqrt{m}}\right)\cdot C_3\frac{1}{\log^{3/2} m}+C_4\mu\frac{\log^2 m}{\sqrt{m}}\cdot\sqrt{\log m}\lesssim C_4\mu\frac{\log^{5/2} m}{\sqrt{m}}.$$

Substitution into (216) gives

$$\left\|\left(\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\widehat{\boldsymbol{x}}^{t,(l)\mathsf{H}}\boldsymbol{a}_l-y_l\right)\boldsymbol{b}_l\boldsymbol{a}_l^{\mathsf{H}}\widehat{\boldsymbol{x}}^{t,(l)}\right\|_2\lesssim C_4\mu\frac{\log^{5/2} m}{\sqrt{m}}\cdot\sqrt{\frac{K}{m}}\cdot\sqrt{\log m}. \quad (217)$$

Similarly, we can also derive

$$\left\|\overline{\left(\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\widehat{\boldsymbol{x}}^{t,(l)\mathsf{H}}\boldsymbol{a}_l-y_l\right)}\boldsymbol{a}_l\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\right\|\leq\left|\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\widehat{\boldsymbol{x}}^{t,(l)\mathsf{H}}\boldsymbol{a}_l-y_l\right|\|\boldsymbol{a}_l\|_2\left|\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\right|$$
$$\lesssim C_4\mu\frac{\log^{5/2} m}{\sqrt{m}}\cdot\sqrt{K}\cdot C_4\frac{\mu}{\sqrt{m}}\log^2 m$$

Putting these bounds together indicates that

$$\|\boldsymbol{v}_3\|_2\lesssim (C_4)^2\frac{\mu}{\sqrt{m}}\sqrt{\frac{\mu^2K\log^9 m}{m}}.$$

The above bounds taken together with (212) and (213) ensure the existence of a constant $C>0$ such that

$$\begin{aligned}
\operatorname{dist}\left(\boldsymbol{z}^{t+1,(l)},\widetilde{\boldsymbol{z}}^{t+1}\right)&\leq\max\left\{\left|\frac{\alpha^{t+1}}{\alpha^t}\right|,\left|\frac{\alpha^t}{\alpha^{t+1}}\right|\right\}\left\{\left(1-\frac{\eta}{16}+CC_1\eta\frac{1}{\log^2 m}\right)\right.\\
&\quad\left.\left\|\widehat{\boldsymbol{z}}^{t,(l)}-\widetilde{\boldsymbol{z}}^t\right\|_2+C\,(C_4)^2\,\eta\frac{\mu}{\sqrt{m}}\sqrt{\frac{\mu^2K\log^9 m}{m}}\right\}\\
&\overset{\text{(i)}}{\leq}\frac{1-\eta/21}{1-\eta/20}\left\{\left(1-\frac{\eta}{20}\right)\left\|\widehat{\boldsymbol{z}}^{t,(l)}-\widetilde{\boldsymbol{z}}^t\right\|_2\right.\\
&\quad\left.+C\,(C_4)^2\,\eta\frac{\mu}{\sqrt{m}}\sqrt{\frac{\mu^2K\log^9 m}{m}}\right\}\\
&\leq\left(1-\frac{\eta}{21}\right)\left\|\widehat{\boldsymbol{z}}^{t,(l)}-\widetilde{\boldsymbol{z}}^t\right\|_2+2C\,(C_4)^2\,\eta\frac{\mu}{\sqrt{m}}\sqrt{\frac{\mu^2K\log^9 m}{m}}\\
&=\left(1-\frac{\eta}{21}\right)\operatorname{dist}\left(\boldsymbol{z}^{t,(l)},\widetilde{\boldsymbol{z}}^t\right)+2C\,(C_4)^2\,\eta\frac{\mu}{\sqrt{m}}\sqrt{\frac{\mu^2K\log^9 m}{m}}\\
&\overset{\text{(ii)}}{\leq}C_2\frac{\mu}{\sqrt{m}}\sqrt{\frac{\mu^2K\log^9 m}{m}}.
\end{aligned}$$

Here, (i) holds as long as m is sufficiently large such that $CC_1 1/\log^2 m \ll 1$ and

$$\max\left\{\left|\frac{\alpha^{t+1}}{\alpha^t}\right|, \left|\frac{\alpha^t}{\alpha^{t+1}}\right|\right\} < \frac{1-\eta/21}{1-\eta/20}, \tag{218}$$

which is guaranteed by Lemma 16. Inequality (ii) arises from the induction hypothesis (90b) and taking $C_2 > 0$ is sufficiently large.

Finally, we establish the second inequality claimed in the lemma. Take $(\boldsymbol{h}_1, \boldsymbol{x}_1) = (\widetilde{\boldsymbol{h}}^{t+1}, \widetilde{\boldsymbol{x}}^{t+1})$ and $(\boldsymbol{h}_2, \boldsymbol{x}_2) = (\widehat{\boldsymbol{h}}^{t+1,(l)}, \widehat{\boldsymbol{x}}^{t+1,(l)})$ in Lemma 55. Since both $(\boldsymbol{h}_1, \boldsymbol{x}_1)$ and $(\boldsymbol{h}_2, \boldsymbol{x}_2)$ are close enough to $(\boldsymbol{h}^\star, \boldsymbol{x}^\star)$, we deduce that

$$\left\|\widetilde{\boldsymbol{z}}^{t+1,(l)} - \widetilde{\boldsymbol{z}}^{t+1}\right\|_2 \lesssim \left\|\widehat{\boldsymbol{z}}^{t+1,(l)} - \widetilde{\boldsymbol{z}}^{t+1}\right\|_2 \lesssim C_2 \frac{\mu}{\sqrt{m}}\sqrt{\frac{\mu^2 K \log^9 m}{m}}$$

as claimed.

C.4 Proof of Lemma 18

Before going forward, we make note of the following inequality

$$\max_{1\le l\le m}\left|\boldsymbol{b}_l^{\mathsf{H}}\frac{1}{\overline{\alpha^{t+1}}}\boldsymbol{h}^{t+1}\right| \le \left|\frac{\alpha^t}{\alpha^{t+1}}\right| \max_{1\le l\le m}\left|\boldsymbol{b}_l^{\mathsf{H}}\frac{1}{\overline{\alpha^t}}\boldsymbol{h}^{t+1}\right| \le (1+\delta)\max_{1\le l\le m}\left|\boldsymbol{b}_l^{\mathsf{H}}\frac{1}{\overline{\alpha^t}}\boldsymbol{h}^{t+1}\right|$$

for some small $\delta \asymp \log^{-2} m$, where the last relation follows from Lemma 16 that

$$\left|\frac{\alpha^{t+1}}{\alpha^t} - 1\right| \lesssim \frac{1}{\log^2 m} \le \delta$$

for m sufficiently large. In view of the above inequality, the focus of our subsequent analysis will be to control $\max_l \left|\boldsymbol{b}_l^{\mathsf{H}}\frac{1}{\overline{\alpha^t}}\boldsymbol{h}^{t+1}\right|$.

The gradient update rule for $\boldsymbol{h}^{t+1}$ (cf. (79a)) gives

$$\frac{1}{\overline{\alpha^t}}\boldsymbol{h}^{t+1} = \widetilde{\boldsymbol{h}}^t - \eta\xi\sum_{j=1}^m \boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\big(\widetilde{\boldsymbol{h}}^t\widetilde{\boldsymbol{x}}^{t\mathsf{H}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\big)\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\widetilde{\boldsymbol{x}}^t,$$

where $\widetilde{\boldsymbol{h}}^t = \frac{1}{\overline{\alpha^t}}\boldsymbol{h}^t$ and $\widetilde{\boldsymbol{x}}^t = \alpha^t\boldsymbol{x}^t$. Here and below, we denote $\xi = 1/\|\widetilde{\boldsymbol{x}}^t\|_2^2$ for notational convenience. The above formula can be further decomposed into the following terms

$$\frac{1}{\overline{\alpha^t}}\boldsymbol{h}^{t+1} = \widetilde{\boldsymbol{h}}^t - \eta\xi\sum_{j=1}^m \boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\left|\boldsymbol{a}_j^{\mathsf{H}}\widetilde{\boldsymbol{x}}^t\right|^2 + \eta\xi\sum_{j=1}^m \boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\widetilde{\boldsymbol{x}}^t$$

$$= \left(1-\eta\xi\left\|\boldsymbol{x}^\star\right\|_2^2\right)\widetilde{\boldsymbol{h}}^t - \underbrace{\eta\xi\sum_{j=1}^m \boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\left(\left|\boldsymbol{a}_j^{\mathsf{H}}\widetilde{\boldsymbol{x}}^t\right|^2-\left|\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}^\star\right|^2\right)}_{:=\boldsymbol{v}_1}$$
$$\underbrace{-\eta\xi\sum_{j=1}^m \boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\left(\left|\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}^\star\right|^2-\left\|\boldsymbol{x}^\star\right\|_2^2\right)}_{:=\boldsymbol{v}_2} + \underbrace{\eta\xi\sum_{j=1}^m \boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\widetilde{\boldsymbol{x}}^t}_{:=\boldsymbol{v}_3},$$

where we use the fact that $\sum_{j=1}^m \boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}} = \boldsymbol{I}_K$. In the sequel, we shall control each term separately.

1. We start with $|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{v}_1|$ by making the observation that

$$\begin{aligned}\frac{1}{\eta\xi}\left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{v}_1\right| &= \left|\sum_{j=1}^m \boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\left[\boldsymbol{a}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^t-\boldsymbol{x}^\star\right)\left(\boldsymbol{a}_j^{\mathsf{H}}\widetilde{\boldsymbol{x}}^t\right)^{\mathsf{H}}+\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}^\star\left(\boldsymbol{a}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^t-\boldsymbol{x}^\star\right)\right)^{\mathsf{H}}\right]\right| \\ &\le \sum_{j=1}^m \left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{b}_j\right|\left\{\max_{1\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right|\right\}\left\{\max_{1\le j\le m}\left|\boldsymbol{a}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^t-\boldsymbol{x}^\star\right)\right|\left(\left|\boldsymbol{a}_j^{\mathsf{H}}\widetilde{\boldsymbol{x}}^t\right|+\left|\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}^\star\right|\right)\right\}.\end{aligned} \tag{219}$$

   Combining the induction hypothesis (90c) and condition (189) yields

$$\begin{aligned}\max_{1\le j\le m}\left|\boldsymbol{a}_j^{\mathsf{H}}\widetilde{\boldsymbol{x}}^t\right| &\le \max_{1\le j\le m}\left|\boldsymbol{a}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^t-\boldsymbol{x}^\star\right)\right|+\max_{1\le j\le m}\left|\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}^\star\right| \\ &\le C_3\frac{1}{\log^{3/2}m}+5\sqrt{\log m}\le 6\sqrt{\log m}\end{aligned}$$

   as long as m is sufficiently large. This further implies

$$\max_{1\le j\le m}\left|\boldsymbol{a}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^t-\boldsymbol{x}^\star\right)\right|\left(\left|\boldsymbol{a}_j^{\mathsf{H}}\widetilde{\boldsymbol{x}}^t\right|+\left|\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}^\star\right|\right)\le C_3\frac{1}{\log^{3/2}m}\cdot 11\sqrt{\log m}\le 11C_3\frac{1}{\log m}.$$

   Substituting it into (219) and taking Lemma 48, we arrive at

$$\frac{1}{\eta\xi}\left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{v}_1\right|\lesssim \log m\cdot\left\{\max_{1\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right|\right\}\cdot C_3\frac{1}{\log m}\lesssim C_3\max_{1\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right|\le 0.1\max_{1\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right|,$$

   with the proviso that C_3 is sufficiently small.
2. We then move on to $|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{v}_3|$, which obeys

$$\frac{1}{\eta\xi}\left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{v}_3\right|\le\left|\sum_{j=1}^m \boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}^\star\right|+\left|\sum_{j=1}^m \boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^t-\boldsymbol{x}^\star\right)\right|. \tag{220}$$

Regarding the first term, we have the following lemma, whose proof is given in Appendix C.4.1.

Lemma 28 *Suppose $m \geq CK\log^2 m$ for some sufficiently large constant $C > 0$. Then with probability at least $1 - O\left(m^{-10}\right)$, one has*

$$\left|\sum_{j=1}^{m} \boldsymbol{b}_l^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star - \boldsymbol{b}_l^{\mathsf{H}} \boldsymbol{h}^\star\right| \lesssim \frac{\mu}{\sqrt{m}}.$$

For the remaining term, we apply the same strategy as in bounding $|\boldsymbol{b}_l^{\mathsf{H}} \boldsymbol{v}_1|$ to get

$$\begin{aligned}
&\left|\sum_{j=1}^{m} \boldsymbol{b}_l^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \left(\widetilde{\boldsymbol{x}}^t - \boldsymbol{x}^\star\right)\right| \\
&\leq \sum_{j=1}^{m} \left|\boldsymbol{b}_l^{\mathsf{H}} \boldsymbol{b}_j\right| \left\{\max_{1\leq j\leq m} \left|\boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star\right|\right\} \left\{\max_{1\leq j\leq m} \left|\boldsymbol{a}_j^{\mathsf{H}} \left(\widetilde{\boldsymbol{x}}^t - \boldsymbol{x}^\star\right)\right|\right\} \left\{\max_{1\leq j\leq m} \left|\boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star\right|\right\} \\
&\leq 4\log m \cdot \frac{\mu}{\sqrt{m}} \cdot C_3 \frac{1}{\log^{3/2} m} \cdot 5\sqrt{\log m} \\
&\lesssim C_3 \frac{\mu}{\sqrt{m}},
\end{aligned}$$

where the second line follows from incoherence (36), the induction hypothesis (90c), condition (189), and Lemma 48. Combining the above three inequalities and incoherence (36) yields

$$\frac{1}{\eta\xi}\left|\boldsymbol{b}_l^{\mathsf{H}} \boldsymbol{v}_3\right| \lesssim \left|\boldsymbol{b}_l^{\mathsf{H}} \boldsymbol{h}^\star\right| + \frac{\mu}{\sqrt{m}} + C_3 \frac{\mu}{\sqrt{m}} \lesssim (1 + C_3) \frac{\mu}{\sqrt{m}}.$$

3. Finally, we need to control $\left|\boldsymbol{b}_l^{\mathsf{H}} \boldsymbol{v}_2\right|$. For convenience of presentation, we will only bound $\left|\boldsymbol{b}_1^{\mathsf{H}} \boldsymbol{v}_2\right|$ in the sequel, but the argument easily extends to all other $\boldsymbol{b}_l$'s. The idea is to group $\{\boldsymbol{b}_j\}_{1\leq j\leq m}$ into bins each containing τ adjacent vectors, and to look at each bin separately. Here, $\tau \asymp \mathrm{poly}\log(m)$ is some integer to be specified later. For notational simplicity, we assume m/τ to be an integer, although all arguments continue to hold when m/τ is not an integer. For each $0 \leq l \leq m - \tau$, the following summation over τ adjacent data obeys

$$\begin{aligned}
&\boldsymbol{b}_1^{\mathsf{H}} \sum_{j=1}^{\tau} \boldsymbol{b}_{l+j} \boldsymbol{b}_{l+j}^{\mathsf{H}} \widetilde{\boldsymbol{h}}^t \left(\left|\boldsymbol{a}_{l+j}^{\mathsf{H}} \boldsymbol{x}^\star\right|^2 - \left\|\boldsymbol{x}^\star\right\|_2^2\right) \\
&= \boldsymbol{b}_1^{\mathsf{H}} \sum_{j=1}^{\tau} \boldsymbol{b}_{l+1} \boldsymbol{b}_{l+1}^{\mathsf{H}} \widetilde{\boldsymbol{h}}^t \left(\left|\boldsymbol{a}_{l+j}^{\mathsf{H}} \boldsymbol{x}^\star\right|^2 - \left\|\boldsymbol{x}^\star\right\|_2^2\right)
\end{aligned}$$

$$
\begin{aligned}
&+ \boldsymbol{b}_1^{\mathsf{H}} \sum_{j=1}^{\tau} \left(\boldsymbol{b}_{l+j} \boldsymbol{b}_{l+j}^{\mathsf{H}} - \boldsymbol{b}_{l+1} \boldsymbol{b}_{l+1}^{\mathsf{H}} \right) \widetilde{\boldsymbol{h}}^t \left(\left| \boldsymbol{a}_{l+j}^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 - \left\| \boldsymbol{x}^\star \right\|_2^2 \right) \\
&= \left\{ \sum_{j=1}^{\tau} \left(\left| \boldsymbol{a}_{l+j}^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 - \left\| \boldsymbol{x}^\star \right\|_2^2 \right) \right\} \boldsymbol{b}_1^{\mathsf{H}} \boldsymbol{b}_{l+1} \boldsymbol{b}_{l+1}^{\mathsf{H}} \widetilde{\boldsymbol{h}}^t \\
&\quad + \boldsymbol{b}_1^{\mathsf{H}} \sum_{j=1}^{\tau} \left(\boldsymbol{b}_{l+j} - \boldsymbol{b}_{l+1} \right) \boldsymbol{b}_{l+j}^{\mathsf{H}} \widetilde{\boldsymbol{h}}^t \left(\left| \boldsymbol{a}_{l+j}^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 - \left\| \boldsymbol{x}^\star \right\|_2^2 \right) \\
&\quad + \boldsymbol{b}_1^{\mathsf{H}} \sum_{j=1}^{\tau} \boldsymbol{b}_{l+1} \left(\boldsymbol{b}_{l+j} - \boldsymbol{b}_{l+1} \right)^{\mathsf{H}} \widetilde{\boldsymbol{h}}^t \left(\left| \boldsymbol{a}_{l+j}^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 - \left\| \boldsymbol{x}^\star \right\|_2^2 \right). \qquad (221)
\end{aligned}
$$

We will now bound each term in (221) separately.

- Before bounding the first term in (221), we first bound the pre-factor $\left| \sum_{j=1}^{\tau} \left(|\boldsymbol{a}_{l+j}^{\mathsf{H}} \boldsymbol{x}^\star|^2 - \|\boldsymbol{x}^\star\|_2^2 \right) \right|$. Notably, the fluctuation of this quantity does not grow fast as it is the sum of i.i.d. random variables over a group of relatively large size, i.e., τ. Since $2 \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right|^2$ follows the χ_2^2 distribution, by standard concentration results (e.g., [95, Theorem 1.1]), with probability exceeding $1 - O\left(m^{-10}\right)$,

$$
\left| \sum_{j=1}^{\tau} \left(\left| \boldsymbol{a}_{l+j}^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 - \left\| \boldsymbol{x}^\star \right\|_2^2 \right) \right| \lesssim \sqrt{\tau \log m}.
$$

  With this result in place, we can bound the first term in (221) as

$$
\left| \left\{ \sum_{j=1}^{\tau} \left(\left| \boldsymbol{a}_{l+j}^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 - \left\| \boldsymbol{x}^\star \right\|_2^2 \right) \right\} \boldsymbol{b}_1^{\mathsf{H}} \boldsymbol{b}_{l+1} \boldsymbol{b}_{l+1}^{\mathsf{H}} \widetilde{\boldsymbol{h}}^t \right| \lesssim \sqrt{\tau \log m} \left| \boldsymbol{b}_1^{\mathsf{H}} \boldsymbol{b}_{l+1} \right| \max_{1 \le l \le m} \left| \boldsymbol{b}_l^{\mathsf{H}} \widetilde{\boldsymbol{h}}^t \right|.
$$

  Taking the summation over all bins gives

$$
\begin{aligned}
&\sum_{k=0}^{\frac{m}{\tau}-1} \left| \left\{ \sum_{j=1}^{\tau} \left(\left| \boldsymbol{a}_{k\tau+j}^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 - \left\| \boldsymbol{x}^\star \right\|_2^2 \right) \right\} \boldsymbol{b}_1^{\mathsf{H}} \boldsymbol{b}_{k\tau+1} \boldsymbol{b}_{k\tau+1}^{\mathsf{H}} \widetilde{\boldsymbol{h}}^t \right| \\
&\lesssim \sqrt{\tau \log m} \sum_{k=0}^{\frac{m}{\tau}-1} \left| \boldsymbol{b}_1^{\mathsf{H}} \boldsymbol{b}_{k\tau+1} \right| \max_{1 \le l \le m} \left| \boldsymbol{b}_l^{\mathsf{H}} \widetilde{\boldsymbol{h}}^t \right|. \qquad (222)
\end{aligned}
$$

  It is straightforward to see from the proof of Lemma 48 that

$$
\sum_{k=0}^{\frac{m}{\tau}-1} \left| \boldsymbol{b}_1^{\mathsf{H}} \boldsymbol{b}_{k\tau+1} \right| = \|\boldsymbol{b}_1\|_2^2 + \sum_{k=1}^{\frac{m}{\tau}-1} \left| \boldsymbol{b}_1^{\mathsf{H}} \boldsymbol{b}_{k\tau+1} \right| \le \frac{K}{m} + O\left(\frac{\log m}{\tau} \right). \qquad (223)
$$

Substituting (223) into the previous inequality (222) gives

$$\begin{aligned}
&\sum_{k=0}^{\frac{m}{\tau}-1}\left|\left\{\sum_{j=1}^{\tau}\left(\left|\boldsymbol{a}_{k\tau+j}^{\mathsf{H}}\boldsymbol{x}^{\star}\right|^2-\|\boldsymbol{x}^{\star}\|_2^2\right)\right\}\boldsymbol{b}_1^{\mathsf{H}}\boldsymbol{b}_{k\tau+1}\boldsymbol{b}_{k\tau+1}^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right| \\
&\quad\lesssim\left(\frac{K\sqrt{\tau\log m}}{m}+\sqrt{\frac{\log^3 m}{\tau}}\right)\max_{1\le l\le m}\left|\boldsymbol{b}_l^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right| \\
&\quad\le 0.1\max_{1\le l\le m}\left|\boldsymbol{b}_l^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right|,
\end{aligned}$$

as long as $m\gg K\sqrt{\tau\log m}$ and $\tau\gg\log^3 m$.

- The second term of (221) obeys

$$\begin{aligned}
&\left|\boldsymbol{b}_1^{\mathsf{H}}\sum_{j=1}^{\tau}\left(\boldsymbol{b}_{l+j}-\boldsymbol{b}_{l+1}\right)\boldsymbol{b}_{l+j}^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\left(\left|\boldsymbol{a}_{l+j}^{\mathsf{H}}\boldsymbol{x}^{\star}\right|^2-\|\boldsymbol{x}^{\star}\|_2^2\right)\right| \\
&\quad\le\max_{1\le l\le m}\left|\boldsymbol{b}_l^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right|\sqrt{\sum_{j=1}^{\tau}\left|\boldsymbol{b}_1^{\mathsf{H}}\left(\boldsymbol{b}_{l+j}-\boldsymbol{b}_{l+1}\right)\right|^2}\sqrt{\sum_{j=1}^{\tau}\left(\left|\boldsymbol{a}_{l+j}^{\mathsf{H}}\boldsymbol{x}^{\star}\right|^2-\|\boldsymbol{x}^{\star}\|_2^2\right)^2} \\
&\quad\lesssim\sqrt{\tau}\max_{1\le l\le m}\left|\boldsymbol{b}_l^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right|\sqrt{\sum_{j=1}^{\tau}\left|\boldsymbol{b}_1^{\mathsf{H}}\left(\boldsymbol{b}_{l+j}-\boldsymbol{b}_{l+1}\right)\right|^2},
\end{aligned}$$

where the first inequality is due to Cauchy–Schwarz, and the second one holds because of the following lemma, whose proof can be found in Appendix C.4.2.

Lemma 29 *Suppose $\tau\ge C\log^4 m$ for some sufficiently large constant $C>0$. Then with probability exceeding $1-O\left(m^{-10}\right)$,*

$$\sum_{j=1}^{\tau}\left(\left|\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}^{\star}\right|^2-\|\boldsymbol{x}^{\star}\|_2^2\right)^2\lesssim\tau.$$

With the above bound in mind, we can sum over all bins of size τ to obtain

$$\begin{aligned}
&\left|\boldsymbol{b}_1^{\mathsf{H}}\sum_{k=0}^{\frac{m}{\tau}-1}\sum_{j=1}^{\tau}\left(\boldsymbol{b}_{k\tau+j}-\boldsymbol{b}_{k\tau+1}\right)\boldsymbol{b}_{k\tau+j}^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\left\{\left|\boldsymbol{a}_{l+j}^{\mathsf{H}}\boldsymbol{x}^{\star}\right|^2-\|\boldsymbol{x}^{\star}\|_2^2\right\}\right| \\
&\quad\lesssim\left\{\sqrt{\tau}\sum_{k=0}^{\frac{m}{\tau}-1}\sqrt{\sum_{j=1}^{\tau}\left|\boldsymbol{b}_1^{\mathsf{H}}\left(\boldsymbol{b}_{k\tau+j}-\boldsymbol{b}_{k\tau+1}\right)\right|^2}\right\}\max_{1\le l\le m}\left|\boldsymbol{b}_l^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right| \\
&\quad\le 0.1\max_{1\le l\le m}\left|\boldsymbol{b}_l^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right|.
\end{aligned}$$

Here, the last line arises from Lemma 51, which says that for any small constant $c > 0$, as long as $m \gg \tau K \log m$

$$\sum_{k=0}^{\frac{m}{\tau}-1} \sqrt{\sum_{j=1}^{\tau} \left| \boldsymbol{b}_1^{\mathsf{H}} \left(\boldsymbol{b}_{k\tau+j} - \boldsymbol{b}_{k\tau+1} \right) \right|^2} \le c \frac{1}{\sqrt{\tau}}.$$

- The third term of (221) obeys

$$\begin{aligned} &\left| \boldsymbol{b}_1^{\mathsf{H}} \sum_{j=1}^{\tau} \boldsymbol{b}_{l+1} \left(\boldsymbol{b}_{l+j} - \boldsymbol{b}_{l+1} \right)^{\mathsf{H}} \widetilde{\boldsymbol{h}}^t \left\{ \left| \boldsymbol{a}_{l+j}^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 - \left\| \boldsymbol{x}^\star \right\|_2^2 \right\} \right| \\ &\quad \le \left| \boldsymbol{b}_1^{\mathsf{H}} \boldsymbol{b}_{l+1} \right| \left\{ \sum_{j=1}^{\tau} \left| \left| \boldsymbol{a}_{l+j}^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 - \left\| \boldsymbol{x}^\star \right\|_2^2 \right| \right\} \max_{0 \le l \le m-\tau,\, 1 \le j \le \tau} \left| \left(\boldsymbol{b}_{l+j} - \boldsymbol{b}_{l+1} \right)^{\mathsf{H}} \widetilde{\boldsymbol{h}}^t \right| \\ &\quad \lesssim \tau \left| \boldsymbol{b}_1^{\mathsf{H}} \boldsymbol{b}_{l+1} \right| \max_{0 \le l \le m-\tau,\, 1 \le j \le \tau} \left| \left(\boldsymbol{b}_{l+j} - \boldsymbol{b}_{l+1} \right)^{\mathsf{H}} \widetilde{\boldsymbol{h}}^t \right|, \end{aligned}$$

where the last line relies on the inequality

$$\sum_{j=1}^{\tau} \left| \left| \boldsymbol{a}_{l+j}^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 - \left\| \boldsymbol{x}^\star \right\|_2^2 \right| \le \sqrt{\tau} \sqrt{\sum_{j=1}^{\tau} \left(\left| \boldsymbol{a}_{l+j}^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 - \left\| \boldsymbol{x}^\star \right\|_2^2 \right)^2} \lesssim \tau$$

owing to Lemma 29 and the Cauchy–Schwarz inequality. Summing over all bins gives

$$\begin{aligned} &\sum_{k=0}^{\frac{m}{\tau}-1} \left| \boldsymbol{b}_1^{\mathsf{H}} \sum_{j=1}^{\tau} \boldsymbol{b}_{k\tau+1} \left(\boldsymbol{b}_{k\tau+j} - \boldsymbol{b}_{k\tau+1} \right)^{\mathsf{H}} \widetilde{\boldsymbol{h}}^t \left\{ \left| \boldsymbol{a}_{k\tau+j}^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 - \left\| \boldsymbol{x}^\star \right\|_2^2 \right\} \right| \\ &\quad \lesssim \tau \sum_{k=0}^{\frac{m}{\tau}-1} \left| \boldsymbol{b}_1^{\mathsf{H}} \boldsymbol{b}_{k\tau+1} \right| \max_{0 \le l \le m-\tau,\, 1 \le j \le \tau} \left| \left(\boldsymbol{b}_{l+j} - \boldsymbol{b}_{l+1} \right)^{\mathsf{H}} \widetilde{\boldsymbol{h}}^t \right| \\ &\quad \lesssim \log m \max_{0 \le l \le m-\tau,\, 1 \le j \le \tau} \left| \left(\boldsymbol{b}_{l+j} - \boldsymbol{b}_{l+1} \right)^{\mathsf{H}} \widetilde{\boldsymbol{h}}^t \right|, \end{aligned}$$

where the last relation makes use of (223) with the proviso that $m \gg K\tau$. It then boils down to bounding $\max_{0 \le l \le m-\tau,\, 1 \le j \le \tau} \left| \left(\boldsymbol{b}_{l+j} - \boldsymbol{b}_{l+1} \right)^{\mathsf{H}} \widetilde{\boldsymbol{h}}^t \right|$. Without loss of generality, it suffices to look at $\left| \left(\boldsymbol{b}_j - \boldsymbol{b}_1 \right)^{\mathsf{H}} \widetilde{\boldsymbol{h}}^t \right|$ for all $1 \le j \le \tau$. Specifically, we claim for the moment that

$$\max_{1 \le j \le \tau} \left| \left(\boldsymbol{b}_j - \boldsymbol{b}_1 \right)^{\mathsf{H}} \widetilde{\boldsymbol{h}}^t \right| \le c C_4 \frac{\mu}{\sqrt{m}} \log m \tag{224}$$

for some sufficiently small constant $c > 0$, provided that $m \gg \tau K \log^4 m$. As a result,

$$\sum_{k=0}^{\frac{m}{\tau}-1} \left| \boldsymbol{b}_1^{\mathsf{H}} \sum_{j=1}^{\tau} \boldsymbol{b}_{k\tau+1} \left(\boldsymbol{b}_{k\tau+j} - \boldsymbol{b}_{k\tau+1} \right)^{\mathsf{H}} \widetilde{\boldsymbol{h}}^t \left\{ \left| \boldsymbol{a}_{k\tau+j}^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 - \left\| \boldsymbol{x}^\star \right\|_2^2 \right\} \right| \lesssim c C_4 \frac{\mu}{\sqrt{m}} \log^2 m.$$

- Putting the above results together, we get

$$\begin{aligned} \frac{1}{\eta\xi} \left| \boldsymbol{b}_1^{\mathsf{H}} \boldsymbol{v}_2 \right| &\le \sum_{k=0}^{\frac{m}{\tau}-1} \left| \boldsymbol{b}_1^{\mathsf{H}} \sum_{j=1}^{\tau} \boldsymbol{b}_{k\tau+j} \boldsymbol{b}_{k\tau+j}^{\mathsf{H}} \widetilde{\boldsymbol{h}}^t \left\{ \left| \boldsymbol{a}_{k\tau+j}^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 - \left\| \boldsymbol{x}^\star \right\|_2^2 \right\} \right| \\ &\le 0.2 \max_{1 \le l \le m} \left| \boldsymbol{b}_l^{\mathsf{H}} \widetilde{\boldsymbol{h}}^t \right| + O\left(c C_4 \frac{\mu}{\sqrt{m}} \log^2 m \right). \end{aligned}$$

4. Combining the preceding bounds guarantees the existence of some constant $C_8 > 0$ such that

$$\begin{aligned} \left| \boldsymbol{b}_l^{\mathsf{H}} \widetilde{\boldsymbol{h}}^{t+1} \right| &\le (1+\delta) \left\{ (1-\eta\xi) \left| \boldsymbol{b}_l^{\mathsf{H}} \widetilde{\boldsymbol{h}}^t \right| + 0.3\eta\xi \max_{1 \le l \le m} \left| \boldsymbol{b}_l^{\mathsf{H}} \widetilde{\boldsymbol{h}}^t \right| \right. \\ &\qquad \left. + C_8(1+C_3)\eta\xi \frac{\mu}{\sqrt{m}} + C_8 \eta\xi c C_4 \frac{\mu}{\sqrt{m}} \log^2 m \right\} \\ &\overset{\text{(i)}}{\le} \left(1 + O\left(\frac{1}{\log^2 m} \right) \right) \left\{ (1 - 0.7\eta\xi)\, C_4 \frac{\mu}{\sqrt{m}} \log^2 m \right. \\ &\qquad \left. + C_8(1+C_3)\eta\xi \frac{\mu}{\sqrt{m}} + C_8 \eta\xi c C_4 \frac{\mu}{\sqrt{m}} \log^2 m \right\} \\ &\overset{\text{(ii)}}{\le} C_4 \frac{\mu}{\sqrt{m}} \log^2 m. \end{aligned}$$

Here, (i) uses the induction hypothesis (90d), and (ii) holds as long as $c > 0$ is sufficiently small (so that $(1+\delta)C_8\eta\xi c \ll 1$) and $\eta > 0$ is some sufficiently small constant. In order for the proof to go through, it suffices to pick

$$\tau = c_{10} \log^4 m$$

for some sufficiently large constant $c_{10} > 0$. Accordingly, we need the sample size to exceed

$$m \gg \mu^2 \tau K \log^4 m \asymp \mu^2 K \log^8 m.$$

Finally, it remains to verify claim (224), which we accomplish in Appendix C.4.3.

C.4.1 Proof of Lemma 28

Denote

$$w_j = \boldsymbol{b}_l^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{x}^{\star \mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star.$$

Recognizing that $\mathbb{E}[\boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}}] = \boldsymbol{I}_K$ and $\sum_{j=1}^m \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} = \boldsymbol{I}_K$, we can write the quantity of interest as the sum of independent random variables, namely

$$\sum_{j=1}^m \boldsymbol{b}_l^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{x}^{\star \mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star - \boldsymbol{b}_l^{\mathsf{H}} \boldsymbol{h}^\star = \sum_{j=1}^m \left(w_j - \mathbb{E}\left[w_j \right] \right).$$

Further, the sub-exponential norm (see definition in [116]) of $w_j - \mathbb{E}\left[w_j \right]$ obeys

$$\left\| w_j - \mathbb{E}\left[w_j \right] \right\|_{\psi_1} \overset{\text{(i)}}{\le} 2 \left\| w_j \right\|_{\psi_1} \overset{\text{(ii)}}{\le} 4 \left| \boldsymbol{b}_l^{\mathsf{H}} \boldsymbol{b}_j \right| \left| \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \right| \left\| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right\|_{\psi_2}^2 \overset{\text{(iii)}}{\lesssim} \left| \boldsymbol{b}_l^{\mathsf{H}} \boldsymbol{b}_j \right| \frac{\mu}{\sqrt{m}} \overset{\text{(iv)}}{\le} \frac{\mu \sqrt{K}}{m},$$

where (i) arises from the centering property of the sub-exponential norm (see [116, Remark 5.18]), (ii) utilizes the relationship between the sub-exponential norm and the sub-Gaussian norm [116, Lemma 5.14], (iii) is a consequence of the incoherence condition (36) and the fact that $\left\| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right\|_{\psi_2} \lesssim 1$, and (iv) follows from $\| \boldsymbol{b}_j \|_2 = \sqrt{K/m}$. Let $M = \max_{j \in [m]} \left\| w_j - \mathbb{E}\left[w_j \right] \right\|_{\psi_1}$ and

$$V^2 = \sum_{j=1}^m \left\| w_j - \mathbb{E}\left[w_j \right] \right\|_{\psi_1}^2 \lesssim \sum_{j=1}^m \left(\left| \boldsymbol{b}_l^{\mathsf{H}} \boldsymbol{b}_j \right| \frac{\mu}{\sqrt{m}} \right)^2 = \frac{\mu^2}{m} \| \boldsymbol{b}_l \|_2^2 = \frac{\mu^2 K}{m^2},$$

which follows since $\sum_{j=1}^m \left| \boldsymbol{b}_l^{\mathsf{H}} \boldsymbol{b}_j \right|^2 = \boldsymbol{b}_l^{\mathsf{H}} \left(\sum_{j=1}^m \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \right) \boldsymbol{b}_l = \| \boldsymbol{b}_l \|_2^2 = K/m$. Let $a_j = \left\| w_j - \mathbb{E}\left[w_j \right] \right\|_{\psi_1}$ and $X_j = (w_j - \mathbb{E}[w_j])/a_j$. Since $\| X_j \|_{\psi_1} = 1$, $\sum_{j=1}^m a_j^2 = V^2$ and $\max_{j \in [m]} |a_j| = M$, we can invoke [116, Proposition 5.16] to obtain that

$$\mathbb{P}\left(\left| \sum_{j=1}^m a_j X_j \right| \ge t \right) \le 2 \exp\left(-c \min\left\{ \frac{t}{M}, \frac{t^2}{V^2} \right\} \right),$$

where $c > 0$ is some universal constant. By taking $t = \mu/\sqrt{m}$, we see there exists some constant c' such that

$$\begin{aligned} &\mathbb{P}\left(\left| \sum_{j=1}^m \boldsymbol{b}_l^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{x}^{\star \mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star - \boldsymbol{b}_l^{\mathsf{H}} \boldsymbol{h}^\star \right| \ge \frac{\mu}{\sqrt{m}} \right) \\ &\quad \le 2 \exp\left(-c \min\left\{ \frac{\mu/\sqrt{m}}{M}, \frac{\mu^2/m}{V^2} \right\} \right) \end{aligned}$$

$$\leq 2\exp\left(-c'\min\left\{\frac{\mu/\sqrt{m}}{\mu\sqrt{K}/m},\frac{\mu^2/m}{\mu^2K/m^2}\right\}\right)$$
$$= 2\exp\left(-c'\min\left\{\sqrt{m/K},m/K\right\}\right).$$

We conclude the proof by observing that $m \gg K\log^2 m$ as stated in the assumption.

C.4.2 Proof of Lemma 29

From the elementary inequality $(a-b)^2 \leq 2\left(a^2+b^2\right)$, we see that

$$\sum_{j=1}^{\tau}\left(\left|\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}^\star\right|^2-\left\|\boldsymbol{x}^\star\right\|_2^2\right)^2 \leq 2\sum_{j=1}^{\tau}\left(\left|\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}^\star\right|^4+\left\|\boldsymbol{x}^\star\right\|_2^4\right)=2\sum_{j=1}^{\tau}\left|\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}^\star\right|^4+2\tau, \tag{225}$$

where the last identity holds true since $\|\boldsymbol{x}^\star\|_2 = 1$. It thus suffices to control $\sum_{j=1}^{\tau}\left|\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}^\star\right|^4$. Let $\xi_i = \boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}^\star$, which is a standard complex Gaussian random variable. Since the ξ_i's are statistically independent, one has

$$\mathsf{Var}\left(\sum_{i=1}^{\tau}|\xi_i|^4\right)\leq C_4\tau$$

for some constant $C_4 > 0$. It then follows from the hypercontractivity concentration result for Gaussian polynomials that [99, Theorem 1.9]

$$\begin{aligned}
&\mathbb{P}\left\{\sum_{i=1}^{\tau}\left(|\xi_i|^4-\mathbb{E}\left[|\xi_i|^4\right]\right)\geq c\tau\right\}\\
&\leq C\exp\left(-c_2\left(\frac{c^2\tau^2}{\mathsf{Var}\left(\sum_{i=1}^{\tau}|\xi_i|^4\right)}\right)^{1/4}\right)\\
&\leq C\exp\left(-c_2\left(\frac{c^2\tau^2}{C_4\tau}\right)^{1/4}\right)=C\exp\left(-c_2\left(\frac{c^2}{C_4}\right)^{1/4}\tau^{1/4}\right)\\
&\leq O(m^{-10}),
\end{aligned}$$

for some constants $c, c_2, C > 0$, with the proviso that $\tau \gg \log^4 m$. As a consequence, with probability at least $1 - O(m^{-10})$,

$$\sum_{j=1}^{\tau}\left|\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}^\star\right|^4 \lesssim \tau+\sum_{j=1}^{\tau}\mathbb{E}\left[\left|\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}^\star\right|^4\right]\asymp\tau,$$

which together with (225) concludes the proof.

C.4.3 Proof of Claim (224)

We will prove the claim by induction. Again, observe that

$$\left|\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right| = \left|\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\frac{1}{\overline{\alpha^t}}\boldsymbol{h}^t\right| = \left|\frac{\alpha^{t-1}}{\alpha^t}\right|\left|\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\frac{1}{\overline{\alpha^{t-1}}}\boldsymbol{h}^t\right| \le (1+\delta)\left|\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\frac{1}{\overline{\alpha^{t-1}}}\boldsymbol{h}^t\right|$$

for some $\delta \asymp \log^{-2} m$, which allows us to look at $\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\frac{1}{\overline{\alpha^{t-1}}}\boldsymbol{h}^t$ instead.

Use the gradient update rule for $\boldsymbol{h}^t$ (cf. (79a)) once again to get

$$\begin{aligned}\frac{1}{\overline{\alpha^{t-1}}}\boldsymbol{h}^t &= \frac{1}{\overline{\alpha^{t-1}}}\left(\boldsymbol{h}^{t-1}-\frac{\eta}{\left\|\boldsymbol{x}^{t-1}\right\|_2^2}\sum_{l=1}^m \boldsymbol{b}_l\boldsymbol{b}_l^{\mathsf{H}}\left(\boldsymbol{h}^{t-1}\boldsymbol{x}^{t-1\mathsf{H}}-\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right)\boldsymbol{a}_l\boldsymbol{a}_l^{\mathsf{H}}\boldsymbol{x}^{t-1}\right)\\ &= \widetilde{\boldsymbol{h}}^{t-1}-\eta\theta\sum_{l=1}^m \boldsymbol{b}_l\boldsymbol{b}_l^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^{t-1}\widetilde{\boldsymbol{x}}^{t-1\mathsf{H}}-\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right)\boldsymbol{a}_l\boldsymbol{a}_l^{\mathsf{H}}\widetilde{\boldsymbol{x}}^{t-1},\end{aligned}$$

where we denote $\theta := 1/\left\|\widetilde{\boldsymbol{x}}^{t-1}\right\|_2^2$. This further gives rise to

$$\begin{aligned}&\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\frac{1}{\overline{\alpha^{t-1}}}\boldsymbol{h}^t\\ &= \left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\widetilde{\boldsymbol{h}}^{t-1}-\eta\theta\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\sum_{l=1}^m \boldsymbol{b}_l\boldsymbol{b}_l^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^{t-1}\widetilde{\boldsymbol{x}}^{t-1\mathsf{H}}-\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right)\boldsymbol{a}_l\boldsymbol{a}_l^{\mathsf{H}}\widetilde{\boldsymbol{x}}^{t-1}\\ &= \left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\widetilde{\boldsymbol{h}}^{t-1}-\eta\theta\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\sum_{l=1}^m \boldsymbol{b}_l\boldsymbol{b}_l^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^{t-1}\widetilde{\boldsymbol{x}}^{t-1\mathsf{H}}-\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right)\widetilde{\boldsymbol{x}}^{t-1}\\ &\quad-\eta\theta\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\sum_{l=1}^m \boldsymbol{b}_l\boldsymbol{b}_l^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^{t-1}\widetilde{\boldsymbol{x}}^{t-1\mathsf{H}}-\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right)\left(\boldsymbol{a}_l\boldsymbol{a}_l^{\mathsf{H}}-\boldsymbol{I}_K\right)\widetilde{\boldsymbol{x}}^{t-1}\\ &= \left(1-\eta\theta\|\widetilde{\boldsymbol{x}}^{t-1}\|_2^2\right)\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\widetilde{\boldsymbol{h}}^{t-1}+\underbrace{\eta\theta\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\boldsymbol{h}^\star\left(\boldsymbol{x}^{\star\mathsf{H}}\widetilde{\boldsymbol{x}}^{t-1}\right)}_{:=\beta_1}\\ &\quad\underbrace{-\eta\theta\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\sum_{l=1}^m \boldsymbol{b}_l\boldsymbol{b}_l^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^{t-1}\widetilde{\boldsymbol{x}}^{t-1\mathsf{H}}-\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right)\left(\boldsymbol{a}_l\boldsymbol{a}_l^{\mathsf{H}}-\boldsymbol{I}_K\right)\widetilde{\boldsymbol{x}}^{t-1}}_{:=\beta_2},\end{aligned}$$

where the last identity makes use of the fact that $\sum_{l=1}^m \boldsymbol{b}_l\boldsymbol{b}_l^{\mathsf{H}} = \boldsymbol{I}_K$. For β_1, one can get

$$\frac{1}{\eta\theta}|\beta_1| \le \left|\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\boldsymbol{h}^\star\right|\|\boldsymbol{x}^\star\|_2\|\widetilde{\boldsymbol{x}}^{t-1}\|_2 \le 4\frac{\mu}{\sqrt{m}},$$

where we utilize the incoherence condition (36) and the fact that $\widetilde{\boldsymbol{x}}^{t-1}$ and $\boldsymbol{x}^\star$ are extremely close, i.e.,

$$\left\|\widetilde{\boldsymbol{x}}^{t-1}-\boldsymbol{x}^\star\right\|_2 \le \operatorname{dist}\left(\boldsymbol{z}^{t-1},\boldsymbol{z}^\star\right) \ll 1 \qquad \Longrightarrow \qquad \|\widetilde{\boldsymbol{x}}^{t-1}\|_2 \le 2.$$

Regarding the second term β_2, we have

$$\frac{1}{\eta\theta}\,|\beta_2| \le \left\{\sum_{l=1}^{m}\left|\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\boldsymbol{b}_l\right|\right\}\underbrace{\max_{1\le l\le m}\left|\boldsymbol{b}_l^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^{t-1}\widetilde{\boldsymbol{x}}^{t-1\mathsf{H}}-\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right)\left(\boldsymbol{a}_l\boldsymbol{a}_l^{\mathsf{H}}-\boldsymbol{I}_K\right)\widetilde{\boldsymbol{x}}^{t-1}\right|}_{:=\psi}.$$

The term ψ can be bounded as follows

$$\begin{aligned}
\psi &\le \max_{1\le l\le m}\left|\boldsymbol{b}_l^{\mathsf{H}}\widetilde{\boldsymbol{h}}^{t-1}\widetilde{\boldsymbol{x}}^{t-1\mathsf{H}}\left(\boldsymbol{a}_l\boldsymbol{a}_l^{\mathsf{H}}-\boldsymbol{I}\right)\widetilde{\boldsymbol{x}}^{t-1}\right| + \max_{1\le l\le m}\left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\left(\boldsymbol{a}_l\boldsymbol{a}_l^{\mathsf{H}}-\boldsymbol{I}_K\right)\widetilde{\boldsymbol{x}}^{t-1}\right| \\
&\le \max_{1\le l\le m}\left|\boldsymbol{b}_l^{\mathsf{H}}\widetilde{\boldsymbol{h}}^{t-1}\right|\max_{1\le l\le m}\left|\widetilde{\boldsymbol{x}}^{t-1\mathsf{H}}\left(\boldsymbol{a}_l\boldsymbol{a}_l^{\mathsf{H}}-\boldsymbol{I}_K\right)\widetilde{\boldsymbol{x}}^{t-1}\right| \\
&\quad + \max_{1\le l\le m}\left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^\star\right|\max_{1\le l\le m}\left|\boldsymbol{x}^{\star\mathsf{H}}\left(\boldsymbol{a}_l\boldsymbol{a}_l^{\mathsf{H}}-\boldsymbol{I}_K\right)\widetilde{\boldsymbol{x}}^{t-1}\right| \\
&\lesssim \log m\left\{\max_{1\le l\le m}\left|\boldsymbol{b}_l^{\mathsf{H}}\widetilde{\boldsymbol{h}}^{t-1}\right| + \frac{\mu}{\sqrt{m}}\right\}.
\end{aligned}$$

Here, we have used the incoherence condition (36) and the facts that

$$\begin{aligned}
\left|(\widetilde{\boldsymbol{x}}^{t-1})^{\mathsf{H}}\left(\boldsymbol{a}_l\boldsymbol{a}_l^{\mathsf{H}}-\boldsymbol{I}\right)\widetilde{\boldsymbol{x}}^{t-1}\right| &\le \left\|\boldsymbol{a}_l^{\mathsf{H}}\widetilde{\boldsymbol{x}}^{t-1}\right\|_2^2 + \left\|\widetilde{\boldsymbol{x}}^{t-1}\right\|_2^2 \lesssim \log m, \\
\left|\boldsymbol{x}^{\star\mathsf{H}}\left(\boldsymbol{a}_l\boldsymbol{a}_l^{\mathsf{H}}-\boldsymbol{I}\right)\widetilde{\boldsymbol{x}}^{t-1}\right| &\le \left\|\boldsymbol{a}_l^{\mathsf{H}}\widetilde{\boldsymbol{x}}^{t-1}\right\|_2\left\|\boldsymbol{a}_l^{\mathsf{H}}\boldsymbol{x}^\star\right\|_2 + \left\|\widetilde{\boldsymbol{x}}^{t-1}\right\|_2\|\boldsymbol{x}^\star\|_2 \lesssim \log m,
\end{aligned}$$

which are immediate consequences of (90c) and (189). Combining this with Lemma 50, we see that for any small constant $c>0$

$$\frac{1}{\eta\theta}\,|\beta_2| \le c\frac{1}{\log m}\left\{\max_{1\le l\le m}\left|\boldsymbol{b}_l^{\mathsf{H}}\widetilde{\boldsymbol{h}}^{t-1}\right| + \frac{\mu}{\sqrt{m}}\right\}$$

holds as long as $m \gg \tau K\log^4 m$.

To summarize, we arrive at

$$\begin{aligned}
\left|\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\widetilde{\boldsymbol{h}}^{t}\right| \le (1+\delta)\Bigg\{&\left(1-\eta\theta\left\|\widetilde{\boldsymbol{x}}^{t-1}\right\|_2^2\right)\left|\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\widetilde{\boldsymbol{h}}^{t-1}\right| \\
&+4\eta\theta\frac{\mu}{\sqrt{m}} + c\eta\theta\frac{1}{\log m}\left[\max_{1\le l\le m}\left|\boldsymbol{b}_l^{\mathsf{H}}\widetilde{\boldsymbol{h}}^{t-1}\right| + \frac{\mu}{\sqrt{m}}\right]\Bigg\}.
\end{aligned}$$

Making use of the induction hypothesis (85c) and the fact that $\left\|\widetilde{\boldsymbol{x}}^{t-1}\right\|_2^2 \geq 0.9$, we reach

$$\left|\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right| \leq (1+\delta)\left\{(1-0.9\eta\theta)\left|\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\widetilde{\boldsymbol{h}}^{t-1}\right|+cC_4\eta\theta\frac{\mu}{\sqrt{m}}\log m+\frac{c\mu\eta\theta}{\sqrt{m}\log m}\right\}.$$

Recall that $\delta \asymp 1/\log^2 m$. As a result, if $\eta > 0$ is some sufficiently small constant and if

$$\left|\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\widetilde{\boldsymbol{h}}^{t-1}\right| \leq 10c\left(C_4\frac{\mu}{\sqrt{m}}\log m+\frac{\mu}{\eta\theta\sqrt{m}\log m}\right) \leq 20cC_4\frac{\mu}{\sqrt{m}}\log m$$

holds, then one has

$$\left|\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\widetilde{\boldsymbol{h}}^{t}\right| \leq 20cC_4\frac{\mu}{\sqrt{m}}\log m.$$

Therefore, this concludes the proof of claim (224) by induction, provided that the base case is true, i.e., for some $c > 0$ sufficiently small

$$\left|\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\widetilde{\boldsymbol{h}}^{0}\right| \leq 20cC_4\frac{\mu}{\sqrt{m}}\log m. \tag{226}$$

Claim (226) is proved in Appendix C.6 (see Lemma 30).

C.5 Proof of Lemma 19

Recall that $\check{\boldsymbol{h}}^0$ and $\check{\boldsymbol{x}}^0$ are the leading left and right singular vectors of $\boldsymbol{M}$, respectively. Applying a variant of Wedin's sinΘ theorem [42, Theorem 2.1], we derive that

$$\min_{\alpha\in\mathbb{C},|\alpha|=1}\left\{\left\|\alpha\check{\boldsymbol{h}}^0-\boldsymbol{h}^\star\right\|_2+\left\|\alpha\check{\boldsymbol{x}}^0-\boldsymbol{x}^\star\right\|_2\right\} \leq \frac{c_1\left\|\boldsymbol{M}-\mathbb{E}\left[\boldsymbol{M}\right]\right\|}{\sigma_1\left(\mathbb{E}\left[\boldsymbol{M}\right]\right)-\sigma_2\left(\boldsymbol{M}\right)}, \tag{227}$$

for some universal constant $c_1 > 0$. Regarding the numerator of (227), it has been shown in [76, Lemma 5.20] that for any $\xi > 0$,

$$\left\|\boldsymbol{M}-\mathbb{E}\left[\boldsymbol{M}\right]\right\| \leq \xi \tag{228}$$

with probability exceeding $1-O(m^{-10})$, provided that

$$m \geq \frac{c_2\mu^2K\log^2 m}{\xi^2}$$

for some universal constant $c_2 > 0$. For the denominator of (227), we can take (228) together with Weyl's inequality to demonstrate that

$$\sigma_1\left(\mathbb{E}\left[\boldsymbol{M}\right]\right)-\sigma_2\left(\boldsymbol{M}\right) \geq \sigma_1\left(\mathbb{E}\left[\boldsymbol{M}\right]\right)-\sigma_2\left(\mathbb{E}\left[\boldsymbol{M}\right]\right)-\left\|\boldsymbol{M}-\mathbb{E}\left[\boldsymbol{M}\right]\right\| \geq 1-\xi,$$

where the last inequality utilizes the facts that $\sigma_1(\mathbb{E}[\boldsymbol{M}]) = 1$ and $\sigma_2(\mathbb{E}[\boldsymbol{M}]) = 0$. These together with (227) reveal that

$$\min_{\alpha\in\mathbb{C},|\alpha|=1} \left\{ \left\|\alpha\check{\boldsymbol{h}}^0 - \boldsymbol{h}^\star\right\|_2 + \left\|\alpha\check{\boldsymbol{x}}^0 - \boldsymbol{x}^\star\right\|_2 \right\} \le \frac{c_1\xi}{1-\xi} \le 2c_1\xi \tag{229}$$

as long as $\xi \le 1/2$.

Now we connect the preceding bound (229) with the scaled singular vectors $\boldsymbol{h}^0 = \sqrt{\sigma_1(\boldsymbol{M})}\,\check{\boldsymbol{h}}^0$ and $\boldsymbol{x}^0 = \sqrt{\sigma_1(\boldsymbol{M})}\,\check{\boldsymbol{x}}^0$. For any $\alpha \in \mathbb{C}$ with $|\alpha| = 1$, from the definition of $\boldsymbol{h}^0$ and $\boldsymbol{x}^0$ we have

$$\left\|\alpha\boldsymbol{h}^0 - \boldsymbol{h}^\star\right\|_2 + \left\|\alpha\boldsymbol{x}^0 - \boldsymbol{x}^\star\right\|_2 = \left\|\sqrt{\sigma_1(\boldsymbol{M})}\left(\alpha\check{\boldsymbol{h}}^0\right) - \boldsymbol{h}^\star\right\|_2 + \left\|\sqrt{\sigma_1(\boldsymbol{M})}\left(\alpha\check{\boldsymbol{x}}^0\right) - \boldsymbol{x}^\star\right\|_2.$$

Since $\alpha\check{\boldsymbol{h}}^0, \alpha\check{\boldsymbol{x}}^0$ are also the leading left and right singular vectors of $\boldsymbol{M}$, we can invoke Lemma 60 to get

$$\begin{aligned}
&\left\|\alpha\boldsymbol{h}^0 - \boldsymbol{h}^\star\right\|_2 + \left\|\alpha\boldsymbol{x}^0 - \boldsymbol{x}^\star\right\|_2 \\
&\le \sqrt{\sigma_1(\mathbb{E}[\boldsymbol{M}])}\left(\left\|\alpha\check{\boldsymbol{h}}^0 - \boldsymbol{h}^\star\right\|_2 + \left\|\alpha\check{\boldsymbol{x}}^0 - \boldsymbol{x}^\star\right\|_2\right) + \frac{2\left|\sigma_1(\boldsymbol{M}) - \sigma_1(\mathbb{E}[\boldsymbol{M}])\right|}{\sqrt{\sigma_1(\boldsymbol{M})} + \sqrt{\sigma_1(\mathbb{E}[\boldsymbol{M}])}} \\
&= \left\|\alpha\check{\boldsymbol{h}}^0 - \boldsymbol{h}^\star\right\|_2 + \left\|\alpha\check{\boldsymbol{x}}^0 - \boldsymbol{x}^\star\right\|_2 + \frac{2\left|\sigma_1(\boldsymbol{M}) - \sigma_1(\mathbb{E}[\boldsymbol{M}])\right|}{\sqrt{\sigma_1(\boldsymbol{M})} + 1}.
\end{aligned} \tag{230}$$

In addition, we can apply Weyl's inequality once again to deduce that

$$\left|\sigma_1(\boldsymbol{M}) - \sigma_1(\mathbb{E}[\boldsymbol{M}])\right| \le \|\boldsymbol{M} - \mathbb{E}[\boldsymbol{M}]\| \le \xi, \tag{231}$$

where the last inequality comes from (228). Substitute (231) into (230) to obtain

$$\left\|\alpha\boldsymbol{h}^0 - \boldsymbol{h}^\star\right\|_2 + \left\|\alpha\boldsymbol{x}^0 - \boldsymbol{x}^\star\right\|_2 \le \left\|\alpha\check{\boldsymbol{h}}^0 - \boldsymbol{h}^\star\right\|_2 + \left\|\alpha\check{\boldsymbol{x}}^0 - \boldsymbol{x}^\star\right\|_2 + 2\xi. \tag{232}$$

Taking the minimum over α, one can thus conclude that

$$\begin{aligned}
&\min_{\alpha\in\mathbb{C},|\alpha|=1} \left\{ \left\|\alpha\boldsymbol{h}^0 - \boldsymbol{h}^\star\right\|_2 + \left\|\alpha\boldsymbol{x}^0 - \boldsymbol{x}^\star\right\|_2 \right\} \\
&\le \min_{\alpha\in\mathbb{C},|\alpha|=1} \left\{ \left\|\alpha\check{\boldsymbol{h}}^0 - \boldsymbol{h}^\star\right\|_2 + \left\|\alpha\check{\boldsymbol{x}}^0 - \boldsymbol{x}^\star\right\|_2 \right\} + 2\xi \le 2c_1\xi + 2\xi,
\end{aligned}$$

where the last inequality comes from (229). Since ξ is arbitrary, by taking $m/(\mu^2 K \log^2 m)$ to be large enough, we finish the proof for (92). Carrying out similar arguments (which we omit here), we can also establish (93).

The last claim in Lemma 19 that $||\alpha_0| - 1| \le 1/4$ is a direct corollary of (92) and Lemma 52.

C.6 Proof of Lemma 20

The proof is composed of three steps:

- In the first step, we show that the normalized singular vectors of $\boldsymbol{M}$ and $\boldsymbol{M}^{(l)}$ are close enough; see (240).
- We then proceed by passing this proximity result to the scaled singular vectors; see (243).
- Finally, we translate the usual ℓ_2 distance metric to the distance function we defined in (34); see (245). Along the way, we also prove the incoherence of $\boldsymbol{h}^0$ with respect to $\{\boldsymbol{b}_l\}$.

Here comes the formal proof. Recall that $\check{\boldsymbol{h}}^0$ and $\check{\boldsymbol{x}}^0$ are, respectively, the leading left and right singular vectors of $\boldsymbol{M}$, and $\check{\boldsymbol{h}}^{0,(l)}$ and $\check{\boldsymbol{x}}^{0,(l)}$ are, respectively, the leading left and right singular vectors of $\boldsymbol{M}^{(l)}$. Invoke Wedin's sinΘ theorem [42, Theorem 2.1] to obtain

$$\begin{aligned}&\min_{\alpha\in\mathbb{C},|\alpha|=1}\left\{\left\|\alpha\check{\boldsymbol{h}}^0-\check{\boldsymbol{h}}^{0,(l)}\right\|_2+\left\|\alpha\check{\boldsymbol{x}}^0-\check{\boldsymbol{x}}^{0,(l)}\right\|_2\right\}\\&\quad\le c_1\frac{\left\|\left(\boldsymbol{M}-\boldsymbol{M}^{(l)}\right)\check{\boldsymbol{x}}^{0,(l)}\right\|_2+\left\|\check{\boldsymbol{h}}^{0,(l)\mathsf{H}}\left(\boldsymbol{M}-\boldsymbol{M}^{(l)}\right)\right\|_2}{\sigma_1\left(\boldsymbol{M}^{(l)}\right)-\sigma_2\left(\boldsymbol{M}\right)}\end{aligned}$$

for some universal constant $c_1>0$. Using the Weyl's inequality we get

$$\begin{aligned}\sigma_1\left(\boldsymbol{M}^{(l)}\right)-\sigma_2\left(\boldsymbol{M}\right)&\ge\sigma_1\left(\mathbb{E}[\boldsymbol{M}^{(l)}]\right)-\|\boldsymbol{M}^{(l)}-\mathbb{E}[\boldsymbol{M}^{(l)}]\|-\sigma_2\left(\mathbb{E}[\boldsymbol{M}]\right)-\|\boldsymbol{M}-\mathbb{E}[\boldsymbol{M}]\|\\&\ge 3/4-\|\boldsymbol{M}^{(l)}-\mathbb{E}[\boldsymbol{M}^{(l)}]\|-\|\boldsymbol{M}-\mathbb{E}[\boldsymbol{M}]\|\ge 1/2,\end{aligned}$$

where the penultimate inequality follows from

$$\sigma_1\left(\mathbb{E}[\boldsymbol{M}^{(l)}]\right)\ge 3/4$$

for m sufficiently large, and the last inequality comes from [76, Lemma 5.20], provided that $m\ge c_2\mu^2K\log^2 m$ for some sufficiently large constant $c_2>0$. As a result, denoting

$$\beta^{0,(l)}:=\operatorname*{argmin}_{\alpha\in\mathbb{C},|\alpha|=1}\left\{\left\|\alpha\check{\boldsymbol{h}}^0-\check{\boldsymbol{h}}^{0,(l)}\right\|_2+\left\|\alpha\check{\boldsymbol{x}}^0-\check{\boldsymbol{x}}^{0,(l)}\right\|_2\right\}\tag{233}$$

allows us to obtain

$$\begin{aligned}&\left\|\beta^{0,(l)}\check{\boldsymbol{h}}^0-\check{\boldsymbol{h}}^{0,(l)}\right\|_2+\left\|\beta^{0,(l)}\check{\boldsymbol{x}}^0-\check{\boldsymbol{x}}^{0,(l)}\right\|_2\\&\quad\le 2c_1\left\{\left\|\left(\boldsymbol{M}-\boldsymbol{M}^{(l)}\right)\check{\boldsymbol{x}}^{0,(l)}\right\|_2+\left\|\check{\boldsymbol{h}}^{0,(l)\mathsf{H}}\left(\boldsymbol{M}-\boldsymbol{M}^{(l)}\right)\right\|_2\right\}.\end{aligned}\tag{234}$$

It then boils down to controlling the two terms on the right-hand side of (234). By construction,

$$\boldsymbol{M}-\boldsymbol{M}^{(l)}=\boldsymbol{b}_l\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_l\boldsymbol{a}_l^{\mathsf{H}}.$$

- To bound the first term, observe that

$$\left\|(\boldsymbol{M}-\boldsymbol{M}^{(l)})\check{\boldsymbol{x}}^{0,(l)}\right\|_2 = \left\|\boldsymbol{b}_l\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_l\boldsymbol{a}_l^{\mathsf{H}}\check{\boldsymbol{x}}^{0,(l)}\right\|_2 = \|\boldsymbol{b}_l\|_2\left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^\star\right|\left|\boldsymbol{a}_l^{\mathsf{H}}\boldsymbol{x}^\star\right|\cdot\left|\boldsymbol{a}_l^{\mathsf{H}}\check{\boldsymbol{x}}^{0,(l)}\right| \le 30\frac{\mu}{\sqrt{m}}\cdot\sqrt{\frac{K\log^2 m}{m}}, \tag{235}$$

  where we use the fact that $\|\boldsymbol{b}_l\|_2=\sqrt{K/m}$, the incoherence condition (36), bound (189), and the fact that with probability exceeding $1-O\left(m^{-10}\right)$,

$$\max_{1\le l\le m}\left|\boldsymbol{a}_l^{\mathsf{H}}\check{\boldsymbol{x}}^{0,(l)}\right| \le 5\sqrt{\log m},$$

  due to the independence between $\check{\boldsymbol{x}}^{0,(l)}$ and $\boldsymbol{a}_l$.
- To bound the second term, for any $\widetilde{\alpha}$ obeying $|\widetilde{\alpha}|=1$, one has

$$\begin{aligned}
&\left\|\check{\boldsymbol{h}}^{0,(l)\mathsf{H}}\left(\boldsymbol{M}-\boldsymbol{M}^{(l)}\right)\right\|_2 \\
&= \left\|\check{\boldsymbol{h}}^{0,(l)\mathsf{H}}\boldsymbol{b}_l\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_l\boldsymbol{a}_l^{\mathsf{H}}\right\|_2 = \|\boldsymbol{a}_l\|_2\left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^\star\right|\left|\boldsymbol{a}_l^{\mathsf{H}}\boldsymbol{x}^\star\right|\cdot\left|\boldsymbol{b}_l^{\mathsf{H}}\check{\boldsymbol{h}}^{0,(l)}\right| \\
&\overset{\text{(i)}}{\le} 3\sqrt{K}\cdot\frac{\mu}{\sqrt{m}}\cdot 5\sqrt{\log m}\cdot\left|\boldsymbol{b}_l^{\mathsf{H}}\check{\boldsymbol{h}}^{0,(l)}\right| \\
&\overset{\text{(ii)}}{\le} 15\sqrt{\frac{\mu^2K\log m}{m}}\left|\widetilde{\alpha}\boldsymbol{b}_l^{\mathsf{H}}\check{\boldsymbol{h}}^0\right| + 15\sqrt{\frac{\mu^2K\log m}{m}}\left|\boldsymbol{b}_l^{\mathsf{H}}\left(\widetilde{\alpha}\check{\boldsymbol{h}}^0-\check{\boldsymbol{h}}^{0,(l)}\right)\right| \\
&\overset{\text{(iii)}}{\le} 15\sqrt{\frac{\mu^2K\log m}{m}}\left|\boldsymbol{b}_l^{\mathsf{H}}\check{\boldsymbol{h}}^0\right| + 15\sqrt{\frac{\mu^2K\log m}{m}}\cdot\sqrt{\frac{K}{m}}\left\|\widetilde{\alpha}\check{\boldsymbol{h}}^0-\check{\boldsymbol{h}}^{0,(l)}\right\|_2.
\end{aligned}$$

  Here, (i) arises from the incoherence condition (36) together with bounds (189) and (190), inequality (ii) comes from the triangle inequality, and the last line (iii) holds since $\|\boldsymbol{b}_l\|_2=\sqrt{K/m}$ and $|\widetilde{\alpha}|=1$.

Substitution of the above bounds into (234) yields

$$\begin{aligned}
&\left\|\beta^{0,(l)}\check{\boldsymbol{h}}^0-\check{\boldsymbol{h}}^{0,(l)}\right\|_2 + \left\|\beta^{0,(l)}\check{\boldsymbol{x}}^0-\check{\boldsymbol{x}}^{0,(l)}\right\|_2 \\
&\le 2c_1\left\{30\frac{\mu}{\sqrt{m}}\cdot\sqrt{\frac{K\log^2 m}{m}} + 15\sqrt{\frac{\mu^2K\log m}{m}}\left|\boldsymbol{b}_l^{\mathsf{H}}\check{\boldsymbol{h}}^0\right|\right. \\
&\quad\left. + 15\sqrt{\frac{\mu^2K\log m}{m}}\cdot\sqrt{\frac{K}{m}}\left\|\widetilde{\alpha}\check{\boldsymbol{h}}^0-\check{\boldsymbol{h}}^{0,(l)}\right\|_2\right\}.
\end{aligned}$$

Since the previous inequality holds for all $|\widetilde{\alpha}| = 1$, we can choose $\widetilde{\alpha} = \beta^{0,(l)}$ and rearrange terms to get

$$\left(1 - 30c_1\sqrt{\frac{\mu^2 K\log m}{m}}\sqrt{\frac{K}{m}}\right)\left(\left\|\beta^{0,(l)}\check{\boldsymbol{h}}^0 - \check{\boldsymbol{h}}^{0,(l)}\right\|_2 + \left\|\beta^{0,(l)}\check{\boldsymbol{x}}^0 - \check{\boldsymbol{x}}^{0,(l)}\right\|_2\right)$$
$$\le 60c_1\frac{\mu}{\sqrt{m}}\cdot\sqrt{\frac{K\log^2 m}{m}} + 30c_1\sqrt{\frac{\mu^2 K\log m}{m}}\left|\boldsymbol{b}_l^{\mathsf{H}}\check{\boldsymbol{h}}^0\right|.$$

Under the condition that $m \gg \mu K\log^{1/2} m$, one has $1 - 30c_1\sqrt{\mu^2 K\log m/m}\cdot\sqrt{K/m} \ge \frac{1}{2}$, and therefore,

$$\left\|\beta^{0,(l)}\check{\boldsymbol{h}}^0 - \check{\boldsymbol{h}}^{0,(l)}\right\|_2 + \left\|\beta^{0,(l)}\check{\boldsymbol{x}}^0 - \check{\boldsymbol{x}}^{0,(l)}\right\|_2$$
$$\le 120c_1\frac{\mu}{\sqrt{m}}\cdot\sqrt{\frac{K\log^2 m}{m}} + 60c_1\sqrt{\frac{\mu^2 K\log m}{m}}\left|\boldsymbol{b}_l^{\mathsf{H}}\check{\boldsymbol{h}}^0\right|,$$

which immediately implies that

$$\max_{1\le l\le m}\left\{\left\|\beta^{0,(l)}\check{\boldsymbol{h}}^0 - \check{\boldsymbol{h}}^{0,(l)}\right\|_2 + \left\|\beta^{0,(l)}\check{\boldsymbol{x}}^0 - \check{\boldsymbol{x}}^{0,(l)}\right\|_2\right\}$$
$$\le 120c_1\frac{\mu}{\sqrt{m}}\cdot\sqrt{\frac{K\log^2 m}{m}} + 60c_1\sqrt{\frac{\mu^2 K\log m}{m}}\max_{1\le l\le m}\left|\boldsymbol{b}_l^{\mathsf{H}}\check{\boldsymbol{h}}^0\right|. \tag{236}$$

We then move on to $\left|\boldsymbol{b}_l^{\mathsf{H}}\check{\boldsymbol{h}}^0\right|$. The aim is to show that $\max_{1\le l\le m}\left|\boldsymbol{b}_l^{\mathsf{H}}\check{\boldsymbol{h}}^0\right|$ can also be upper bounded by the left-hand side of (236). By construction, we have $\boldsymbol{M}\check{\boldsymbol{x}}^0 = \sigma_1(\boldsymbol{M})\check{\boldsymbol{h}}^0$, which further leads to

$$\left|\boldsymbol{b}_l^{\mathsf{H}}\check{\boldsymbol{h}}^0\right| = \frac{1}{\sigma_1(\boldsymbol{M})}\left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{M}\check{\boldsymbol{x}}^0\right|$$
$$\overset{\text{(i)}}{\le} 2\left|\sum_{j=1}^m\left(\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{b}_j\right)\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\check{\boldsymbol{x}}^0\right|$$
$$\le 2\left(\sum_{j=1}^m\left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{b}_j\right|\right)\max_{1\le j\le m}\left\{\left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\right|\left|\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}^\star\right|\left|\boldsymbol{a}_j^{\mathsf{H}}\check{\boldsymbol{x}}^0\right|\right\}$$
$$\overset{\text{(ii)}}{\le} 8\log m\cdot\frac{\mu}{\sqrt{m}}\cdot\left(5\sqrt{\log m}\right)\max_{1\le j\le m}\left\{\left|\boldsymbol{a}_j^{\mathsf{H}}\check{\boldsymbol{x}}^{0,(j)}\right| + \|\boldsymbol{a}_j\|_2\left\|\beta^{0,(j)}\check{\boldsymbol{x}}^0 - \check{\boldsymbol{x}}^{0,(j)}\right\|_2\right\}$$
$$\le 200\frac{\mu\log^2 m}{\sqrt{m}} + 120\sqrt{\frac{\mu^2 K\log^3 m}{m}}\max_{1\le j\le m}\left\|\beta^{0,(j)}\check{\boldsymbol{x}}^0 - \check{\boldsymbol{x}}^{0,(j)}\right\|_2, \tag{237}$$

where $\beta^{0,(j)}$ is as defined in (233). Here, (i) comes from the lower bound $\sigma_1(\boldsymbol{M}) \geq 1/2$. Bound (ii) follows by combining the incoherence condition (36), bound (189), the triangle inequality, as well as the estimate $\sum_{j=1}^{m} \left|\boldsymbol{b}_l^{\mathsf{H}} \boldsymbol{b}_j\right| \leq 4 \log m$ from Lemma 48. The last line uses the upper estimate $\max_{1\leq j\leq m} \left|\boldsymbol{a}_j^{\mathsf{H}} \check{\boldsymbol{x}}^{0,(j)}\right| \leq 5\sqrt{\log m}$ and (190). Our bound (237) further implies

$$\max_{1\leq l\leq m} \left|\boldsymbol{b}_l^{\mathsf{H}} \check{\boldsymbol{h}}^0\right| \leq 200 \frac{\mu \log^2 m}{\sqrt{m}} + 120\sqrt{\frac{\mu^2 K \log^3 m}{m}} \max_{1\leq j\leq m} \left\|\beta^{0,(j)} \check{\boldsymbol{x}}^0 - \check{\boldsymbol{x}}^{0,(j)}\right\|_2. \tag{238}$$

The above bound (238) taken together with (236) gives

$$\begin{aligned}\max_{1\leq l\leq m} \left\{\left\|\beta^{0,(l)} \check{\boldsymbol{h}}^0 - \check{\boldsymbol{h}}^{0,(l)}\right\|_2 + \left\|\beta^{0,(l)} \check{\boldsymbol{x}}^0 - \check{\boldsymbol{x}}^{0,(l)}\right\|_2\right\} &\leq 120 c_1 \frac{\mu}{\sqrt{m}} \cdot \sqrt{\frac{K \log^2 m}{m}} \\ &+ 60 c_1 \sqrt{\frac{\mu^2 K \log m}{m}} \left(200 \frac{\mu \log^2 m}{\sqrt{m}} + 120\sqrt{\frac{\mu^2 K \log^3 m}{m}} \max_{1\leq j\leq m} \left\|\beta^{0,(j)} \check{\boldsymbol{x}}^0 - \check{\boldsymbol{x}}^{0,(j)}\right\|_2\right).\end{aligned} \tag{239}$$

As long as $m \gg \mu^2 K \log^2 m$, we have $60 c_1 \sqrt{\mu^2 K \log m/m} \cdot 120\sqrt{\mu^2 K \log^3 m/m} \leq 1/2$. Rearranging terms, we are left with

$$\max_{1\leq l\leq m} \left\{\left\|\beta^{0,(l)} \check{\boldsymbol{h}}^0 - \check{\boldsymbol{h}}^{0,(l)}\right\|_2 + \left\|\beta^{0,(l)} \check{\boldsymbol{x}}^0 - \check{\boldsymbol{x}}^{0,(l)}\right\|_2\right\} \leq c_3 \frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^5 m}{m}} \tag{240}$$

for some constant $c_3 > 0$. Further, this bound combined with (238) yields

$$\max_{1\leq l\leq m} \left|\boldsymbol{b}_l^{\mathsf{H}} \check{\boldsymbol{h}}^0\right| \leq 200 \frac{\mu \log^2 m}{\sqrt{m}} + 120\sqrt{\frac{\mu^2 K \log^3 m}{m}} \cdot c_3 \frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^5 m}{m}} \leq c_2 \frac{\mu \log^2 m}{\sqrt{m}} \tag{241}$$

for some constant $c_2 > 0$, with the proviso that $m \gg \mu^2 K \log^2 m$.

We now translate the preceding bounds to the scaled version. Recall from bound (231) that

$$1/2 \leq 1 - \xi \leq \|\boldsymbol{M}\| = \sigma_1(\boldsymbol{M}) \leq 1 + \xi \leq 2, \tag{242}$$

as long as $\xi \leq 1/2$. For any $\alpha \in \mathbb{C}$ with $|\alpha| = 1$, $\alpha \check{\boldsymbol{h}}^0, \alpha \check{\boldsymbol{x}}^0$ are still the leading left and right singular vectors of $\boldsymbol{M}$. Hence, we can use Lemma 60 to derive that

$$\begin{aligned}&\left|\sigma_1(\boldsymbol{M}) - \sigma_1\big(\boldsymbol{M}^{(l)}\big)\right| \\ &\leq \left\|\big(\boldsymbol{M} - \boldsymbol{M}^{(l)}\big)\check{\boldsymbol{x}}^{0,(l)}\right\|_2 + \left\{\left\|\alpha \check{\boldsymbol{h}}^0 - \check{\boldsymbol{h}}^{0,(l)}\right\|_2 + \left\|\alpha \check{\boldsymbol{x}}^0 - \check{\boldsymbol{x}}^{0,(l)}\right\|_2\right\} \|\boldsymbol{M}\| \\ &\leq \left\|\big(\boldsymbol{M} - \boldsymbol{M}^{(l)}\big)\check{\boldsymbol{x}}^{0,(l)}\right\|_2 + 2\left\{\left\|\alpha \check{\boldsymbol{h}}^0 - \check{\boldsymbol{h}}^{0,(l)}\right\|_2 + \left\|\alpha \check{\boldsymbol{x}}^0 - \check{\boldsymbol{x}}^{0,(l)}\right\|_2\right\}\end{aligned}$$

and

$$\begin{aligned}
&\left\|\alpha \boldsymbol{h}^0-\boldsymbol{h}^{0,(l)}\right\|_2+\left\|\alpha \boldsymbol{x}^0-\boldsymbol{x}^{0,(l)}\right\|_2 \\
&=\left\|\sqrt{\sigma_1(\boldsymbol{M})}\left(\alpha \check{\boldsymbol{h}}^0\right)-\sqrt{\sigma_1\left(\boldsymbol{M}^{(l)}\right)}\check{\boldsymbol{h}}^{0,(l)}\right\|_2+\left\|\sqrt{\sigma_1(\boldsymbol{M})}\alpha \check{\boldsymbol{x}}^0-\sqrt{\sigma_1\left(\boldsymbol{M}^{(l)}\right)}\check{\boldsymbol{x}}^{0,(l)}\right\|_2 \\
&\leq \sqrt{\sigma_1(\boldsymbol{M})}\left\{\left\|\alpha \check{\boldsymbol{h}}^0-\check{\boldsymbol{h}}^{0,(l)}\right\|_2+\left\|\alpha \check{\boldsymbol{x}}^0-\check{\boldsymbol{x}}^{0,(l)}\right\|_2\right\}+\frac{2\left|\sigma_1(\boldsymbol{M})-\sigma_1(\boldsymbol{M}^{(l)})\right|}{\sqrt{\sigma_1(\boldsymbol{M})}+\sqrt{\sigma_1(\boldsymbol{M}^{(l)})}} \\
&\leq \sqrt{2}\left\{\left\|\alpha \check{\boldsymbol{h}}^0-\check{\boldsymbol{h}}^{0,(l)}\right\|_2+\left\|\alpha \check{\boldsymbol{x}}^0-\check{\boldsymbol{x}}^{0,(l)}\right\|_2\right\}+\sqrt{2}\left|\sigma_1(\boldsymbol{M})-\sigma_1(\boldsymbol{M}^{(l)})\right|.
\end{aligned}$$

Taking the previous two bounds collectively yields

$$\begin{aligned}
&\left\|\alpha \boldsymbol{h}^0-\boldsymbol{h}^{0,(l)}\right\|_2+\left\|\alpha \boldsymbol{x}^0-\boldsymbol{x}^{0,(l)}\right\|_2 \leq \sqrt{2}\left\|\left(\boldsymbol{M}-\boldsymbol{M}^{(l)}\right)\check{\boldsymbol{x}}^{0,(l)}\right\|_2 \\
&\quad +6\left\{\left\|\alpha \check{\boldsymbol{h}}^0-\check{\boldsymbol{h}}^{0,(l)}\right\|_2+\left\|\alpha \check{\boldsymbol{x}}^0-\check{\boldsymbol{x}}^{0,(l)}\right\|_2\right\},
\end{aligned}$$

which together with (235) and (240) implies

$$\min_{\alpha\in\mathbb{C},|\alpha|=1}\left\{\left\|\alpha \boldsymbol{h}^0-\boldsymbol{h}^{0,(l)}\right\|_2+\left\|\alpha \boldsymbol{x}^0-\boldsymbol{x}^{0,(l)}\right\|_2\right\} \leq c_5\frac{\mu}{\sqrt{m}}\sqrt{\frac{\mu^2 K\log^5 m}{m}} \tag{243}$$

for some constant $c_5>0$, as long as ξ is sufficiently small. Moreover, we have

$$\left\|\frac{1}{\overline{\alpha^0}}\boldsymbol{h}^0-\frac{\alpha}{\overline{\alpha^0}}\boldsymbol{h}^{0,(l)}\right\|_2+\left\|\alpha^0\boldsymbol{x}^0-\alpha\alpha^0\boldsymbol{x}^{0,(l)}\right\|_2 \leq 2\left\{\left\|\boldsymbol{h}^0-\alpha\boldsymbol{h}^{0,(l)}\right\|_2+\left\|\boldsymbol{x}^0-\alpha\boldsymbol{x}^{0,(l)}\right\|_2\right\}$$

for any $|\alpha|=1$, where α^0 is defined in (38) and, according to Lemma 19, satisfies

$$1/2\leq|\alpha^0|\leq 2. \tag{244}$$

Therefore,

$$\begin{aligned}
&\min_{\alpha\in\mathbb{C},|\alpha|=1}\sqrt{\left\|\frac{1}{\overline{\alpha^0}}\boldsymbol{h}^0-\frac{\alpha}{\overline{\alpha^0}}\boldsymbol{h}^{0,(l)}\right\|_2^2+\left\|\alpha^0\boldsymbol{x}^0-\alpha\alpha^0\boldsymbol{x}^{0,(l)}\right\|_2^2} \\
&\leq \min_{\alpha\in\mathbb{C},|\alpha|=1}\left\{\left\|\frac{1}{\overline{\alpha^0}}\boldsymbol{h}^0-\frac{\alpha}{\overline{\alpha^0}}\boldsymbol{h}^{0,(l)}\right\|_2+\left\|\alpha^0\boldsymbol{x}^0-\alpha\alpha^0\boldsymbol{x}^{0,(l)}\right\|_2\right\} \\
&\leq 2\min_{\alpha\in\mathbb{C},|\alpha|=1}\left\{\left\|\boldsymbol{h}^0-\alpha\boldsymbol{h}^{0,(l)}\right\|_2+\left\|\boldsymbol{x}^0-\alpha\boldsymbol{x}^{0,(l)}\right\|_2\right\} \\
&\leq 2c_5\frac{\mu}{\sqrt{m}}\sqrt{\frac{\mu^2 K\log^5 m}{m}}.
\end{aligned}$$

Furthermore, we have

$$\begin{aligned}\operatorname{dist}\left(z^{0,(l)},\widetilde{z}^{0}\right) &= \min_{\alpha\in\mathbb{C}}\sqrt{\left\|\frac{1}{\overline{\alpha}}\boldsymbol{h}^{0,(l)}-\frac{1}{\overline{\alpha^{0}}}\boldsymbol{h}^{0}\right\|_2^2+\left\|\alpha\boldsymbol{x}^{0,(l)}-\alpha^{0}\boldsymbol{x}^{0}\right\|_2^2}\\ &\le \min_{\alpha\in\mathbb{C},|\alpha|=1}\sqrt{\left\|\frac{1}{\overline{\alpha^{0}}}\boldsymbol{h}^{0}-\frac{\alpha}{\overline{\alpha^{0}}}\boldsymbol{h}^{0,(l)}\right\|_2^2+\left\|\alpha^{0}\boldsymbol{x}^{0}-\alpha\alpha^{0}\boldsymbol{x}^{0,(l)}\right\|_2^2}\\ &\le 2c_5\frac{\mu}{\sqrt{m}}\sqrt{\frac{\mu^2K\log^5 m}{m}}, \end{aligned} \tag{245}$$

where the second line follows since the latter is minimizing over a smaller feasible set. This completes the proof for claim (96).

Regarding $\left|\boldsymbol{b}_l^{\mathsf{H}}\widetilde{\boldsymbol{h}}^{0}\right|$, one first sees that

$$\left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^{0}\right| = \left|\sqrt{\sigma_1(\boldsymbol{M})}\boldsymbol{b}_l^{\mathsf{H}}\check{\boldsymbol{h}}^{0}\right| \le \sqrt{2}c_2\frac{\mu\log^2 m}{\sqrt{m}},$$

where the last relation holds due to (241) and (242). Hence, using the property (244), we have

$$\left|\boldsymbol{b}_l^{\mathsf{H}}\widetilde{\boldsymbol{h}}^{0}\right| = \left|\boldsymbol{b}_l^{\mathsf{H}}\frac{1}{\overline{\alpha^{0}}}\boldsymbol{h}^{0}\right| \le \left|\frac{1}{\overline{\alpha^{0}}}\right|\left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^{0}\right| \le 2\sqrt{2}c_2\frac{\mu\log^2 m}{\sqrt{m}},$$

which finishes the proof of claim (97).

Before concluding this section, we note a by-product of the proof. Specifically, we can establish the claim required in (226) using many results derived in this section. This is formally stated in the following lemma.

Lemma 30 *Fix any small constant $c > 0$. Suppose the number of samples obeys $m \gg \tau K\log^4 m$. Then with probability at least $1 - O\left(m^{-10}\right)$, we have*

$$\max_{1\le j\le\tau}\left|\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\widetilde{\boldsymbol{h}}^{0}\right| \le c\frac{\mu}{\sqrt{m}}\log m.$$

Proof Instate the notation and hypotheses in Appendix C.6. Recognize that

$$\begin{aligned}\left|\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\widetilde{\boldsymbol{h}}^{0}\right| &= \left|\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\frac{1}{\overline{\alpha^{0}}}\boldsymbol{h}^{0}\right| = \left|\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\frac{1}{\overline{\alpha^{0}}}\sqrt{\sigma_1(\boldsymbol{M})}\check{\boldsymbol{h}}^{0}\right|\\ &\le \left|\frac{1}{\overline{\alpha^{0}}}\right|\sqrt{\sigma_1(\boldsymbol{M})}\left|\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\check{\boldsymbol{h}}^{0}\right|\\ &\le 4\left|\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\check{\boldsymbol{h}}^{0}\right|,\end{aligned}$$

where the last inequality comes from (242) and (244). It thus suffices to prove that $\left|\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\check{\boldsymbol{h}}^{0}\right| \le c\mu\log m/\sqrt{m}$ for some $c > 0$ small enough. To this end, it can be

seen that

$$\begin{aligned}\left|\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\check{\boldsymbol{h}}^0\right| &= \frac{1}{\sigma_1\left(\boldsymbol{M}\right)}\left|\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\boldsymbol{M}\check{\boldsymbol{x}}^0\right| \\ &\le 2\left|\sum_{k=1}^{m}\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\boldsymbol{b}_k\boldsymbol{b}_k^{\mathsf{H}}\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_k\boldsymbol{a}_k^{\mathsf{H}}\check{\boldsymbol{x}}^0\right| \\ &\le 2\left(\sum_{k=1}^{m}\left|\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\boldsymbol{b}_k\right|\right)\max_{1\le k\le m}\left\{\left|\boldsymbol{b}_k^{\mathsf{H}}\boldsymbol{h}^\star\right|\left|\boldsymbol{a}_k^{\mathsf{H}}\boldsymbol{x}^\star\right|\left|\boldsymbol{a}_k^{\mathsf{H}}\check{\boldsymbol{x}}^0\right|\right\} \\ &\overset{\text{(i)}}{\le} c\frac{1}{\log^2 m}\cdot\frac{\mu}{\sqrt{m}}\cdot\left(5\sqrt{\log m}\right)\max_{1\le j\le m}\left\{\left|\boldsymbol{a}_j^{\mathsf{H}}\check{\boldsymbol{x}}^{0,(j)}\right|\right. \\ &\quad\left.+\left\|\boldsymbol{a}_j\right\|_2\left\|\alpha^{0,(j)}\check{\boldsymbol{x}}^0-\check{\boldsymbol{x}}^{0,(j)}\right\|_2\right\} \\ &\overset{\text{(ii)}}{\lesssim} c\frac{\mu}{\sqrt{m}}\frac{1}{\log m}\le c\frac{\mu}{\sqrt{m}}\log m,\end{aligned}\tag{246}$$

where (i) comes from Lemma 50, the incoherence condition (36), and estimate (189). The last line (ii) holds since we have already established (see (237) and (240))

$$\max_{1\le j\le m}\left\{\left|\boldsymbol{a}_j^{\mathsf{H}}\check{\boldsymbol{x}}^{0,(j)}\right|+\left\|\boldsymbol{a}_j\right\|_2\left\|\alpha^{0,(j)}\check{\boldsymbol{x}}^0-\check{\boldsymbol{x}}^{0,(j)}\right\|_2\right\}\lesssim\sqrt{\log m}.$$

The proof is then complete. □

C.7 Proof of Lemma 21

Recall that α^0 and $\alpha^{0,(l)}$ are the alignment parameters between $\boldsymbol{z}^0$ and $\boldsymbol{z}^\star$, and between $\boldsymbol{z}^{0,(l)}$ and $\boldsymbol{z}^\star$, respectively, that is,

$$\begin{aligned}\alpha^0 &:= \operatorname*{argmin}_{\alpha\in\mathbb{C}}\left\{\left\|\frac{1}{\overline{\alpha}}\boldsymbol{h}^0-\boldsymbol{h}^\star\right\|_2^2+\left\|\alpha\boldsymbol{x}^0-\boldsymbol{x}^\star\right\|_2^2\right\},\\ \alpha^{0,(l)} &:= \operatorname*{argmin}_{\alpha\in\mathbb{C}}\left\{\left\|\frac{1}{\overline{\alpha}}\boldsymbol{h}^{0,(l)}-\boldsymbol{h}^\star\right\|_2^2+\left\|\alpha\boldsymbol{x}^{0,(l)}-\boldsymbol{x}^\star\right\|_2^2\right\}.\end{aligned}$$

Also, we let

$$\alpha_{\text{mutual}}^{0,(l)} := \operatorname*{argmin}_{\alpha\in\mathbb{C}}\left\{\left\|\frac{1}{\overline{\alpha}}\boldsymbol{h}^{0,(l)}-\frac{1}{\overline{\alpha^0}}\boldsymbol{h}^0\right\|_2^2+\left\|\alpha\boldsymbol{x}^{0,(l)}-\alpha^0\boldsymbol{x}^0\right\|_2^2\right\}.$$

The triangle inequality together with (94) and (245) then tells us that

$$\sqrt{\left\|\frac{1}{\overline{\alpha_{\text{mutual}}^{0,(l)}}}\boldsymbol{h}^{0,(l)}-\boldsymbol{h}^\star\right\|_2^2+\left\|\alpha_{\text{mutual}}^{0,(l)}\boldsymbol{x}^{0,(l)}-\boldsymbol{x}^\star\right\|_2^2}$$

$$\begin{aligned}
&\le \sqrt{\left\| \frac{1}{\overline{\alpha^0}} \boldsymbol{h}^0 - \frac{1}{\overline{\alpha^{0,(l)}_{\text{mutual}}}} \boldsymbol{h}^{0,(l)} \right\|_2^2 + \left\| \alpha^0 \boldsymbol{x}^0 - \alpha^{0,(l)}_{\text{mutual}} \boldsymbol{x}^{0,(l)} \right\|_2^2} \\
&\quad + \sqrt{\left\| \frac{1}{\overline{\alpha^0}} \boldsymbol{h}^0 - \boldsymbol{h}^\star \right\|_2^2 + \left\| \alpha^0 \boldsymbol{x}^0 - \boldsymbol{x}^\star \right\|_2^2} \\
&\le 2c_5 \frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^5 m}{m}} + C_1 \frac{1}{\log^2 m} \\
&\le 2C_1 \frac{1}{\log^2 m},
\end{aligned}$$

where the last relation holds as long as $m \gg \mu^2 \sqrt{K} \log^{9/2} m$.

Let

$$\boldsymbol{x}_1 = \alpha^0 \boldsymbol{x}^0, \quad \boldsymbol{h}_1 = \frac{1}{\overline{\alpha^0}} \boldsymbol{h}_0 \qquad \text{and} \qquad \boldsymbol{x}_2 = \alpha^{0,(l)}_{\text{mutual}} \boldsymbol{x}^{0,(l)}, \quad \boldsymbol{h}_2 = \frac{1}{\overline{\alpha^{0,(l)}_{\text{mutual}}}} \boldsymbol{h}^{0,(l)}.$$

It is easy to see that $\boldsymbol{x}_1, \boldsymbol{h}_1, \boldsymbol{x}_2, \boldsymbol{h}_2$ satisfy the assumptions in Lemma 55, which implies

$$\begin{aligned}
&\sqrt{\left\| \frac{1}{\overline{\alpha^{0,(l)}}} \boldsymbol{h}^{0,(l)} - \frac{1}{\overline{\alpha^0}} \boldsymbol{h}^0 \right\|_2^2 + \left\| \alpha^{0,(l)} \boldsymbol{x}^{0,(l)} - \alpha^0 \boldsymbol{x}^0 \right\|_2^2} \\
&\lesssim \sqrt{\left\| \frac{1}{\overline{\alpha^0}} \boldsymbol{h}^0 - \frac{1}{\overline{\alpha^{0,(l)}_{\text{mutual}}}} \boldsymbol{h}^{0,(l)} \right\|_2^2 + \left\| \alpha^0 \boldsymbol{x}^0 - \alpha^{0,(l)}_{\text{mutual}} \boldsymbol{x}^{0,(l)} \right\|_2^2} \\
&\lesssim \frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^5 m}{m}},
\end{aligned} \tag{247}$$

where the last line comes from (245). With this upper estimate at hand, we are now ready to show that with high probability,

$$\begin{aligned}
\left| \boldsymbol{a}_l^{\mathsf{H}} \left(\alpha^0 \boldsymbol{x}^0 - \boldsymbol{x}^\star \right) \right| &\overset{\text{(i)}}{\le} \left| \boldsymbol{a}_l^{\mathsf{H}} \left(\alpha^{0,(l)} \boldsymbol{x}^{0,(l)} - \boldsymbol{x}^\star \right) \right| + \left| \boldsymbol{a}_l^{\mathsf{H}} \left(\alpha^0 \boldsymbol{x}^0 - \alpha^{0,(l)} \boldsymbol{x}^{0,(l)} \right) \right| \\
&\overset{\text{(ii)}}{\le} 5\sqrt{\log m} \left\| \alpha^{0,(l)} \boldsymbol{x}^{0,(l)} - \boldsymbol{x}^\star \right\|_2 + \| \boldsymbol{a}_l \|_2 \left\| \alpha^0 \boldsymbol{x}^0 - \alpha^{0,(l)} \boldsymbol{x}^{0,(l)} \right\|_2 \\
&\overset{\text{(iii)}}{\lesssim} \sqrt{\log m} \cdot \frac{1}{\log^2 m} + \sqrt{K} \frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^5 m}{m}} \\
&\overset{\text{(iv)}}{\lesssim} \frac{1}{\log^{3/2} m},
\end{aligned}$$

where (i) follows from the triangle inequality, (ii) uses Cauchy–Schwarz and the independence between $\boldsymbol{x}^{0,(l)}$ and $\boldsymbol{a}_l$, (iii) holds because of (95) and (247) under the condition $m \gg \mu^2 K \log^6 m$, and (iv) holds true as long as $m \gg \mu^2 K \log^4 m$.

D Technical Lemmas

D.1 Technical Lemmas for Phase Retrieval

D.1.1 Matrix Concentration Inequalities

Lemma 31 *Suppose that* $\boldsymbol{a}_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$ *for every* $1 \le j \le m$. *Fix any small constant* $\delta > 0$. *With probability at least* $1 - C_2 e^{-c_2 m}$, *one has*

$$\left\| \frac{1}{m} \sum_{j=1}^{m} \boldsymbol{a}_j \boldsymbol{a}_j^\top - \boldsymbol{I}_n \right\| \le \delta,$$

as long as $m \ge c_0 n$ *for some sufficiently large constant* $c_0 > 0$. *Here,* $C_2, c_2 > 0$ *are some universal constants.*

Proof This is an immediate consequence of [116, Corollary 5.35]. □

Lemma 32 *Suppose that* $\boldsymbol{a}_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, *for every* $1 \le j \le m$. *Fix any small constant* $\delta > 0$. *With probability at least* $1 - O(n^{-10})$, *we have*

$$\left\| \frac{1}{m} \sum_{j=1}^{m} \left(\boldsymbol{a}_j^\top \boldsymbol{x}^\star \right)^2 \boldsymbol{a}_j \boldsymbol{a}_j^\top - \left(\|\boldsymbol{x}^\star\|_2^2 \boldsymbol{I}_n + 2\boldsymbol{x}^\star \boldsymbol{x}^{\star\top} \right) \right\| \le \delta \|\boldsymbol{x}^\star\|_2^2,$$

provided that $m \ge c_0 n \log n$ *for some sufficiently large constant* $c_0 > 0$.

Proof This is adapted from [18, Lemma 7.4]. □

Lemma 33 *Suppose that* $\boldsymbol{a}_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, *for every* $1 \le j \le m$. *Fix any small constant* $\delta > 0$ *and any constant* $C > 0$. *Suppose* $m \ge c_0 n$ *for some sufficiently large constant* $c_0 > 0$. *Then with probability at least* $1 - C_2 e^{-c_2 m}$,

$$\left\| \frac{1}{m} \sum_{j=1}^{m} \left(\boldsymbol{a}_j^\top \boldsymbol{x} \right)^2 \mathbb{1}_{\{|\boldsymbol{a}_j^\top \boldsymbol{x}| \le C\}} \boldsymbol{a}_j \boldsymbol{a}_j^\top - \left(\beta_1 \boldsymbol{x}\boldsymbol{x}^\top + \beta_2 \|\boldsymbol{x}\|_2^2 \boldsymbol{I}_n \right) \right\| \le \delta \|\boldsymbol{x}\|_2^2, \quad \forall \boldsymbol{x} \in \mathbb{R}^n,$$

holds for some absolute constants $c_2, C_2 > 0$, *where*

$$\beta_1 := \mathbb{E}\left[\xi^4 \mathbb{1}_{\{|\xi| \le C\}} \right] - \mathbb{E}\left[\xi^2 \mathbb{1}_{|\xi| \le C} \right] \quad \text{and} \quad \beta_2 = \mathbb{E}\left[\xi^2 \mathbb{1}_{|\xi| \le C} \right]$$

with ξ *being a standard Gaussian random variable.*

Proof This is supplied in [25, supplementary material]. □

D.1.2 Matrix Perturbation Bounds

Lemma 34 *Let $\lambda_1(\boldsymbol{A})$, $\boldsymbol{u}$ be the leading eigenvalue and eigenvector of a symmetric matrix $\boldsymbol{A}$, respectively, and $\lambda_1(\widetilde{\boldsymbol{A}})$, $\widetilde{\boldsymbol{u}}$ be the leading eigenvalue and eigenvector of a symmetric matrix $\widetilde{\boldsymbol{A}}$, respectively. Suppose that $\lambda_1(\boldsymbol{A}), \lambda_1(\widetilde{\boldsymbol{A}}), \|\boldsymbol{A}\|, \|\widetilde{\boldsymbol{A}}\| \in [C_1, C_2]$ for some $C_1, C_2 > 0$. Then,*

$$\left\| \sqrt{\lambda_1(\boldsymbol{A})}\, \boldsymbol{u} - \sqrt{\lambda_1(\widetilde{\boldsymbol{A}})}\, \widetilde{\boldsymbol{u}} \right\|_2 \le \frac{\left\|(\boldsymbol{A} - \widetilde{\boldsymbol{A}})\boldsymbol{u}\right\|_2}{2\sqrt{C_1}} + \left(\sqrt{C_2} + \frac{C_2}{\sqrt{C_1}} \right) \|\boldsymbol{u} - \widetilde{\boldsymbol{u}}\|_2 .$$

Proof Observe that

$$\begin{aligned} &\left\| \sqrt{\lambda_1(\boldsymbol{A})}\, \boldsymbol{u} - \sqrt{\lambda_1(\widetilde{\boldsymbol{A}})}\, \widetilde{\boldsymbol{u}} \right\|_2 \\ &\le \left\| \sqrt{\lambda_1(\boldsymbol{A})}\, \boldsymbol{u} - \sqrt{\lambda_1(\widetilde{\boldsymbol{A}})}\, \boldsymbol{u} \right\|_2 + \left\| \sqrt{\lambda_1(\widetilde{\boldsymbol{A}})}\, \boldsymbol{u} - \sqrt{\lambda_1(\widetilde{\boldsymbol{A}})}\, \widetilde{\boldsymbol{u}} \right\|_2 \\ &\le \left| \sqrt{\lambda_1(\boldsymbol{A})} - \sqrt{\lambda_1(\widetilde{\boldsymbol{A}})} \right| + \sqrt{\lambda_1(\widetilde{\boldsymbol{A}})}\, \|\boldsymbol{u} - \widetilde{\boldsymbol{u}}\|_2 , \end{aligned} \tag{248}$$

where the last inequality follows since $\|\boldsymbol{u}\|_2 = 1$. Using the identity $\sqrt{a} - \sqrt{b} = (a-b)/(\sqrt{a} + \sqrt{b})$, we have

$$\left| \sqrt{\lambda_1(\boldsymbol{A})} - \sqrt{\lambda_1(\widetilde{\boldsymbol{A}})} \right| = \frac{\left|\lambda_1(\boldsymbol{A}) - \lambda_1(\widetilde{\boldsymbol{A}})\right|}{\left| \sqrt{\lambda_1(\boldsymbol{A})} + \sqrt{\lambda_1(\widetilde{\boldsymbol{A}})} \right|} \le \frac{\left|\lambda_1(\boldsymbol{A}) - \lambda_1(\widetilde{\boldsymbol{A}})\right|}{2\sqrt{C_1}},$$

where the last inequality comes from our assumptions on $\lambda_1(\boldsymbol{A})$ and $\lambda_1(\widetilde{\boldsymbol{A}})$. This combined with (248) yields

$$\left\| \sqrt{\lambda_1(\boldsymbol{A})}\, \boldsymbol{u} - \sqrt{\lambda_1(\widetilde{\boldsymbol{A}})}\, \widetilde{\boldsymbol{u}} \right\|_2 \le \frac{\left|\lambda_1(\boldsymbol{A}) - \lambda_1(\widetilde{\boldsymbol{A}})\right|}{2\sqrt{C_1}} + \sqrt{C_2}\, \|\boldsymbol{u} - \widetilde{\boldsymbol{u}}\|_2 . \tag{249}$$

To control $\left|\lambda_1(\boldsymbol{A}) - \lambda_1(\widetilde{\boldsymbol{A}})\right|$, use the relationship between the eigenvalue and the eigenvector to obtain

$$\begin{aligned} \left|\lambda_1(\boldsymbol{A}) - \lambda_1(\widetilde{\boldsymbol{A}})\right| &= \left| \boldsymbol{u}^\top \boldsymbol{A}\boldsymbol{u} - \widetilde{\boldsymbol{u}}^\top \widetilde{\boldsymbol{A}} \widetilde{\boldsymbol{u}} \right| \\ &\le \left| \boldsymbol{u}^\top (\boldsymbol{A} - \widetilde{\boldsymbol{A}})\boldsymbol{u} \right| + \left| \boldsymbol{u}^\top \widetilde{\boldsymbol{A}} \boldsymbol{u} - \widetilde{\boldsymbol{u}}^\top \widetilde{\boldsymbol{A}} \boldsymbol{u} \right| + \left| \widetilde{\boldsymbol{u}}^\top \widetilde{\boldsymbol{A}} \boldsymbol{u} - \widetilde{\boldsymbol{u}}^\top \widetilde{\boldsymbol{A}} \widetilde{\boldsymbol{u}} \right| \\ &\le \left\| (\boldsymbol{A} - \widetilde{\boldsymbol{A}})\boldsymbol{u} \right\|_2 + 2\, \|\boldsymbol{u} - \widetilde{\boldsymbol{u}}\|_2 \left\| \widetilde{\boldsymbol{A}} \right\|, \end{aligned}$$

which together with (249) gives

$$\left\|\sqrt{\lambda_1(\boldsymbol{A})}\,\boldsymbol{u}-\sqrt{\lambda_1(\widetilde{\boldsymbol{A}})}\,\widetilde{\boldsymbol{u}}\right\|_2 \le \frac{\left\|(\boldsymbol{A}-\widetilde{\boldsymbol{A}})\boldsymbol{u}\right\|_2+2\left\|\boldsymbol{u}-\widetilde{\boldsymbol{u}}\right\|_2\left\|\widetilde{\boldsymbol{A}}\right\|}{2\sqrt{C_1}}+\sqrt{C_2}\left\|\boldsymbol{u}-\widetilde{\boldsymbol{u}}\right\|_2$$
$$\le \frac{\left\|(\boldsymbol{A}-\widetilde{\boldsymbol{A}})\boldsymbol{u}\right\|_2}{2\sqrt{C_1}}+\left(\frac{C_2}{\sqrt{C_1}}+\sqrt{C_2}\right)\left\|\boldsymbol{u}-\widetilde{\boldsymbol{u}}\right\|_2$$

as claimed. □

D.2 Technical Lemmas for Matrix Completion

D.2.1 Orthogonal Procrustes Problem

The orthogonal Procrustes problem is a matrix approximation problem which seeks an orthogonal matrix $\boldsymbol{R}$ to best "align" two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. Specifically, for $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times r}$, define $\widehat{\boldsymbol{R}}$ to be the minimizer of

$$\text{minimize}_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \quad \|\boldsymbol{A}\boldsymbol{R}-\boldsymbol{B}\|_{\mathrm{F}}\,. \tag{250}$$

The first lemma is concerned with the characterization of the minimizer $\widehat{\boldsymbol{R}}$ of (250).

Lemma 35 *For $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times r}$, $\widehat{\boldsymbol{R}}$ is the minimizer of (250) if and only if $\widehat{\boldsymbol{R}}^\top \boldsymbol{A}^\top \boldsymbol{B}$ is symmetric and positive semidefinite.*

Proof This is an immediate consequence of [112, Theorem 2]. □

Let $\boldsymbol{A}^\top\boldsymbol{B}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ be the singular value decomposition of $\boldsymbol{A}^\top\boldsymbol{B}\in\mathbb{R}^{r\times r}$. It is easy to check that $\widehat{\boldsymbol{R}}:=\boldsymbol{U}\boldsymbol{V}^\top$ satisfies the conditions that $\widehat{\boldsymbol{R}}^\top\boldsymbol{A}^\top\boldsymbol{B}$ is both symmetric and positive semidefinite. In view of Lemma 35, $\widehat{\boldsymbol{R}}=\boldsymbol{U}\boldsymbol{V}^\top$ is the minimizer of (250). In the special case when $\boldsymbol{C}:=\boldsymbol{A}^\top\boldsymbol{B}$ is invertible, $\widehat{\boldsymbol{R}}$ enjoys the following equivalent form:

$$\widehat{\boldsymbol{R}}=\widehat{\boldsymbol{H}}(\boldsymbol{C}):=\boldsymbol{C}\left(\boldsymbol{C}^\top\boldsymbol{C}\right)^{-1/2}, \tag{251}$$

where $\widehat{\boldsymbol{H}}(\cdot)$ is an $\mathbb{R}^{r\times r}$-valued function on $\mathbb{R}^{r\times r}$. This motivates us to look at the perturbation bounds for the matrix-valued function $\widehat{\boldsymbol{H}}(\cdot)$, which is formulated in the following lemma.

Lemma 36 *Let $\boldsymbol{C}\in\mathbb{R}^{r\times r}$ be a nonsingular matrix. Then for any matrix $\boldsymbol{E}\in\mathbb{R}^{r\times r}$ with $\|\boldsymbol{E}\|\le\sigma_{\min}(\boldsymbol{C})$ and any unitarily invariant norm $|||\cdot|||$, one has*

$$\left|\left|\left|\widehat{\boldsymbol{H}}(\boldsymbol{C}+\boldsymbol{E})-\widehat{\boldsymbol{H}}(\boldsymbol{C})\right|\right|\right| \le \frac{2}{\sigma_{r-1}(\boldsymbol{C})+\sigma_r(\boldsymbol{C})}|||\boldsymbol{E}|||,$$

where $\widehat{\boldsymbol{H}}(\cdot)$ is defined above.

Proof This is an immediate consequence of [85, Theorem 2.3]. □

With Lemma 36 in place, we are ready to present the following bounds on two matrices after "aligning" them with $\boldsymbol{X}^\star$.

Lemma 37 *Instate the notation in Sect. 3.2. Suppose $\boldsymbol{X}_1, \boldsymbol{X}_2 \in \mathbb{R}^{n\times r}$ are two matrices such that*

$$\left\|\boldsymbol{X}_1 - \boldsymbol{X}^\star\right\| \left\|\boldsymbol{X}^\star\right\| \le \sigma_{\min}/2, \tag{252a}$$
$$\left\|\boldsymbol{X}_1 - \boldsymbol{X}_2\right\| \left\|\boldsymbol{X}^\star\right\| \le \sigma_{\min}/4. \tag{252b}$$

Denote

$$\boldsymbol{R}_1 := \operatorname*{argmin}_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \left\|\boldsymbol{X}_1\boldsymbol{R} - \boldsymbol{X}^\star\right\|_{\mathrm{F}} \quad \text{and} \quad \boldsymbol{R}_2 := \operatorname*{argmin}_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \left\|\boldsymbol{X}_2\boldsymbol{R} - \boldsymbol{X}^\star\right\|_{\mathrm{F}}.$$

Then the following two inequalities hold true:

$$\left\|\boldsymbol{X}_1\boldsymbol{R}_1 - \boldsymbol{X}_2\boldsymbol{R}_2\right\| \le 5\kappa \left\|\boldsymbol{X}_1 - \boldsymbol{X}_2\right\| \quad \text{and} \quad \left\|\boldsymbol{X}_1\boldsymbol{R}_1 - \boldsymbol{X}_2\boldsymbol{R}_2\right\|_{\mathrm{F}} \le 5\kappa \left\|\boldsymbol{X}_1 - \boldsymbol{X}_2\right\|_{\mathrm{F}}.$$

Proof Before proving the claims, we first gather some immediate consequences of assumptions (252). Denote $\boldsymbol{C} = \boldsymbol{X}_1^\top \boldsymbol{X}^\star$ and $\boldsymbol{E} = (\boldsymbol{X}_2 - \boldsymbol{X}_1)^\top \boldsymbol{X}^\star$. It is easily seen that $\boldsymbol{C}$ is invertible since

$$\left\|\boldsymbol{C} - \boldsymbol{X}^{\star\top}\boldsymbol{X}^\star\right\| \le \left\|\boldsymbol{X}_1 - \boldsymbol{X}^\star\right\| \left\|\boldsymbol{X}^\star\right\| \overset{\text{(i)}}{\le} \sigma_{\min}/2 \quad \overset{\text{(ii)}}{\Longrightarrow} \quad \sigma_r(\boldsymbol{C}) \ge \sigma_{\min}/2, \tag{253}$$

where (i) follows from assumption (252a) and (ii) is a direct application of Weyl's inequality. In addition, $\boldsymbol{C} + \boldsymbol{E} = \boldsymbol{X}_2^\top \boldsymbol{X}^\star$ is also invertible since

$$\|\boldsymbol{E}\| \le \left\|\boldsymbol{X}_1 - \boldsymbol{X}_2\right\| \left\|\boldsymbol{X}^\star\right\| \overset{\text{(i)}}{\le} \sigma_{\min}/4 \overset{\text{(ii)}}{<} \sigma_r(\boldsymbol{C}),$$

where (i) arises from assumption (252b) and (ii) holds because of (253). When both $\boldsymbol{C}$ and $\boldsymbol{C} + \boldsymbol{E}$ are invertible, the orthonormal matrices $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ admit closed-form expressions as follows

$$\boldsymbol{R}_1 = \boldsymbol{C}\left(\boldsymbol{C}^\top\boldsymbol{C}\right)^{-1/2} \quad \text{and} \quad \boldsymbol{R}_2 = (\boldsymbol{C}+\boldsymbol{E})\left[(\boldsymbol{C}+\boldsymbol{E})^\top(\boldsymbol{C}+\boldsymbol{E})\right]^{-1/2}.$$

Moreover, we have the following bound on $\|\boldsymbol{X}_1\|$:

$$\|\boldsymbol{X}_1\| \overset{\text{(i)}}{\le} \left\|\boldsymbol{X}_1 - \boldsymbol{X}^\star\right\| + \left\|\boldsymbol{X}^\star\right\| \overset{\text{(ii)}}{\le} \frac{\sigma_{\min}}{2\left\|\boldsymbol{X}^\star\right\|} + \left\|\boldsymbol{X}^\star\right\| \le \frac{\sigma_{\max}}{2\left\|\boldsymbol{X}^\star\right\|} + \left\|\boldsymbol{X}^\star\right\| \overset{\text{(iii)}}{\le} 2\left\|\boldsymbol{X}^\star\right\|, \tag{254}$$

where (i) is the triangle inequality, (ii) uses assumption (252a), and (iii) arises from the fact that $\left\|\boldsymbol{X}^\star\right\| = \sqrt{\sigma_{\max}}$.

With these in place, we turn to establishing the claimed bounds. We will focus on the upper bound on $\|\boldsymbol{X}_1\boldsymbol{R}_1 - \boldsymbol{X}_2\boldsymbol{R}_2\|_{\mathrm{F}}$, as the bound on $\|\boldsymbol{X}_1\boldsymbol{R}_1 - \boldsymbol{X}_2\boldsymbol{R}_2\|$ can be

easily obtained using the same argument. Simple algebra reveals that

$$\begin{aligned}\|\boldsymbol{X}_1\boldsymbol{R}_1-\boldsymbol{X}_2\boldsymbol{R}_2\|_{\mathrm{F}} &= \|(\boldsymbol{X}_1-\boldsymbol{X}_2)\,\boldsymbol{R}_2+\boldsymbol{X}_1\,(\boldsymbol{R}_1-\boldsymbol{R}_2)\|_{\mathrm{F}}\\ &\le \|\boldsymbol{X}_1-\boldsymbol{X}_2\|_{\mathrm{F}}+\|\boldsymbol{X}_1\|\,\|\boldsymbol{R}_1-\boldsymbol{R}_2\|_{\mathrm{F}}\\ &\le \|\boldsymbol{X}_1-\boldsymbol{X}_2\|_{\mathrm{F}}+2\left\|\boldsymbol{X}^\star\right\|\|\boldsymbol{R}_1-\boldsymbol{R}_2\|_{\mathrm{F}}, \end{aligned} \tag{255}$$

where the first inequality uses the fact that $\|\boldsymbol{R}_2\|=1$ and the last inequality comes from (254). An application of Lemma 36 leads us to conclude that

$$\begin{aligned}\|\boldsymbol{R}_1-\boldsymbol{R}_2\|_{\mathrm{F}} &\le \frac{2}{\sigma_r(\boldsymbol{C})+\sigma_{r-1}(\boldsymbol{C})}\|\boldsymbol{E}\|_{\mathrm{F}}\\ &\le \frac{2}{\sigma_{\min}}\left\|(\boldsymbol{X}_2-\boldsymbol{X}_1)^\top\boldsymbol{X}^\star\right\|_{\mathrm{F}} \qquad (256)\\ &\le \frac{2}{\sigma_{\min}}\|\boldsymbol{X}_2-\boldsymbol{X}_1\|_{\mathrm{F}}\left\|\boldsymbol{X}^\star\right\|, \qquad (257)\end{aligned}$$

where (256) utilizes (253). Combine (255) and (257) to reach

$$\begin{aligned}\|\boldsymbol{X}_1\boldsymbol{R}_1-\boldsymbol{X}_2\boldsymbol{R}_2\|_{\mathrm{F}} &\le \|\boldsymbol{X}_1-\boldsymbol{X}_2\|_{\mathrm{F}}+\frac{4}{\sigma_{\min}}\|\boldsymbol{X}_2-\boldsymbol{X}_1\|_{\mathrm{F}}\left\|\boldsymbol{X}^\star\right\|^2\\ &\le (1+4\kappa)\,\|\boldsymbol{X}_1-\boldsymbol{X}_2\|_{\mathrm{F}},\end{aligned}$$

which finishes the proof by noting that $\kappa\ge 1$. □

D.2.2 Matrix Concentration Inequalities

This section collects various measure concentration results regarding the Bernoulli random variables $\{\delta_{j,k}\}_{1\le j,k\le n}$, which is ubiquitous in the analysis for matrix completion.

Lemma 38 *Fix any small constant $\delta>0$, and suppose that $m\gg\delta^{-2}\mu nr\log n$. Then with probability exceeding $1-O\left(n^{-10}\right)$, one has*

$$(1-\delta)\|\boldsymbol{B}\|_{\mathrm{F}}\le\frac{1}{\sqrt{p}}\|\mathcal{P}_\Omega(\boldsymbol{B})\|_{\mathrm{F}}\le(1+\delta)\|\boldsymbol{B}\|_{\mathrm{F}}$$

which holds simultaneously for all $\boldsymbol{B}\in\mathbb{R}^{n\times n}$ lying within the tangent space of $\boldsymbol{M}^\star$.

Proof This result has been established in [19, Section 4.2] for asymmetric sampling patterns (where each $(i,j), i\ne j$, is included in Ω independently). It is straightforward to extend the proof and the result to symmetric sampling patterns (where each (i,j), $i\ge j$, is included in Ω independently). We omit the proof for conciseness. □

Lemma 39 *Fix a matrix $\boldsymbol{M} \in \mathbb{R}^{n\times n}$. Suppose $n^2 p \geq c_0 n \log n$ for some sufficiently large constant $c_0 > 0$. With probability at least $1 - O\left(n^{-10}\right)$, one has*

$$\left\| \frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{M}\right) - \boldsymbol{M} \right\| \leq C\sqrt{\frac{n}{p}}\left\|\boldsymbol{M}\right\|_{\infty},$$

where $C > 0$ is some absolute constant.

Proof See [64, Lemma 3.2]. Similar to Lemma 38, the result therein was provided for the asymmetric sampling patterns but can be easily extended to the symmetric case. □

Lemma 40 *Recall from Sect. 3.2 that $\boldsymbol{E} \in \mathbb{R}^{n\times n}$ is the symmetric noise matrix. Suppose the sample size obeys $n^2 p \geq c_0 n \log^2 n$ for some sufficiently large constant $c_0 > 0$. With probability at least $1 - O\left(n^{-10}\right)$, one has*

$$\left\| \frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{E}\right) \right\| \leq C\sigma\sqrt{\frac{n}{p}},$$

where $C > 0$ is some universal constant.

Proof See [32, Lemma 11]. □

Lemma 41 *Fix some matrix $\boldsymbol{A} \in \mathbb{R}^{n\times r}$ with $n \geq 2r$ and some $1 \leq l \leq n$. Suppose $\left\{\delta_{l,j}\right\}_{1\leq j\leq n}$ are independent Bernoulli random variables with means $\left\{p_j\right\}_{1\leq j\leq n}$ no more than p. Define*

$$\boldsymbol{G}_l\left(\boldsymbol{A}\right) := \left[\delta_{l,1}\boldsymbol{A}_{1,\cdot}^{\top}, \delta_{l,2}\boldsymbol{A}_{2,\cdot}^{\top}, \cdots, \delta_{l,n}\boldsymbol{A}_{n,\cdot}^{\top}\right] \in \mathbb{R}^{r\times n}.$$

Then one has

$$\text{Median}\left[\left\|\boldsymbol{G}_l\left(\boldsymbol{A}\right)\right\|\right] \leq \sqrt{p\left\|\boldsymbol{A}\right\|^2 + \sqrt{2p\left\|\boldsymbol{A}\right\|_{2,\infty}^2\left\|\boldsymbol{A}\right\|^2\log\left(4r\right)} + \frac{2\left\|\boldsymbol{A}\right\|_{2,\infty}^2}{3}\log\left(4r\right)}$$

and for any constant $C \geq 3$, with probability exceeding $1 - n^{-(1.5C-1)}$

$$\left\|\sum_{j=1}^{n}(\delta_{l,j} - p)\boldsymbol{A}_{j,\cdot}^{\top}\boldsymbol{A}_{j,\cdot}\right\| \leq C\left(\sqrt{p\left\|\boldsymbol{A}\right\|_{2,\infty}^2\left\|\boldsymbol{A}\right\|^2\log n} + \left\|\boldsymbol{A}\right\|_{2,\infty}^2\log n\right),$$

and

$$\left\|\boldsymbol{G}_l\left(\boldsymbol{A}\right)\right\| \leq \sqrt{p\left\|\boldsymbol{A}\right\|^2 + C\left(\sqrt{p\left\|\boldsymbol{A}\right\|_{2,\infty}^2\left\|\boldsymbol{A}\right\|^2\log n} + \left\|\boldsymbol{A}\right\|_{2,\infty}^2\log n\right)}.$$

Proof By the definition of $\boldsymbol{G}_l(\boldsymbol{A})$ and the triangle inequality, one has

$$\begin{aligned}\|\boldsymbol{G}_l(\boldsymbol{A})\|^2 &= \left\|\boldsymbol{G}_l(\boldsymbol{A})\,\boldsymbol{G}_l(\boldsymbol{A})^\top\right\| = \left\|\sum_{j=1}^n \delta_{l,j}\boldsymbol{A}_{j,\cdot}^\top \boldsymbol{A}_{j,\cdot}\right\| \\ &\le \left\|\sum_{j=1}^n \left(\delta_{l,j}-p_j\right)\boldsymbol{A}_{j,\cdot}^\top \boldsymbol{A}_{j,\cdot}\right\| + p\,\|\boldsymbol{A}\|^2.\end{aligned}$$

Therefore, it suffices to control the first term. It can be seen that $\left\{\left(\delta_{l,j}-p_j\right)\boldsymbol{A}_{j,\cdot}^\top \boldsymbol{A}_{j,\cdot}\right\}_{1\le j\le n}$ are i.i.d. zero-mean random matrices. Letting

$$\begin{aligned}L &:= \max_{1\le j\le n}\left\|\left(\delta_{l,j}-p_j\right)\boldsymbol{A}_{j,\cdot}^\top \boldsymbol{A}_{j,\cdot}\right\| \le \|\boldsymbol{A}\|_{2,\infty}^2 \\ \text{and}\quad V &:= \left\|\sum_{j=1}^n \mathbb{E}\left[\left(\delta_{l,j}-p_j\right)^2 \boldsymbol{A}_{j,\cdot}^\top \boldsymbol{A}_{j,\cdot}\boldsymbol{A}_{j,\cdot}^\top \boldsymbol{A}_{j,\cdot}\right]\right\| \\ &\le \mathbb{E}\left[\left(\delta_{l,j}-p_j\right)^2\right]\|\boldsymbol{A}\|_{2,\infty}^2\left\|\sum_{j=1}^n \boldsymbol{A}_{j,\cdot}^\top \boldsymbol{A}_{j,\cdot}\right\| \le p\,\|\boldsymbol{A}\|_{2,\infty}^2\,\|\boldsymbol{A}\|^2\end{aligned}$$

and invoking matrix Bernstein's inequality [114, Theorem 6.1.1], one has for all $t \ge 0$,

$$\mathbb{P}\left\{\left\|\sum_{j=1}^n \left(\delta_{l,j}-p_j\right)\boldsymbol{A}_{j,\cdot}^\top \boldsymbol{A}_{j,\cdot}\right\| \ge t\right\} \le 2r\cdot\exp\left(\frac{-t^2/2}{p\,\|\boldsymbol{A}\|_{2,\infty}^2\,\|\boldsymbol{A}\|^2 + \|\boldsymbol{A}\|_{2,\infty}^2\cdot t/3}\right). \tag{258}$$

We can thus find an upper bound on $\mathsf{Median}\left[\left\|\sum_{j=1}^n \left(\delta_{l,j}-p_j\right)\boldsymbol{A}_{j,\cdot}^\top \boldsymbol{A}_{j,\cdot}\right\|\right]$ by finding a value t that ensures the right-hand side of (258) is smaller than $1/2$. Using this strategy and some simple calculations, we get

$$\mathsf{Median}\left[\left\|\sum_{j=1}^n \left(\delta_{l,j}-p_j\right)\boldsymbol{A}_{j,\cdot}^\top \boldsymbol{A}_{j,\cdot}\right\|\right] \le \sqrt{2p\,\|\boldsymbol{A}\|_{2,\infty}^2\,\|\boldsymbol{A}\|^2\log(4r)} + \frac{2\,\|\boldsymbol{A}\|_{2,\infty}^2}{3}\log(4r)$$

and for any $C \ge 3$,

$$\left\|\sum_{j=1}^n \left(\delta_{l,j}-p_j\right)\boldsymbol{A}_{j,\cdot}^\top \boldsymbol{A}_{j,\cdot}\right\| \le C\left(\sqrt{p\,\|\boldsymbol{A}\|_{2,\infty}^2\,\|\boldsymbol{A}\|^2\log n} + \|\boldsymbol{A}\|_{2,\infty}^2\log n\right)$$

holds with probability at least $1 - n^{-(1.5C-1)}$. As a consequence, we have

$$\text{Median}\left[\|\boldsymbol{G}_l(\boldsymbol{A})\|\right] \le \sqrt{p\|\boldsymbol{A}\|^2 + \sqrt{2p\|\boldsymbol{A}\|_{2,\infty}^2\|\boldsymbol{A}\|^2\log(4r)} + \frac{2\|\boldsymbol{A}\|_{2,\infty}^2}{3}\log(4r)},$$

and with probability exceeding $1 - n^{-(1.5C-1)}$,

$$\|\boldsymbol{G}_l(\boldsymbol{A})\|^2 \le p\|\boldsymbol{A}\|^2 + C\left(\sqrt{p\|\boldsymbol{A}\|_{2,\infty}^2\|\boldsymbol{A}\|^2\log n} + \|\boldsymbol{A}\|_{2,\infty}^2\log n\right).$$

This completes the proof. □

Lemma 42 *Let $\{\delta_{l,j}\}_{1\le l\le j\le n}$ be i.i.d. Bernoulli random variables with mean p and $\delta_{l,j} = \delta_{j,l}$. For any $\boldsymbol{\Delta} \in \mathbb{R}^{n\times r}$, define*

$$\boldsymbol{G}_l(\boldsymbol{\Delta}) := \left[\delta_{l,1}\boldsymbol{\Delta}_{1,\cdot}^\top, \delta_{l,2}\boldsymbol{\Delta}_{2,\cdot}^\top, \cdots, \delta_{l,n}\boldsymbol{\Delta}_{n,\cdot}^\top\right] \in \mathbb{R}^{r\times n}.$$

Suppose the sample size obeys $n^2p \gg \kappa\mu rn\log^2 n$. Then for any $k > 0$ and $\alpha > 0$ large enough, with probability at least $1 - c_1e^{-\alpha Cnr\log n/2}$,

$$\sum_{l=1}^{n}\mathbb{1}_{\left\{\|\boldsymbol{G}_l(\boldsymbol{\Delta})\|\ge 4\sqrt{p}\psi + 2\sqrt{kr}\xi\right\}} \le \frac{2\alpha n\log n}{k}$$

holds simultaneously for all $\boldsymbol{\Delta} \in \mathbb{R}^{n\times r}$ obeying

$$\|\boldsymbol{\Delta}\|_{2,\infty} \le C_5\rho^t\mu r\sqrt{\frac{\log n}{np}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty} + C_8\sigma\sqrt{\frac{n\log n}{p}}\left\|\boldsymbol{X}^\star\right\|_{2,\infty} := \xi$$

$$\text{and}\qquad \|\boldsymbol{\Delta}\| \le C_9\rho^t\mu r\frac{1}{\sqrt{np}}\left\|\boldsymbol{X}^\star\right\| + C_{10}\sigma\sqrt{\frac{n}{p}}\left\|\boldsymbol{X}^\star\right\| := \psi,$$

where $c_1, C_5, C_8, C_9, C_{10} > 0$ are some absolute constants.

Proof For simplicity of presentation, we will prove the claim for the asymmetric case where $\{\delta_{l,j}\}_{1\le l,j\le n}$ are independent. The results immediately carry over to the symmetric case as claimed in this lemma. To see this, note that we can always divide $\boldsymbol{G}_l(\boldsymbol{\Delta})$ into

$$\boldsymbol{G}_l(\boldsymbol{\Delta}) = \boldsymbol{G}_l^{\text{upper}}(\boldsymbol{\Delta}) + \boldsymbol{G}_l^{\text{lower}}(\boldsymbol{\Delta}),$$

where all nonzero components of $\boldsymbol{G}_l^{\text{upper}}(\boldsymbol{\Delta})$ come from the upper triangular part (those blocks with $l \le j$), while all nonzero components of $\boldsymbol{G}_l^{\text{lower}}(\boldsymbol{\Delta})$ are from the lower triangular part (those blocks with $l > j$). We can then look at $\left\{\boldsymbol{G}_l^{\text{upper}}(\boldsymbol{\Delta}) \mid 1 \le l \le n\right\}$ and $\left\{\boldsymbol{G}_l^{\text{upper}}(\boldsymbol{\Delta}) \mid 1 \le l \le n\right\}$ separately using the argument we develop for the asymmetric case. From now on, we assume that $\{\delta_{l,j}\}_{1\le l,j\le n}$ are independent.

Suppose for the moment that $\boldsymbol{\Delta}$ is statistically independent of $\{\delta_{l,j}\}$. Clearly, for any $\boldsymbol{\Delta}, \widetilde{\boldsymbol{\Delta}} \in \mathbb{R}^{n\times r}$,

$$\begin{aligned}\left|\left\|\boldsymbol{G}_l(\boldsymbol{\Delta})\right\| - \left\|\boldsymbol{G}_l(\widetilde{\boldsymbol{\Delta}})\right\|\right| &\le \left\|\boldsymbol{G}_l(\boldsymbol{\Delta}) - \boldsymbol{G}_l(\widetilde{\boldsymbol{\Delta}})\right\| \le \left\|\boldsymbol{G}_l(\boldsymbol{\Delta}) - \boldsymbol{G}_l(\widetilde{\boldsymbol{\Delta}})\right\|_{\mathrm{F}} \\ &\le \sqrt{\sum_{j=1}^{n}\left\|\boldsymbol{\Delta}_{j,\cdot} - \widetilde{\boldsymbol{\Delta}}_{j,\cdot}\right\|_2^2} \\ &:= d\big(\boldsymbol{\Delta}, \widetilde{\boldsymbol{\Delta}}\big),\end{aligned}$$

which implies that $\|\boldsymbol{G}_l(\boldsymbol{\Delta})\|$ is 1-Lipschitz with respect to the metric $d(\cdot,\cdot)$. Moreover,

$$\max_{1\le j\le n}\left\|\delta_{l,j}\boldsymbol{\Delta}_{j,\cdot}\right\|_2 \le \|\boldsymbol{\Delta}\|_{2,\infty} \le \xi$$

according to our assumption. Hence, Talagrand's inequality [24, Proposition 1] reveals the existence of some absolute constants $C, c > 0$ such that for all $\lambda > 0$

$$\mathbb{P}\left\{\|\boldsymbol{G}_l(\boldsymbol{\Delta})\| - \mathsf{Median}\left[\|\boldsymbol{G}_l(\boldsymbol{\Delta})\|\right] \ge \lambda\xi\right\} \le C\exp\left(-c\lambda^2\right). \tag{259}$$

We then proceed to control $\mathsf{Median}\left[\|\boldsymbol{G}_l(\boldsymbol{\Delta})\|\right]$. A direct application of Lemma 41 yields

$$\mathsf{Median}\left[\|\boldsymbol{G}_l(\boldsymbol{\Delta})\|\right] \le \sqrt{2p\psi^2 + \sqrt{p\log(4r)}\xi\psi + \frac{2\xi^2}{3}\log(4r)} \le 2\sqrt{p}\psi,$$

where the last relation holds since $p\psi^2 \gg \xi^2\log r$, which follows by combining the definitions of ψ and ξ, the sample size condition $np \gg \kappa\mu r\log^2 n$, and the incoherence condition (114). Thus, substitution into (259) and taking $\lambda = \sqrt{kr}$ give

$$\mathbb{P}\left\{\|\boldsymbol{G}_l(\boldsymbol{\Delta})\| \ge 2\sqrt{p}\psi + \sqrt{kr}\xi\right\} \le C\exp(-ckr) \tag{260}$$

for any $k \ge 0$. Furthermore, invoking [4, Corollary A.1.14] and using bound (260), one has

$$\mathbb{P}\left(\sum_{l=1}^{n}\mathbb{1}_{\left\{\|\boldsymbol{G}_l(\boldsymbol{\Delta})\|\ge 2\sqrt{p}\psi+\sqrt{kr}\xi\right\}} \ge tnC\exp(-ckr)\right) \le 2\exp\left(-\frac{t\log t}{2}nC\exp(-ckr)\right)$$

for any $t \ge 6$. Choose $t = \alpha\log n/\left[kC\exp(-ckr)\right] \ge 6$ to obtain

$$\mathbb{P}\left(\sum_{l=1}^{n}\mathbb{1}_{\left\{\|\boldsymbol{G}_l(\boldsymbol{\Delta})\|\ge 2\sqrt{p}\psi+\sqrt{kr}\xi\right\}} \ge \frac{\alpha n\log n}{k}\right) \le 2\exp\left(-\frac{\alpha C}{2}nr\log n\right). \tag{261}$$

So far, we have demonstrated that for any fixed $\boldsymbol{\Delta}$ obeying our assumptions, $\sum_{l=1}^{n} \mathbb{1}_{\left\{\|\boldsymbol{G}_l(\boldsymbol{\Delta})\| \geq 2\sqrt{p}\psi + \sqrt{kr}\xi\right\}}$ is well controlled with exponentially high probability. In order to extend the results to all feasible $\boldsymbol{\Delta}$, we resort to the standard ϵ-net argument. Clearly, due to the homogeneity property of $\|\boldsymbol{G}_l(\boldsymbol{\Delta})\|$, it suffices to restrict attention to the following set:

$$\mathcal{S} = \{\boldsymbol{\Delta} \mid \min\{\xi, \psi\} \leq \|\boldsymbol{\Delta}\| \leq \psi\}, \tag{262}$$

where $\psi/\xi \lesssim \|\boldsymbol{X}^\star\|/\|\boldsymbol{X}^\star\|_{2,\infty} \lesssim \sqrt{n}$. We then proceed with the following steps.

1. Introduce the auxiliary function

$$\chi_l(\boldsymbol{\Delta}) = \begin{cases} 1, & \text{if } \|\boldsymbol{G}_l(\boldsymbol{\Delta})\| \geq 4\sqrt{p}\psi + 2\sqrt{kr}\xi, \\ \frac{\|\boldsymbol{G}_l(\boldsymbol{\Delta})\| - 2\sqrt{p}\psi - \sqrt{kr}\xi}{2\sqrt{p}\psi + \sqrt{kr}\xi}, & \text{if } \|\boldsymbol{G}_l(\boldsymbol{\Delta})\| \in [2\sqrt{p}\psi + \sqrt{kr}\xi,\ 4\sqrt{p}\psi + 2\sqrt{kr}\xi], \\ 0, & \text{else.} \end{cases}$$

   Clearly, this function is sandwiched between two indicator functions

$$\mathbb{1}_{\left\{\|\boldsymbol{G}_l(\boldsymbol{\Delta})\| \geq 4\sqrt{p}\psi + 2\sqrt{kr}\xi\right\}} \leq \chi_l(\boldsymbol{\Delta}) \leq \mathbb{1}_{\left\{\|\boldsymbol{G}_l(\boldsymbol{\Delta})\| \geq 2\sqrt{p}\psi + \sqrt{kr}\xi\right\}}.$$

   Note that χ_l is more convenient to work with due to continuity.
2. Consider an ϵ-net $\mathcal{N}_\epsilon$ [111, Section 2.3.1] of the set $\mathcal{S}$ as defined in (262). For any $\epsilon = 1/n^{O(1)}$, one can find such a net with cardinality $\log|\mathcal{N}_\epsilon| \lesssim nr\log n$. Apply the union bound and (261) to yield

$$\begin{aligned} &\mathbb{P}\left(\sum_{l=1}^{n} \chi_l(\boldsymbol{\Delta}) \geq \frac{\alpha n\log n}{k},\ \forall \boldsymbol{\Delta} \in \mathcal{N}_\epsilon\right) \\ &\leq \mathbb{P}\left(\sum_{l=1}^{n} \mathbb{1}_{\left\{\|\boldsymbol{G}_l(\boldsymbol{\Delta})\| \geq 2\sqrt{p}\psi + \sqrt{kr}\xi\right\}} \geq \frac{\alpha n\log n}{k},\ \forall \boldsymbol{\Delta} \in \mathcal{N}_\epsilon\right) \\ &\leq 2|\mathcal{N}_\epsilon| \exp\left(-\frac{\alpha C}{2} nr\log n\right) \leq 2\exp\left(-\frac{\alpha C}{4} nr\log n\right), \end{aligned}$$

   as long as α is chosen to be sufficiently large.
3. One can then use the continuity argument to extend the bound to all $\boldsymbol{\Delta}$ outside the ϵ-net, i.e., with exponentially high probability,

$$\begin{aligned} &\sum_{l=1}^{n} \chi_l(\boldsymbol{\Delta}) \leq \frac{2\alpha n\log n}{k}, \quad \forall \boldsymbol{\Delta} \in \mathcal{S} \\ \Longrightarrow \quad &\sum_{l=1}^{n} \mathbb{1}_{\left\{\|\boldsymbol{G}_l(\boldsymbol{\Delta})\| \geq 4\sqrt{p}\psi + 2\sqrt{kr}\xi\right\}} \leq \sum_{l=1}^{n} \chi_l(\boldsymbol{\Delta}) \leq \frac{2\alpha n\log n}{k}, \quad \forall \boldsymbol{\Delta} \in \mathcal{S}. \end{aligned}$$

   This is fairly standard (see, e.g., [111, Section 2.3.1]) and is thus omitted here.

We have thus concluded the proof. □

Lemma 43 *Suppose the sample size obeys $n^2 p \geq C\kappa\mu r n \log n$ for some sufficiently large constant $C > 0$. Then with probability at least $1 - O\left(n^{-10}\right)$,*

$$\left\| \frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{X}^\star\boldsymbol{X}^{\star\top}\right)\right\| \leq 2n\epsilon^2 \left\|\boldsymbol{X}^\star\right\|_{2,\infty}^2 + 4\epsilon\sqrt{n}\log n \left\|\boldsymbol{X}^\star\right\|_{2,\infty} \left\|\boldsymbol{X}^\star\right\|$$

holds simultaneously for all $\boldsymbol{X} \in \mathbb{R}^{n\times r}$ satisfying

$$\left\|\boldsymbol{X} - \boldsymbol{X}^\star\right\|_{2,\infty} \leq \epsilon \left\|\boldsymbol{X}^\star\right\|_{2,\infty}, \tag{263}$$

where $\epsilon > 0$ is any fixed constant.

Proof To simplify the notations hereafter, we denote $\boldsymbol{\Delta} := \boldsymbol{X} - \boldsymbol{X}^\star$. With this notation in place, one can decompose

$$\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{X}^\star\boldsymbol{X}^{\star\top} = \boldsymbol{\Delta}\boldsymbol{X}^{\star\top} + \boldsymbol{X}^\star\boldsymbol{\Delta}^\top + \boldsymbol{\Delta}\boldsymbol{\Delta}^\top,$$

which together with the triangle inequality implies that

$$\begin{aligned}\left\| \frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{X}^\star\boldsymbol{X}^{\star\top}\right)\right\| &\leq \left\| \frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{\Delta}\boldsymbol{X}^{\star\top}\right)\right\| + \left\| \frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{X}^\star\boldsymbol{\Delta}^\top\right)\right\| + \left\| \frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{\Delta}\boldsymbol{\Delta}^\top\right)\right\| \\ &= \underbrace{\left\| \frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{\Delta}\boldsymbol{\Delta}^\top\right)\right\|}_{:=\alpha_1} + 2\underbrace{\left\| \frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{\Delta}\boldsymbol{X}^{\star\top}\right)\right\|}_{:=\alpha_2}. \end{aligned} \tag{264}$$

In the sequel, we bound α_1 and α_2 separately.

1. Recall from [84, Theorem 2.5] the elementary inequality that

$$\|\boldsymbol{C}\| \leq \big\||\boldsymbol{C}|\big\|, \tag{265}$$

where $|\boldsymbol{C}| := [|c_{i,j}|]_{1\leq i,j\leq n}$ for any matrix $\boldsymbol{C} = [c_{i,j}]_{1\leq i,j\leq n}$. In addition, for any matrix $\boldsymbol{D} := [d_{i,j}]_{1\leq i,j\leq n}$ such that $|d_{i,j}| \geq |c_{i,j}|$ for all i and j, one has $\big\||\boldsymbol{C}|\big\| \leq \big\||\boldsymbol{D}|\big\|$. Therefore,

$$\alpha_1 \leq \left\| \frac{1}{p}\mathcal{P}_\Omega\left(\left|\boldsymbol{\Delta}\boldsymbol{\Delta}^\top\right|\right)\right\| \leq \|\boldsymbol{\Delta}\|_{2,\infty}^2 \left\| \frac{1}{p}\mathcal{P}_\Omega\left(\mathbf{1}\mathbf{1}^\top\right)\right\|.$$

Lemma 39 then tells us that with probability at least $1 - O(n^{-10})$,

$$\left\| \frac{1}{p}\mathcal{P}_\Omega\left(\mathbf{1}\mathbf{1}^\top\right) - \mathbf{1}\mathbf{1}^\top\right\| \leq C\sqrt{\frac{n}{p}} \tag{266}$$

for some universal constant $C > 0$, as long as $p \gg \log n/n$. This together with the triangle inequality yields

$$\left\| \frac{1}{p}\mathcal{P}_\Omega\left(\mathbf{1}\mathbf{1}^\top\right)\right\| \leq \left\| \frac{1}{p}\mathcal{P}_\Omega\left(\mathbf{1}\mathbf{1}^\top\right) - \mathbf{1}\mathbf{1}^\top\right\| + \left\|\mathbf{1}\mathbf{1}^\top\right\| \leq C\sqrt{\frac{n}{p}} + n \leq 2n, \tag{267}$$

provided that $p \gg 1/n$. Putting together the previous bounds, we arrive at

$$\alpha_1 \le 2n \left\|\boldsymbol{\Delta}\right\|_{2,\infty}^2. \tag{268}$$

2. Regarding the second term α_2, apply the elementary inequality (265) once again to get

$$\left\|\mathcal{P}_\Omega\left(\boldsymbol{\Delta}\boldsymbol{X}^{\star\top}\right)\right\| \le \left\|\mathcal{P}_\Omega\left(\left|\boldsymbol{\Delta}\boldsymbol{X}^{\star\top}\right|\right)\right\|,$$

which motivates us to look at $\left\|\mathcal{P}_\Omega\left(\left|\boldsymbol{\Delta}\boldsymbol{X}^{\star\top}\right|\right)\right\|$ instead. A key step of this part is to take advantage of the $\ell_{2,\infty}$ norm constraint of $\mathcal{P}_\Omega\left(\left|\boldsymbol{\Delta}\boldsymbol{X}^{\star\top}\right|\right)$. Specifically, we claim for the moment that with probability exceeding $1 - O(n^{-10})$,

$$\left\|\mathcal{P}_\Omega\left(\left|\boldsymbol{\Delta}\boldsymbol{X}^{\star\top}\right|\right)\right\|_{2,\infty}^2 \le 2p\sigma_{\max}\left\|\boldsymbol{\Delta}\right\|_{2,\infty}^2 := \theta \tag{269}$$

holds under our sample size condition. In addition, we also have the following trivial ℓ_∞ norm bound

$$\left\|\mathcal{P}_\Omega\left(\left|\boldsymbol{\Delta}\boldsymbol{X}^{\star\top}\right|\right)\right\|_{\infty} \le \left\|\boldsymbol{\Delta}\right\|_{2,\infty}\left\|\boldsymbol{X}^\star\right\|_{2,\infty} := \gamma. \tag{270}$$

In what follows, for simplicity of presentation, we will denote

$$\boldsymbol{A} := \mathcal{P}_\Omega\left(\left|\boldsymbol{\Delta}\boldsymbol{X}^{\star\top}\right|\right). \tag{271}$$

(a) To facilitate the analysis of $\left\|\boldsymbol{A}\right\|$, we first introduce $k_0 + 1 = \frac{1}{2}\log(\kappa\mu r)$ auxiliary matrices[9] $\boldsymbol{B}_s \in \mathbb{R}^{n\times n}$ that satisfy

$$\left\|\boldsymbol{A}\right\| \le \left\|\boldsymbol{B}_{k_0}\right\| + \sum_{s=0}^{k_0-1}\left\|\boldsymbol{B}_s\right\|. \tag{272}$$

To be precise, each $\boldsymbol{B}_s$ is defined such that

$$[\boldsymbol{B}_s]_{j,k} = \begin{cases} \frac{1}{2^s}\gamma, & \text{if } A_{j,k} \in (\frac{1}{2^{s+1}}\gamma, \frac{1}{2^s}\gamma], \\ 0, & \text{else,} \end{cases} \quad \text{for } 0 \le s \le k_0 - 1 \quad \text{and}$$

$$\left[\boldsymbol{B}_{k_0}\right]_{j,k} = \begin{cases} \frac{1}{2^{k_0}}\gamma, & \text{if } A_{j,k} \le \frac{1}{2^{k_0}}\gamma, \\ 0, & \text{else,} \end{cases}$$

which clearly satisfy (272); in words, $\boldsymbol{B}_s$ is constructed by rounding up those entries of $\boldsymbol{A}$ within a prescribed magnitude interval. Thus, it suffices to bound

[9] For simplicity, we assume $\frac{1}{2}\log(\kappa\mu r)$ is an integer. The argument here can be easily adapted to the case when $\frac{1}{2}\log(\kappa\mu r)$ is not an integer.

$\|\boldsymbol{B}_s\|$ for every s. To this end, we start with $s = k_0$ and use the definition of $\boldsymbol{B}_{k_0}$ to get

$$\left\|\boldsymbol{B}_{k_0}\right\| \overset{\text{(i)}}{\le} \left\|\boldsymbol{B}_{k_0}\right\|_\infty \sqrt{(2np)^2} \overset{\text{(ii)}}{\le} 4np\frac{1}{\sqrt{\kappa\mu r}}\|\boldsymbol{\Delta}\|_{2,\infty}\left\|\boldsymbol{X}^\star\right\|_{2,\infty}$$
$$\overset{\text{(iii)}}{\le} 4\sqrt{n}p\|\boldsymbol{\Delta}\|_{2,\infty}\left\|\boldsymbol{X}^\star\right\|,$$

where (i) arises from Lemma 44, with $2np$ being a crude upper bound on the number of nonzero entries in each row and each column. This can be derived by applying the standard Chernoff bound on Ω. The second inequality (ii) relies on the definitions of γ and k_0. The last one (iii) follows from the incoherence condition (114). Besides, for any $0 \le s \le k_0 - 1$, by construction one has

$$\|\boldsymbol{B}_s\|_{2,\infty}^2 \le 4\theta = 8p\sigma_{\max}\|\boldsymbol{\Delta}\|_{2,\infty}^2 \qquad \text{and} \qquad \|\boldsymbol{B}_s\|_\infty = \frac{1}{2^s}\gamma,$$

where θ is as defined in (269). Here, we have used the fact that the magnitude of each entry of $\boldsymbol{B}_s$ is at most two times that of $\boldsymbol{A}$. An immediate implication is that there are at most

$$\frac{\|\boldsymbol{B}_s\|_{2,\infty}^2}{\|\boldsymbol{B}_s\|_\infty^2} \le \frac{8p\sigma_{\max}\|\boldsymbol{\Delta}\|_{2,\infty}^2}{\left(\frac{1}{2^s}\gamma\right)^2} := k_{\mathrm{r}}$$

nonzero entries in each row of $\boldsymbol{B}_s$ and at most

$$k_{\mathrm{c}} = 2np$$

nonzero entries in each column of $\boldsymbol{B}_s$, where k_{c} is derived from the standard Chernoff bound on Ω. Utilizing Lemma 44 once more, we discover that

$$\|\boldsymbol{B}_s\| \le \|\boldsymbol{B}_s\|_\infty\sqrt{k_{\mathrm{r}}k_{\mathrm{c}}} = \frac{1}{2^s}\gamma\sqrt{k_{\mathrm{r}}k_{\mathrm{c}}} = \sqrt{16np^2\sigma_{\max}\|\boldsymbol{\Delta}\|_{2,\infty}^2}$$
$$= 4\sqrt{n}p\|\boldsymbol{\Delta}\|_{2,\infty}\left\|\boldsymbol{X}^\star\right\|$$

for each $0 \le s \le k_0 - 1$. Combining all, we arrive at

$$\|\boldsymbol{A}\| \le \sum_{s=0}^{k_0-1}\|\boldsymbol{B}_s\| + \left\|\boldsymbol{B}_{k_0}\right\| \le (k_0+1)\,4\sqrt{n}p\|\boldsymbol{\Delta}\|_{2,\infty}\left\|\boldsymbol{X}^\star\right\|$$
$$\le 2\sqrt{n}p\log(\kappa\mu r)\|\boldsymbol{\Delta}\|_{2,\infty}\left\|\boldsymbol{X}^\star\right\|$$
$$\le 2\sqrt{n}p\log n\|\boldsymbol{\Delta}\|_{2,\infty}\left\|\boldsymbol{X}^\star\right\|,$$

where the last relation holds under the condition $n \ge \kappa\mu r$. This further gives

$$\alpha_2 \le \frac{1}{p}\|\boldsymbol{A}\| \le 2\sqrt{n}\log n\|\boldsymbol{\Delta}\|_{2,\infty}\left\|\boldsymbol{X}^\star\right\|. \tag{273}$$

(b) In order to finish the proof of this part, we need to justify claim (269). Observe that

$$\begin{aligned}\left\|\left[\mathcal{P}_\Omega\left(\left|\boldsymbol{\Delta}\boldsymbol{X}^{\star\top}\right|\right)\right]_{l,\cdot}\right\|_2^2 &= \sum\nolimits_{j=1}^{n}\left(\boldsymbol{\Delta}_{l,\cdot}\boldsymbol{X}_{j,\cdot}^{\star\top}\delta_{l,j}\right)^2\\ &= \boldsymbol{\Delta}_{l,\cdot}\left(\sum\nolimits_{j=1}^{n}\delta_{l,j}\boldsymbol{X}_{j,\cdot}^{\star\top}\boldsymbol{X}_{j,\cdot}^{\star}\right)\boldsymbol{\Delta}_{l,\cdot}^{\top}\\ &\le \|\boldsymbol{\Delta}\|_{2,\infty}^2\left\|\sum\nolimits_{j=1}^{n}\delta_{l,j}\boldsymbol{X}_{j,\cdot}^{\star\top}\boldsymbol{X}_{j,\cdot}^{\star}\right\| \end{aligned} \tag{274}$$

for every $1 \le l \le n$, where $\delta_{l,j}$ indicates whether the entry with the index (l, j) is observed or not. Invoke Lemma 41 to yield

$$\begin{aligned}\left\|\sum\nolimits_{j=1}^{n}\delta_{l,j}\boldsymbol{X}_{j,\cdot}^{\star\top}\boldsymbol{X}_{j,\cdot}^{\star}\right\| &= \left\|\left[\delta_{l,1}\boldsymbol{X}_{1,\cdot}^{\star\top},\delta_{l,2}\boldsymbol{X}_{2,\cdot}^{\star\top},\cdots,\delta_{l,n}\boldsymbol{X}_{n,\cdot}^{\star\top}\right]\right\|^2\\ &\le p\sigma_{\max}+C\left(\sqrt{p\left\|\boldsymbol{X}^{\star}\right\|_{2,\infty}^2\left\|\boldsymbol{X}^{\star}\right\|^2\log n}+\left\|\boldsymbol{X}^{\star}\right\|_{2,\infty}^2\log n\right)\\ &\le\left(p+C\sqrt{\frac{p\kappa\mu r\log n}{n}}+C\frac{\kappa\mu r\log n}{n}\right)\sigma_{\max}\\ &\le 2p\sigma_{\max}, \end{aligned} \tag{275}$$

with high probability, as soon as $np \gg \kappa\mu r\log n$. Combining (274) and (275) yields

$$\left\|\left[\mathcal{P}_\Omega\left(\left|\boldsymbol{\Delta}\boldsymbol{X}^{\star\top}\right|\right)\right]_{l,\cdot}\right\|_2^2 \le 2p\sigma_{\max}\|\boldsymbol{\Delta}\|_{2,\infty}^2, \qquad 1\le l\le n,$$

as claimed in (269).

3. Taken together, the preceding bounds (264), (268), and (273) yield

$$\left\|\frac{1}{p}\mathcal{P}_\Omega\left(\boldsymbol{X}\boldsymbol{X}^{\top}-\boldsymbol{X}^{\star}\boldsymbol{X}^{\star\top}\right)\right\| \le \alpha_1+2\alpha_2 \le 2n\|\boldsymbol{\Delta}\|_{2,\infty}^2+4\sqrt{n}\log n\|\boldsymbol{\Delta}\|_{2,\infty}\left\|\boldsymbol{X}^{\star}\right\|.$$

The proof is completed by substituting the assumption $\|\boldsymbol{\Delta}\|_{2,\infty} \le \epsilon\left\|\boldsymbol{X}^{\star}\right\|_{2,\infty}$. □

In the end of this subsection, we record a useful lemma to bound the spectral norm of a sparse Bernoulli matrix.

Lemma 44 *Let $\boldsymbol{A} \in \{0,1\}^{n_1\times n_2}$ be a binary matrix, and suppose that there are at most k_{r} and k_{c} nonzero entries in each row and column of $\boldsymbol{A}$, respectively. Then one has $\|\boldsymbol{A}\| \le \sqrt{k_{\mathrm{c}}k_{\mathrm{r}}}$.*

Proof This immediately follows from the elementary inequality $\|\boldsymbol{A}\|^2 \le \|\boldsymbol{A}\|_{1\to 1}\|\boldsymbol{A}\|_{\infty\to\infty}$ (see [56, equation (1.11)]), where $\|\boldsymbol{A}\|_{1\to 1}$ and $\|\boldsymbol{A}\|_{\infty\to\infty}$ are the induced 1-norm (or maximum absolute column sum norm) and the induced ∞-norm (or maximum absolute row sum norm), respectively. □

D.2.3 Matrix Perturbation Bounds

Lemma 45 *Let $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ be a symmetric matrix with the top-r eigendecomposition $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^\top$. Assume $\|\boldsymbol{M}-\boldsymbol{M}^\star\| \le \sigma_{\min}/2$, and denote*

$$\widehat{\boldsymbol{Q}} := \operatorname*{argmin}_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \|\boldsymbol{U}\boldsymbol{R}-\boldsymbol{U}^\star\|_{\mathrm{F}}.$$

Then there is some numerical constant $c_3 > 0$ such that

$$\|\boldsymbol{U}\widehat{\boldsymbol{Q}}-\boldsymbol{U}^\star\| \le \frac{c_3}{\sigma_{\min}}\|\boldsymbol{M}-\boldsymbol{M}^\star\|.$$

Proof Define $\boldsymbol{Q} = \boldsymbol{U}^\top\boldsymbol{U}^\star$. The triangle inequality gives

$$\|\boldsymbol{U}\widehat{\boldsymbol{Q}}-\boldsymbol{U}^\star\| \le \|\boldsymbol{U}(\widehat{\boldsymbol{Q}}-\boldsymbol{Q})\| + \|\boldsymbol{U}\boldsymbol{Q}-\boldsymbol{U}^\star\| \le \|\widehat{\boldsymbol{Q}}-\boldsymbol{Q}\| + \left\|\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{U}^\star-\boldsymbol{U}^\star\right\|. \tag{276}$$

[1, Lemma 3] asserts that

$$\|\widehat{\boldsymbol{Q}}-\boldsymbol{Q}\| \le 4\left(\|\boldsymbol{M}-\boldsymbol{M}^\star\|/\sigma_{\min}\right)^2$$

as long as $\|\boldsymbol{M}-\boldsymbol{M}^\star\| \le \sigma_{\min}/2$. For the remaining term in (276), one can use $\boldsymbol{U}^{\star\top}\boldsymbol{U}^\star = \boldsymbol{I}_r$ to obtain

$$\left\|\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{U}^\star-\boldsymbol{U}^\star\right\| = \left\|\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{U}^\star-\boldsymbol{U}^\star\boldsymbol{U}^{\star\top}\boldsymbol{U}^\star\right\| \le \left\|\boldsymbol{U}\boldsymbol{U}^\top-\boldsymbol{U}^\star\boldsymbol{U}^{\star\top}\right\|,$$

which together with the Davis–Kahan sinΘ theorem [39] reveals that

$$\left\|\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{U}^\star-\boldsymbol{U}^\star\right\| \le \frac{c_2}{\sigma_{\min}}\|\boldsymbol{M}-\boldsymbol{M}^\star\|$$

for some constant $c_2 > 0$. Combine the estimates on $\|\widehat{\boldsymbol{Q}}-\boldsymbol{Q}\|$, $\|\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{U}^\star-\boldsymbol{U}^\star\|$ and (276) to reach

$$\|\boldsymbol{U}\widehat{\boldsymbol{Q}}-\boldsymbol{U}^\star\| \le \left(\frac{4}{\sigma_{\min}}\|\boldsymbol{M}-\boldsymbol{M}^\star\|\right)^2 + \frac{c_2}{\sigma_{\min}}\|\boldsymbol{M}-\boldsymbol{M}^\star\| \le \frac{c_3}{\sigma_{\min}}\|\boldsymbol{M}-\boldsymbol{M}^\star\|$$

for some numerical constant $c_3 > 0$, where we have utilized the fact that $\|\boldsymbol{M}-\boldsymbol{M}^\star\|/\sigma_{\min} \le 1/2$. □

Lemma 46 *Let $\boldsymbol{M}, \widetilde{\boldsymbol{M}} \in \mathbb{R}^{n\times n}$ be two symmetric matrices with top-r eigendecompositions $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^\top$ and $\widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{U}}^\top$, respectively. Assume $\|\boldsymbol{M}-\boldsymbol{M}^\star\| \le \sigma_{\min}/4$ and $\|\widetilde{\boldsymbol{M}}-\boldsymbol{M}^\star\| \le \sigma_{\min}/4$, and suppose $\sigma_{\max}/\sigma_{\min}$ is bounded by some constant $c_1 > 0$,*

with $\sigma_{\max}$ and $\sigma_{\min}$ the largest and the smallest singular values of $\boldsymbol{M}^\star$, respectively. If we denote

$$\boldsymbol{Q} := \operatorname*{argmin}_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \left\|\boldsymbol{U}\boldsymbol{R} - \widetilde{\boldsymbol{U}}\right\|_{\mathrm{F}},$$

then there exists some numerical constant $c_3 > 0$ such that

$$\left\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{Q} - \boldsymbol{Q}\widetilde{\boldsymbol{\Sigma}}^{1/2}\right\| \le \frac{c_3}{\sqrt{\sigma_{\min}}}\left\|\widetilde{\boldsymbol{M}} - \boldsymbol{M}\right\| \quad \text{and}$$
$$\left\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{Q} - \boldsymbol{Q}\widetilde{\boldsymbol{\Sigma}}^{1/2}\right\|_{\mathrm{F}} \le \frac{c_3}{\sqrt{\sigma_{\min}}}\left\|(\widetilde{\boldsymbol{M}} - \boldsymbol{M})\boldsymbol{U}\right\|_{\mathrm{F}}.$$

Proof Here, we focus on the Frobenius norm; the bound on the operator norm follows from the same argument, and hence we omit the proof. Since $\|\cdot\|_{\mathrm{F}}$ is unitarily invariant, we have

$$\left\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{Q} - \boldsymbol{Q}\widetilde{\boldsymbol{\Sigma}}^{1/2}\right\|_{\mathrm{F}} = \left\|\boldsymbol{Q}^\top\boldsymbol{\Sigma}^{1/2}\boldsymbol{Q} - \widetilde{\boldsymbol{\Sigma}}^{1/2}\right\|_{\mathrm{F}},$$

where $\boldsymbol{Q}^\top\boldsymbol{\Sigma}^{1/2}\boldsymbol{Q}$ and $\widetilde{\boldsymbol{\Sigma}}^{1/2}$ are the matrix square roots of $\boldsymbol{Q}^\top\boldsymbol{\Sigma}\boldsymbol{Q}$ and $\widetilde{\boldsymbol{\Sigma}}$, respectively. In view of the matrix square root perturbation bound [97, Lemma 2.1],

$$\begin{aligned}\left\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{Q} - \boldsymbol{Q}\widetilde{\boldsymbol{\Sigma}}^{1/2}\right\|_{\mathrm{F}} &\le \frac{1}{\sigma_{\min}\left[(\boldsymbol{\Sigma})^{1/2}\right] + \sigma_{\min}\left[(\widetilde{\boldsymbol{\Sigma}})^{1/2}\right]}\left\|\boldsymbol{Q}^\top\boldsymbol{\Sigma}\boldsymbol{Q} - \widetilde{\boldsymbol{\Sigma}}\right\|_{\mathrm{F}}\\ &\le \frac{1}{\sqrt{\sigma_{\min}}}\left\|\boldsymbol{Q}^\top\boldsymbol{\Sigma}\boldsymbol{Q} - \widetilde{\boldsymbol{\Sigma}}\right\|_{\mathrm{F}},\end{aligned} \tag{277}$$

where the last inequality follows from the lower estimates

$$\sigma_{\min}(\boldsymbol{\Sigma}) \ge \sigma_{\min}\left(\boldsymbol{\Sigma}^\star\right) - \|\boldsymbol{M} - \boldsymbol{M}^\star\| \ge \sigma_{\min}/4$$

and, similarly, $\sigma_{\min}(\widetilde{\boldsymbol{\Sigma}}) \ge \sigma_{\min}/4$. Recognizing that $\boldsymbol{\Sigma} = \boldsymbol{U}^\top\boldsymbol{M}\boldsymbol{U}$ and $\widetilde{\boldsymbol{\Sigma}} = \widetilde{\boldsymbol{U}}^\top\widetilde{\boldsymbol{M}}\widetilde{\boldsymbol{U}}$, one gets

$$\begin{aligned}&\left\|\boldsymbol{Q}^\top\boldsymbol{\Sigma}\boldsymbol{Q} - \widetilde{\boldsymbol{\Sigma}}\right\|_{\mathrm{F}} = \left\|(\boldsymbol{U}\boldsymbol{Q})^\top\boldsymbol{M}(\boldsymbol{U}\boldsymbol{Q}) - \widetilde{\boldsymbol{U}}^\top\widetilde{\boldsymbol{M}}\widetilde{\boldsymbol{U}}\right\|_{\mathrm{F}}\\ &\le \left\|(\boldsymbol{U}\boldsymbol{Q})^\top\boldsymbol{M}(\boldsymbol{U}\boldsymbol{Q}) - (\boldsymbol{U}\boldsymbol{Q})^\top\widetilde{\boldsymbol{M}}(\boldsymbol{U}\boldsymbol{Q})\right\|_{\mathrm{F}} + \left\|(\boldsymbol{U}\boldsymbol{Q})^\top\widetilde{\boldsymbol{M}}(\boldsymbol{U}\boldsymbol{Q}) - \widetilde{\boldsymbol{U}}^\top\widetilde{\boldsymbol{M}}(\boldsymbol{U}\boldsymbol{Q})\right\|_{\mathrm{F}}\\ &\quad + \left\|\widetilde{\boldsymbol{U}}^\top\widetilde{\boldsymbol{M}}(\boldsymbol{U}\boldsymbol{Q}) - \widetilde{\boldsymbol{U}}^\top\widetilde{\boldsymbol{M}}\widetilde{\boldsymbol{U}}\right\|_{\mathrm{F}}\\ &\le \left\|(\widetilde{\boldsymbol{M}} - \boldsymbol{M})\boldsymbol{U}\right\|_{\mathrm{F}} + 2\left\|\boldsymbol{U}\boldsymbol{Q} - \widetilde{\boldsymbol{U}}\right\|_{\mathrm{F}}\left\|\widetilde{\boldsymbol{M}}\right\| \le \left\|(\widetilde{\boldsymbol{M}} - \boldsymbol{M})\boldsymbol{U}\right\|_{\mathrm{F}} + 4\sigma_{\max}\left\|\boldsymbol{U}\boldsymbol{Q} - \widetilde{\boldsymbol{U}}\right\|_{\mathrm{F}},\end{aligned} \tag{278}$$

where the last relation holds due to the upper estimate

$$\left\|\widetilde{\boldsymbol{M}}\right\| \le \left\|\boldsymbol{M}^\star\right\| + \left\|\widetilde{\boldsymbol{M}} - \boldsymbol{M}^\star\right\| \le \sigma_{\max} + \sigma_{\min}/4 \le 2\sigma_{\max}.$$

Invoke the Davis–Kahan sinΘ theorem [39] to obtain

$$\left\|\boldsymbol{U}\,\boldsymbol{Q}-\widetilde{\boldsymbol{U}}\right\|_{\mathrm{F}} \leq \frac{c_2}{\sigma_r\left(\boldsymbol{M}\right)-\sigma_{r+1}(\widetilde{\boldsymbol{M}})}\left\|(\widetilde{\boldsymbol{M}}-\boldsymbol{M})\boldsymbol{U}\right\|_{\mathrm{F}} \leq \frac{2c_2}{\sigma_{\min}}\left\|(\widetilde{\boldsymbol{M}}-\boldsymbol{M})\boldsymbol{U}\right\|_{\mathrm{F}}, \tag{279}$$

for some constant $c_2 > 0$, where the last inequality follows from the bounds

$$\begin{aligned}\sigma_r\left(\boldsymbol{M}\right) &\geq \sigma_r\left(\boldsymbol{M}^\star\right)-\|\boldsymbol{M}-\boldsymbol{M}^\star\| \geq 3\sigma_{\min}/4,\\ \sigma_{r+1}(\widetilde{\boldsymbol{M}}) &\leq \sigma_{r+1}\left(\boldsymbol{M}^\star\right)+\|\widetilde{\boldsymbol{M}}-\boldsymbol{M}^\star\| \leq \sigma_{\min}/4.\end{aligned}$$

Combine (277), (278), (279), and the fact $\sigma_{\max}/\sigma_{\min} \leq c_1$ to reach

$$\left\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{Q}-\boldsymbol{Q}\widetilde{\boldsymbol{\Sigma}}^{1/2}\right\|_{\mathrm{F}} \leq \frac{c_3}{\sqrt{\sigma_{\min}}}\left\|(\widetilde{\boldsymbol{M}}-\boldsymbol{M})\boldsymbol{U}\right\|_{\mathrm{F}}$$

for some constant $c_3 > 0$. □

Lemma 47 *Let $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ be a symmetric matrix with the top-r eigendecomposition $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^\top$. Denote $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}^{1/2}$ and $\boldsymbol{X}^\star = \boldsymbol{U}^\star(\boldsymbol{\Sigma}^\star)^{1/2}$, and define*

$$\widehat{\boldsymbol{Q}} := \operatorname*{argmin}_{\boldsymbol{R}\in\mathcal{O}^{r\times r}}\left\|\boldsymbol{U}\boldsymbol{R}-\boldsymbol{U}^\star\right\|_{\mathrm{F}} \quad \text{and} \quad \widehat{\boldsymbol{H}} := \operatorname*{argmin}_{\boldsymbol{R}\in\mathcal{O}^{r\times r}}\left\|\boldsymbol{X}\boldsymbol{R}-\boldsymbol{X}^\star\right\|_{\mathrm{F}}.$$

Assume $\left\|\boldsymbol{M}-\boldsymbol{M}^\star\right\| \leq \sigma_{\min}/2$, and suppose $\sigma_{\max}/\sigma_{\min}$ is bounded by some constant $c_1 > 0$. Then there exists a numerical constant $c_3 > 0$ such that

$$\left\|\widehat{\boldsymbol{Q}}-\widehat{\boldsymbol{H}}\right\| \leq \frac{c_3}{\sigma_{\min}}\left\|\boldsymbol{M}-\boldsymbol{M}^\star\right\|.$$

Proof We first collect several useful facts about the spectrum of $\boldsymbol{\Sigma}$. Weyl's inequality tells us that $\left\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}^\star\right\| \leq \left\|\boldsymbol{M}-\boldsymbol{M}^\star\right\| \leq \sigma_{\min}/2$, which further implies that

$$\sigma_r\left(\boldsymbol{\Sigma}\right) \geq \sigma_r\left(\boldsymbol{\Sigma}^\star\right)-\left\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}^\star\right\| \geq \sigma_{\min}/2 \quad \text{and} \quad \|\boldsymbol{\Sigma}\| \leq \left\|\boldsymbol{\Sigma}^\star\right\|+\left\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}^\star\right\| \leq 2\sigma_{\max}.$$

Denote

$$\boldsymbol{Q} = \boldsymbol{U}^\top\boldsymbol{U}^\star \quad \text{and} \quad \boldsymbol{H} = \boldsymbol{X}^\top\boldsymbol{X}^\star.$$

Simple algebra yields

$$\boldsymbol{H} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{Q}\left(\boldsymbol{\Sigma}^\star\right)^{1/2} = \underbrace{\boldsymbol{\Sigma}^{1/2}\left(\boldsymbol{Q}-\widehat{\boldsymbol{Q}}\right)\left(\boldsymbol{\Sigma}^\star\right)^{1/2}+\left(\boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{Q}}-\widehat{\boldsymbol{Q}}\boldsymbol{\Sigma}^{1/2}\right)\left(\boldsymbol{\Sigma}^\star\right)^{1/2}}_{:=\boldsymbol{E}} + \underbrace{\widehat{\boldsymbol{Q}}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\star\right)^{1/2}}_{:=\boldsymbol{A}}.$$

It can be easily seen that $\sigma_{r-1}(\boldsymbol{A}) \geq \sigma_r(\boldsymbol{A}) \geq \sigma_{\min}/2$, and

$$\begin{aligned}\|\boldsymbol{E}\| &\leq \left\|\boldsymbol{\Sigma}^{1/2}\right\| \cdot \left\|\boldsymbol{Q}-\widehat{\boldsymbol{Q}}\right\| \cdot \left\|\left(\boldsymbol{\Sigma}^{\star}\right)^{1/2}\right\| + \left\|\boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{Q}} - \widehat{\boldsymbol{Q}}\boldsymbol{\Sigma}^{1/2}\right\| \cdot \left\|\left(\boldsymbol{\Sigma}^{\star}\right)^{1/2}\right\| \\ &\leq 2\sigma_{\max}\underbrace{\left\|\boldsymbol{Q}-\widehat{\boldsymbol{Q}}\right\|}_{:=\alpha} + \sqrt{\sigma_{\max}}\underbrace{\left\|\boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{Q}} - \widehat{\boldsymbol{Q}}\boldsymbol{\Sigma}^{1/2}\right\|}_{:=\beta},\end{aligned}$$

which can be controlled as follows.

- Regarding α, use [1, Lemma 3] to reach

$$\alpha = \left\|\boldsymbol{Q}-\widehat{\boldsymbol{Q}}\right\| \leq 4\left\|\boldsymbol{M}-\boldsymbol{M}^{\star}\right\|^2/\sigma_{\min}^2.$$

- For β, one has

$$\beta \overset{\text{(i)}}{=} \left\|\widehat{\boldsymbol{Q}}^{\top}\boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{Q}} - \boldsymbol{\Sigma}^{1/2}\right\| \overset{\text{(ii)}}{\leq} \frac{1}{2\sigma_r\left(\boldsymbol{\Sigma}^{1/2}\right)}\left\|\widehat{\boldsymbol{Q}}^{\top}\boldsymbol{\Sigma}\widehat{\boldsymbol{Q}} - \boldsymbol{\Sigma}\right\| \overset{\text{(iii)}}{=} \frac{1}{2\sigma_r\left(\boldsymbol{\Sigma}^{1/2}\right)}\left\|\boldsymbol{\Sigma}\widehat{\boldsymbol{Q}} - \widehat{\boldsymbol{Q}}\boldsymbol{\Sigma}\right\|,$$

  where (i) and (iii) come from the unitary invariance of $\|\cdot\|$ and (ii) follows from the matrix square root perturbation bound [97, Lemma 2.1]. We can further take the triangle inequality to obtain

$$\begin{aligned}\left\|\boldsymbol{\Sigma}\widehat{\boldsymbol{Q}} - \widehat{\boldsymbol{Q}}\boldsymbol{\Sigma}\right\| &= \left\|\boldsymbol{\Sigma}\boldsymbol{Q} - \boldsymbol{Q}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}(\widehat{\boldsymbol{Q}}-\boldsymbol{Q}) - (\widehat{\boldsymbol{Q}}-\boldsymbol{Q})\boldsymbol{\Sigma}\right\| \\ &\leq \|\boldsymbol{\Sigma}\boldsymbol{Q} - \boldsymbol{Q}\boldsymbol{\Sigma}\| + 2\|\boldsymbol{\Sigma}\|\left\|\boldsymbol{Q}-\widehat{\boldsymbol{Q}}\right\| \\ &= \left\|\boldsymbol{U}\left(\boldsymbol{M}-\boldsymbol{M}^{\star}\right)\boldsymbol{U}^{\star\top} + \boldsymbol{Q}\left(\boldsymbol{\Sigma}^{\star}-\boldsymbol{\Sigma}\right)\right\| + 2\|\boldsymbol{\Sigma}\|\left\|\boldsymbol{Q}-\widehat{\boldsymbol{Q}}\right\| \\ &\leq \left\|\boldsymbol{U}\left(\boldsymbol{M}-\boldsymbol{M}^{\star}\right)\boldsymbol{U}^{\star\top}\right\| + \left\|\boldsymbol{Q}\left(\boldsymbol{\Sigma}^{\star}-\boldsymbol{\Sigma}\right)\right\| + 2\|\boldsymbol{\Sigma}\|\left\|\boldsymbol{Q}-\widehat{\boldsymbol{Q}}\right\| \\ &\leq 2\left\|\boldsymbol{M}-\boldsymbol{M}^{\star}\right\| + 4\sigma_{\max}\alpha,\end{aligned}$$

  where the last inequality uses the Weyl's inequality $\|\boldsymbol{\Sigma}^{\star}-\boldsymbol{\Sigma}\| \leq \|\boldsymbol{M}-\boldsymbol{M}^{\star}\|$ and the fact that $\|\boldsymbol{\Sigma}\| \leq 2\sigma_{\max}$.
- Rearrange the previous bounds to arrive at

$$\|\boldsymbol{E}\| \leq 2\sigma_{\max}\alpha + \sqrt{\sigma_{\max}}\frac{1}{\sqrt{\sigma_{\min}}}\left(2\left\|\boldsymbol{M}-\boldsymbol{M}^{\star}\right\| + 4\sigma_{\max}\alpha\right) \leq c_2\left\|\boldsymbol{M}-\boldsymbol{M}^{\star}\right\|$$

  for some numerical constant $c_2 > 0$, where we have used the assumption that $\sigma_{\max}/\sigma_{\min}$ is bounded.

Recognizing that $\widehat{\boldsymbol{Q}} = \operatorname{sgn}(\boldsymbol{A})$ (see definition in (177)), we are ready to invoke Lemma 36 to deduce that

$$\left\|\widehat{\boldsymbol{Q}}-\widehat{\boldsymbol{H}}\right\| \leq \frac{2}{\sigma_{r-1}(\boldsymbol{A}) + \sigma_r(\boldsymbol{A})}\|\boldsymbol{E}\| \leq \frac{c_3}{\sigma_{\min}}\left\|\boldsymbol{M}-\boldsymbol{M}^{\star}\right\|$$

for some constant $c_3 > 0$. □

D.3 Technical Lemmas for Blind Deconvolution

D.3.1 Wirtinger Calculus

In this section, we formally prove the fundamental theorem of calculus and the mean value form of Taylor's theorem under the Wirtinger calculus; see (283) and (284), respectively.

Let $f : \mathbb{C}^n \to \mathbb{R}$ be a real-valued function. Denote $\boldsymbol{z} = \boldsymbol{x} + i\,\boldsymbol{y} \in \mathbb{C}^n$, and then $f(\cdot)$ can alternatively be viewed as a function $\mathbb{R}^{2n} \to \mathbb{R}$. There is a one-to-one mapping connecting the Wirtinger derivatives and the conventional derivatives [69]:

$$\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} = \boldsymbol{J}^{-1} \begin{bmatrix} \boldsymbol{z} \\ \overline{\boldsymbol{z}} \end{bmatrix}, \tag{280a}$$

$$\nabla_{\mathbb{R}} f\left(\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix}\right) = \boldsymbol{J}^{\mathsf{H}} \nabla_{\mathbb{C}} f\left(\begin{bmatrix} \boldsymbol{z} \\ \overline{\boldsymbol{z}} \end{bmatrix}\right), \tag{280b}$$

$$\nabla^2_{\mathbb{R}} f\left(\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix}\right) = \boldsymbol{J}^{\mathsf{H}} \nabla^2_{\mathbb{C}} f\left(\begin{bmatrix} \boldsymbol{z} \\ \overline{\boldsymbol{z}} \end{bmatrix}\right) \boldsymbol{J}, \tag{280c}$$

where the subscripts $\mathbb{R}$ and $\mathbb{C}$ represent calculus in the real (conventional) sense and in the complex (Wirtinger) sense, respectively, and

$$\boldsymbol{J} = \begin{bmatrix} \boldsymbol{I}_n & i\boldsymbol{I}_n \\ \boldsymbol{I}_n & -i\boldsymbol{I}_n \end{bmatrix}.$$

With these relationships in place, we are ready to verify the fundamental theorem of calculus using the Wirtinger derivatives. Recall from [70, Chapter XIII, Theorem 4.2] that

$$\nabla_{\mathbb{R}} f\left(\begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{y}_1 \end{bmatrix}\right) - \nabla_{\mathbb{R}} f\left(\begin{bmatrix} \boldsymbol{x}_2 \\ \boldsymbol{y}_2 \end{bmatrix}\right) = \left[\int_0^1 \nabla^2_{\mathbb{R}} f\left(\begin{bmatrix} \boldsymbol{x}(\tau) \\ \boldsymbol{y}(\tau) \end{bmatrix}\right) \mathrm{d}\tau\right] \left(\begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{y}_1 \end{bmatrix} - \begin{bmatrix} \boldsymbol{x}_2 \\ \boldsymbol{y}_2 \end{bmatrix}\right), \tag{281}$$

where

$$\begin{bmatrix} \boldsymbol{x}(\tau) \\ \boldsymbol{y}(\tau) \end{bmatrix} := \begin{bmatrix} \boldsymbol{x}_2 \\ \boldsymbol{y}_2 \end{bmatrix} + \tau \left(\begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{y}_1 \end{bmatrix} - \begin{bmatrix} \boldsymbol{x}_2 \\ \boldsymbol{y}_2 \end{bmatrix}\right).$$

Substitute identities (280) into (281) to arrive at

$$\begin{aligned} &\boldsymbol{J}^{\mathsf{H}} \nabla_{\mathbb{C}} f\left(\begin{bmatrix} \boldsymbol{z}_1 \\ \overline{\boldsymbol{z}_1} \end{bmatrix}\right) - \boldsymbol{J}^{\mathsf{H}} \nabla_{\mathbb{C}} f\left(\begin{bmatrix} \boldsymbol{z}_2 \\ \overline{\boldsymbol{z}_2} \end{bmatrix}\right) \\ &= \boldsymbol{J}^{\mathsf{H}} \left[\int_0^1 \nabla^2_{\mathbb{C}} f\left(\begin{bmatrix} \boldsymbol{z}(\tau) \\ \overline{\boldsymbol{z}(\tau)} \end{bmatrix}\right) \mathrm{d}\tau\right] \boldsymbol{J}\boldsymbol{J}^{-1} \left(\begin{bmatrix} \boldsymbol{z}_1 \\ \overline{\boldsymbol{z}_1} \end{bmatrix} - \begin{bmatrix} \boldsymbol{z}_2 \\ \overline{\boldsymbol{z}_2} \end{bmatrix}\right) \\ &= \boldsymbol{J}^{\mathsf{H}} \left[\int_0^1 \nabla^2_{\mathbb{C}} f\left(\begin{bmatrix} \boldsymbol{z}(\tau) \\ \overline{\boldsymbol{z}(\tau)} \end{bmatrix}\right) \mathrm{d}\tau\right] \left(\begin{bmatrix} \boldsymbol{z}_1 \\ \overline{\boldsymbol{z}_1} \end{bmatrix} - \begin{bmatrix} \boldsymbol{z}_2 \\ \overline{\boldsymbol{z}_2} \end{bmatrix}\right), \end{aligned} \tag{282}$$

where $z_1 = \boldsymbol{x}_1 + i\,\boldsymbol{y}_1$, $z_2 = \boldsymbol{x}_2 + i\,\boldsymbol{y}_2$ and

$$\begin{bmatrix} z(\tau) \\ \overline{z(\tau)} \end{bmatrix} := \begin{bmatrix} z_2 \\ \overline{z_2} \end{bmatrix} + \tau \left(\begin{bmatrix} z_1 \\ \overline{z_1} \end{bmatrix} - \begin{bmatrix} z_2 \\ \overline{z_2} \end{bmatrix} \right).$$

Simplification of (282) gives

$$\nabla_{\mathbb{C}} f \left(\begin{bmatrix} z_1 \\ \overline{z_1} \end{bmatrix} \right) - \nabla_{\mathbb{C}} f \left(\begin{bmatrix} z_2 \\ \overline{z_2} \end{bmatrix} \right) = \left[\int_0^1 \nabla_{\mathbb{C}}^2 f \left(\begin{bmatrix} z(\tau) \\ \overline{z(\tau)} \end{bmatrix} \right) \mathrm{d}\tau \right] \left(\begin{bmatrix} z_1 \\ \overline{z_1} \end{bmatrix} - \begin{bmatrix} z_2 \\ \overline{z_2} \end{bmatrix} \right). \tag{283}$$

Repeating the above arguments, one can also show that

$$f(z_1) - f(z_2) = \nabla_{\mathbb{C}} f(z_2)^{\mathsf{H}} \begin{bmatrix} z_1 - z_2 \\ \overline{z_1 - z_2} \end{bmatrix} + \frac{1}{2} \begin{bmatrix} z_1 - z_2 \\ \overline{z_1 - z_2} \end{bmatrix}^{\mathsf{H}} \nabla_{\mathbb{C}}^2 f(\widetilde{z}) \begin{bmatrix} z_1 - z_2 \\ \overline{z_1 - z_2} \end{bmatrix}, \tag{284}$$

where $\widetilde{z}$ is some point lying on the vector connecting z_1 and z_2. This is the mean value form of Taylor's theorem under the Wirtinger calculus.

D.3.2 Discrete Fourier Transform Matrices

Let $\boldsymbol{B} \in \mathbb{C}^{m \times K}$ be the first K columns of a discrete Fourier transform (DFT) matrix $\boldsymbol{F} \in \mathbb{C}^{m \times m}$, and denote by $\boldsymbol{b}_l$ the lth column of the matrix $\boldsymbol{B}^{\mathsf{H}}$. By definition,

$$\boldsymbol{b}_l = \frac{1}{\sqrt{m}} \left(1, \omega^{(l-1)}, \omega^{2(l-1)}, \cdots, \omega^{(K-1)(l-1)} \right)^{\mathsf{H}},$$

where $\omega := e^{-i\frac{2\pi}{m}}$ with i representing the imaginary unit. It is seen that for any $j \neq l$,

$$\begin{aligned} \boldsymbol{b}_l^{\mathsf{H}} \boldsymbol{b}_j = \frac{1}{m} \sum_{k=0}^{K-1} \omega^{k(l-1)} \cdot \overline{\omega^{k(j-1)}} &\overset{\text{(i)}}{=} \frac{1}{m} \sum_{k=0}^{K-1} \omega^{k(l-1)} \cdot \omega^{k(1-j)} = \frac{1}{m} \sum_{k=0}^{K-1} \left(\omega^{l-j} \right)^k \\ &\overset{\text{(ii)}}{=} \frac{1}{m} \frac{1 - \omega^{K(l-j)}}{1 - \omega^{l-j}}. \end{aligned} \tag{285}$$

Here, (i) uses $\overline{\omega^{\alpha}} = \omega^{-\alpha}$ for all $\alpha \in \mathbb{R}$, while the last identity (ii) follows from the formula for the sum of a finite geometric series when $\omega^{l-j} \neq 1$. This leads to the following lemma.

Lemma 48 *For any $m \geq 3$ and any $1 \leq l \leq m$, we have*

$$\sum_{j=1}^{m} \left| \boldsymbol{b}_l^{\mathsf{H}} \boldsymbol{b}_j \right| \leq 4 \log m.$$

Proof We first make use of identity (285) to obtain

$$\sum_{j=1}^{m}\left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{b}_j\right| = \|\boldsymbol{b}_l\|_2^2 + \frac{1}{m}\sum_{j:j\neq l}^{m}\left|\frac{1-\omega^{K(l-j)}}{1-\omega^{l-j}}\right| = \frac{K}{m} + \frac{1}{m}\sum_{j:j\neq l}^{m}\left|\frac{\sin\left[K\,(l-j)\,\frac{\pi}{m}\right]}{\sin\left[(l-j)\,\frac{\pi}{m}\right]}\right|,$$

where the last identity follows since $\|\boldsymbol{b}_l\|_2^2 = K/m$ and, for all $\alpha \in \mathbb{R}$,

$$\left|1-\omega^{\alpha}\right| = \left|1-e^{-i\frac{2\pi}{m}\alpha}\right| = \left|e^{-i\frac{\pi}{m}\alpha}\left(e^{i\frac{\pi}{m}\alpha}-e^{-i\frac{\pi}{m}\alpha}\right)\right| = 2\left|\sin\left(\alpha\frac{\pi}{m}\right)\right|. \tag{286}$$

Without loss of generality, we focus on the case when $l = 1$ in the sequel. Recall that for $c > 0$, we denote by $\lfloor c \rfloor$ the largest integer that does not exceed c. We can continue the derivation to get

$$\begin{aligned}\sum_{j=1}^{m}\left|\boldsymbol{b}_1^{\mathsf{H}}\boldsymbol{b}_j\right| &= \frac{K}{m} + \frac{1}{m}\sum_{j=2}^{m}\left|\frac{\sin\left[K\,(1-j)\,\frac{\pi}{m}\right]}{\sin\left[(1-j)\,\frac{\pi}{m}\right]}\right| \overset{\text{(i)}}{\leq} \frac{1}{m}\sum_{j=2}^{m}\left|\frac{1}{\sin\left[(j-1)\,\frac{\pi}{m}\right]}\right| + \frac{K}{m}\\ &= \frac{1}{m}\left(\sum_{j=2}^{\lfloor\frac{m}{2}\rfloor+1}\left|\frac{1}{\sin\left[(j-1)\,\frac{\pi}{m}\right]}\right| + \sum_{j=\lfloor\frac{m}{2}\rfloor+2}^{m}\left|\frac{1}{\sin\left[(j-1)\,\frac{\pi}{m}\right]}\right|\right) + \frac{K}{m}\\ &\overset{\text{(ii)}}{=} \frac{1}{m}\left(\sum_{j=2}^{\lfloor\frac{m}{2}\rfloor+1}\left|\frac{1}{\sin\left[(j-1)\,\frac{\pi}{m}\right]}\right| + \sum_{j=\lfloor\frac{m}{2}\rfloor+2}^{m}\left|\frac{1}{\sin\left[(m+1-j)\,\frac{\pi}{m}\right]}\right|\right) + \frac{K}{m},\end{aligned}$$

where (i) follows from $\left|\sin\left(K\,(1-j)\,\frac{\pi}{m}\right)\right| \leq 1$ and $|\sin(x)| = |\sin(-x)|$, and (ii) relies on the fact that $\sin(x) = \sin(\pi - x)$. The property that $\sin(x) \geq x/2$ for any $x \in [0, \pi/2]$ allows one to further derive

$$\begin{aligned}\sum_{j=1}^{m}\left|\boldsymbol{b}_1^{\mathsf{H}}\boldsymbol{b}_j\right| &\leq \frac{1}{m}\left(\sum_{j=2}^{\lfloor\frac{m}{2}\rfloor+1}\frac{2m}{(j-1)\,\pi} + \sum_{j=\lfloor\frac{m}{2}\rfloor+2}^{m}\frac{2m}{(m+1-j)\,\pi}\right)\\ &\quad + \frac{K}{m} = \frac{2}{\pi}\left(\sum_{k=1}^{\lfloor\frac{m}{2}\rfloor}\frac{1}{k} + \sum_{k=1}^{\lfloor\frac{m+1}{2}\rfloor-1}\frac{1}{k}\right) + \frac{K}{m}\\ &\overset{\text{(i)}}{\leq} \frac{4}{\pi}\sum_{k=1}^{m}\frac{1}{k} + \frac{K}{m} \overset{\text{(ii)}}{\leq} \frac{4}{\pi}(1+\log m) + 1 \overset{\text{(iii)}}{\leq} 4\log m,\end{aligned}$$

where in (i) we extend the range of the summation, (ii) uses the elementary inequality $\sum_{k=1}^{m} k^{-1} \leq 1 + \log m$, and (iii) holds true as long as $m \geq 3$. □

The next lemma considers the difference of two inner products, namely $(\boldsymbol{b}_l - \boldsymbol{b}_1)^{\mathsf{H}}\boldsymbol{b}_j$.

Lemma 49 *For all* $0 \le l-1 \le \tau \le \lfloor \frac{m}{10} \rfloor$, *we have*

$$\left|(\boldsymbol{b}_l - \boldsymbol{b}_1)^{\mathsf{H}} \boldsymbol{b}_j\right| \le \begin{cases} \frac{4\tau}{(j-l)}\frac{K}{m} + \frac{8\tau/\pi}{(j-l)^2} & \text{for } \ l+\tau \le j \le \lfloor \frac{m}{2} \rfloor + 1, \\ \frac{4\tau}{m-(j-l)}\frac{K}{m} + \frac{8\tau/\pi}{[m-(j-1)]^2} & \text{for } \ \lfloor \frac{m}{2} \rfloor + l \le j \le m-\tau. \end{cases}$$

In addition, for any j *and* l, *the following uniform upper bound holds*

$$\left|(\boldsymbol{b}_l - \boldsymbol{b}_1)^{\mathsf{H}} \boldsymbol{b}_j\right| \le 2\frac{K}{m}.$$

Proof Given (285), we can obtain for $j \ne l$ and $j \ne 1$,

$$\begin{aligned} \left|(\boldsymbol{b}_l - \boldsymbol{b}_1)^{\mathsf{H}} \boldsymbol{b}_j\right| &= \frac{1}{m}\left|\frac{1-\omega^{K(l-j)}}{1-\omega^{l-j}} - \frac{1-\omega^{K(1-j)}}{1-\omega^{1-j}}\right| \\ &= \frac{1}{m}\left|\frac{1-\omega^{K(l-j)}}{1-\omega^{l-j}} - \frac{1-\omega^{K(1-j)}}{1-\omega^{l-j}} + \frac{1-\omega^{K(1-j)}}{1-\omega^{l-j}} - \frac{1-\omega^{K(1-j)}}{1-\omega^{1-j}}\right| \\ &= \frac{1}{m}\left|\frac{\omega^{K(1-j)}-\omega^{K(l-j)}}{1-\omega^{l-j}} + \left(\omega^{l-j}-\omega^{1-j}\right)\frac{1-\omega^{K(1-j)}}{\left(1-\omega^{l-j}\right)\left(1-\omega^{1-j}\right)}\right| \\ &\le \frac{1}{m}\left|\frac{1-\omega^{K(l-1)}}{1-\omega^{l-j}}\right| + \frac{2}{m}\left|\left(1-\omega^{1-l}\right)\frac{1}{\left(1-\omega^{l-j}\right)\left(1-\omega^{1-j}\right)}\right|, \end{aligned}$$

where the last line is due to the triangle inequality and $|\omega^{\alpha}| = 1$ for all $\alpha \in \mathbb{R}$. Identity (286) allows us to rewrite this bound as

$$\left|(\boldsymbol{b}_l - \boldsymbol{b}_1)^{\mathsf{H}} \boldsymbol{b}_j\right| \le \frac{1}{m}\left|\frac{1}{\sin\left[(l-j)\frac{\pi}{m}\right]}\right|\left\{\left|\sin\left[K(l-1)\frac{\pi}{m}\right]\right| + \left|\frac{\sin\left[(1-l)\frac{\pi}{m}\right]}{\sin\left[(1-j)\frac{\pi}{m}\right]}\right|\right\}. \tag{287}$$

Combined with the fact that $|\sin x| \le 2|x|$ for all $x \in \mathbb{R}$, we can upper bound (287) as

$$\left|(\boldsymbol{b}_l - \boldsymbol{b}_1)^{\mathsf{H}} \boldsymbol{b}_j\right| \le \frac{1}{m}\left|\frac{1}{\sin\left[(l-j)\frac{\pi}{m}\right]}\right|\left\{2K\tau\frac{\pi}{m} + \left|\frac{2\tau\frac{\pi}{m}}{\sin\left[(1-j)\frac{\pi}{m}\right]}\right|\right\},$$

where we also utilize the assumption $0 \le l-1 \le \tau$. Then for $l+\tau \le j \le \lfloor m/2 \rfloor + 1$, one has

$$\left|(l-j)\frac{\pi}{m}\right| \le \frac{\pi}{2} \qquad \text{and} \qquad \left|(1-j)\frac{\pi}{m}\right| \le \frac{\pi}{2}.$$

Therefore, utilizing the property $\sin(x) \ge x/2$ for any $x \in [0, \pi/2]$, we arrive at

$$\left|(\boldsymbol{b}_l - \boldsymbol{b}_1)^{\mathsf{H}} \boldsymbol{b}_j\right| \le \frac{2}{(j-l)\pi}\left(2K\tau\frac{\pi}{m} + \frac{4\tau}{j-1}\right) \le \frac{4\tau}{(j-l)}\frac{K}{m} + \frac{8\tau/\pi}{(j-l)^2},$$

where the last inequality holds since $j-1 > j-l$. Similarly, we can obtain the upper bound for $\lfloor m/2 \rfloor + l \le j \le m-\tau$ using nearly identical argument (which is omitted for brevity).

The uniform upper bound can be justified as follows

$$\left|(\boldsymbol{b}_l - \boldsymbol{b}_1)^{\mathsf{H}} \boldsymbol{b}_j\right| \le (\|\boldsymbol{b}_l\|_2 + \|\boldsymbol{b}_1\|_2) \left\|\boldsymbol{b}_j\right\|_2 \le 2K/m.$$

The last relation holds since $\|\boldsymbol{b}_l\|_2^2 = K/m$ for all $1 \le l \le m$. □

Next, we list two consequences of the above estimates in Lemmas 50 and 51.

Lemma 50 *Fix any constant $c > 0$ that is independent of m and K. Suppose $m \ge C\tau K \log^4 m$ for some sufficiently large constant $C > 0$, which solely depends on c. If $0 \le l-1 \le \tau$, then one has*

$$\sum_{j=1}^{m} \left|(\boldsymbol{b}_l - \boldsymbol{b}_1)^{\mathsf{H}} \boldsymbol{b}_j\right| \le \frac{c}{\log^2 m}.$$

Proof For some constant $c_0 > 0$, we can split the index set $[m]$ into the following three disjoint sets

$$\begin{aligned} \mathcal{A}_1 &= \left\{ j : l + c_0\tau \log^2 m \le j \le \left\lfloor \frac{m}{2} \right\rfloor \right\}, \\ \mathcal{A}_2 &= \left\{ j : \left\lfloor \frac{m}{2} \right\rfloor + l \le j \le m - c_0\tau \log^2 m \right\}, \\ \text{and} \quad \mathcal{A}_3 &= [m] \setminus (\mathcal{A}_1 \cup \mathcal{A}_2). \end{aligned}$$

With this decomposition in place, we can write

$$\sum_{j=1}^{m} \left|(\boldsymbol{b}_l - \boldsymbol{b}_1)^{\mathsf{H}} \boldsymbol{b}_j\right| = \sum_{j\in\mathcal{A}_1} \left|(\boldsymbol{b}_l - \boldsymbol{b}_1)^{\mathsf{H}} \boldsymbol{b}_j\right| + \sum_{j\in\mathcal{A}_2} \left|(\boldsymbol{b}_l - \boldsymbol{b}_1)^{\mathsf{H}} \boldsymbol{b}_j\right| + \sum_{j\in\mathcal{A}_3} \left|(\boldsymbol{b}_l - \boldsymbol{b}_1)^{\mathsf{H}} \boldsymbol{b}_j\right|.$$

We first look at $\mathcal{A}_1$. By Lemma 49, one has for any $j \in \mathcal{A}_1$,

$$\left|(\boldsymbol{b}_l - \boldsymbol{b}_1)^{\mathsf{H}} \boldsymbol{b}_j\right| \le \frac{4\tau}{j-l}\frac{K}{m} + \frac{8\tau/\pi}{(j-l)^2},$$

and hence

$$\begin{aligned} \sum_{j\in\mathcal{A}_1} \left|(\boldsymbol{b}_l - \boldsymbol{b}_1)^{\mathsf{H}} \boldsymbol{b}_j\right| &\le \sum_{j=l+c_0\tau\log^2 m}^{\lfloor \frac{m}{2} \rfloor + 1} \left(\frac{4\tau}{j-l}\frac{K}{m} + \frac{8\tau/\pi}{(j-l)^2} \right) \\ &\le \frac{4\tau K}{m} \sum_{k=1}^{m} \frac{1}{k} + \frac{8\tau}{\pi} \sum_{k=c_0\tau\log^2 m}^{m} \frac{1}{k^2} \end{aligned}$$

$$\leq 8\tau \frac{K}{m} \log m + \frac{16\tau}{\pi} \frac{1}{c_0 \tau \log^2 m},$$

where the last inequality arises from $\sum_{k=1}^{m} k^{-1} \leq 1 + \log m \leq 2 \log m$ and $\sum_{k=c}^{m} k^{-2} \leq 2/c$.

Similarly, for $j \in \mathcal{A}_2$, we have

$$\left| (\boldsymbol{b}_l - \boldsymbol{b}_1)^{\mathsf{H}} \boldsymbol{b}_j \right| \leq \frac{4\tau}{m - (j-l)} \frac{K}{m} + \frac{8\tau/\pi}{[m - (j-1)]^2},$$

which in turn implies

$$\sum_{j \in \mathcal{A}_2} \left| (\boldsymbol{b}_l - \boldsymbol{b}_1)^{\mathsf{H}} \boldsymbol{b}_j \right| \leq 8\tau \frac{K}{m} \log m + \frac{16\tau}{\pi} \frac{1}{c_0 \tau \log^2 m}.$$

Regarding $j \in \mathcal{A}_3$, we observe that

$$|\mathcal{A}_3| \leq 2 \left(c_0 \tau \log^2 m + l \right) \leq 2 \left(c_0 \tau \log^2 m + \tau + 1 \right) \leq 4 c_0 \tau \log^2 m.$$

This together with the simple bound $\left| (\boldsymbol{b}_l - \boldsymbol{b}_1)^{\mathsf{H}} \boldsymbol{b}_j \right| \leq 2K/m$ gives

$$\sum_{j \in \mathcal{A}_3} \left| (\boldsymbol{b}_l - \boldsymbol{b}_1)^{\mathsf{H}} \boldsymbol{b}_j \right| \leq 2 \frac{K}{m} |\mathcal{A}_3| \leq \frac{8 c_0 \tau K \log^2 m}{m}.$$

The previous three estimates taken collectively yield

$$\sum_{j=1}^{m} \left| (\boldsymbol{b}_l - \boldsymbol{b}_1)^{\mathsf{H}} \boldsymbol{b}_j \right| \leq \frac{16\tau K \log m}{m} + \frac{32\tau}{\pi} \frac{1}{c_0 \tau \log^2 m} + \frac{8 c_0 \tau K \log^2 m}{m} \leq c \frac{1}{\log^2 m}$$

as long as $c_0 \geq (32/\pi) \cdot (1/c)$ and $m \geq 8 c_0 \tau K \log^4 m / c$. □

Lemma 51 *Fix any constant $c > 0$ that is independent of m and K. Consider an integer $\tau > 0$, and suppose that $m \geq C \tau K \log m$ for some large constant $C > 0$, which depends solely on c. Then we have*

$$\sum_{k=0}^{\lfloor m/\tau \rfloor} \sqrt{\sum_{j=1}^{\tau} \left| \boldsymbol{b}_1^{\mathsf{H}} \left(\boldsymbol{b}_{k\tau+j} - \boldsymbol{b}_{k\tau+1} \right) \right|^2} \leq \frac{c}{\sqrt{\tau}}.$$

Proof The proof strategy is similar to the one used in Lemma 50. First, notice that

$$\left| \boldsymbol{b}_1^{\mathsf{H}} \left(\boldsymbol{b}_{k\tau+j} - \boldsymbol{b}_{k\tau+1} \right) \right| = \left| \left(\boldsymbol{b}_m - \boldsymbol{b}_{m+1-j} \right)^{\mathsf{H}} \boldsymbol{b}_{k\tau} \right|.$$

As before, for some $c_1 > 0$, we can split the index set $\{1, \cdots, \lfloor m/\tau \rfloor\}$ into three disjoint sets

$$\begin{aligned}
\mathcal{B}_1 &= \left\{k : c_1 \le k \le \left\lfloor \left(\left\lfloor \frac{m}{2} \right\rfloor + 1 - j \right) / \tau \right\rfloor \right\}, \\
\mathcal{B}_2 &= \left\{k : \left\lfloor \left(\left\lfloor \frac{m}{2} \right\rfloor + 1 - j \right) / \tau \right\rfloor + 1 \le k \le \lfloor (m+1-j)/\tau \rfloor - c_1 \right\}, \\
\text{and} \quad \mathcal{B}_3 &= \left\{1, \cdots, \left\lfloor \frac{m}{\tau} \right\rfloor \right\} \setminus (\mathcal{B}_1 \cup \mathcal{B}_2),
\end{aligned}$$

where $1 \le j \le \tau$.

By Lemma 49, one has

$$\left| \left(\boldsymbol{b}_m - \boldsymbol{b}_{m+1-j} \right)^{\mathsf{H}} \boldsymbol{b}_{k\tau} \right| \le \frac{4\tau}{k\tau} \frac{K}{m} + \frac{8\tau/\pi}{(k\tau)^2}, \qquad k \in \mathcal{B}_1.$$

Hence, for any $k \in \mathcal{B}_1$,

$$\sqrt{\sum_{j=1}^{\tau} \left| \boldsymbol{b}_1^{\mathsf{H}} \left(\boldsymbol{b}_{k\tau+j} - \boldsymbol{b}_{k\tau+1} \right) \right|^2} \le \sqrt{\tau} \left(\frac{4\tau}{k\tau} \frac{K}{m} + \frac{8\tau/\pi}{(k\tau)^2} \right) = \sqrt{\tau} \left(\frac{4}{k} \frac{K}{m} + \frac{8/\pi}{k^2 \tau} \right),$$

which further implies that

$$\begin{aligned}
\sum_{k \in \mathcal{B}_1} \sqrt{\sum_{j=1}^{\tau} \left| \boldsymbol{b}_1^{\mathsf{H}} \left(\boldsymbol{b}_{k\tau+j} - \boldsymbol{b}_{k\tau+1} \right) \right|^2} &\le \sqrt{\tau} \sum_{k=c_1}^{m} \left(\frac{4}{k} \frac{K}{m} + \frac{8/\pi}{k^2 \tau} \right) \\
&\le 8\sqrt{\tau} \frac{K \log m}{m} + \frac{16}{\pi} \frac{1}{\sqrt{\tau}} \frac{1}{c_1},
\end{aligned}$$

where the last inequality follows since $\sum_{k=1}^{m} k^{-1} \le 2 \log m$ and $\sum_{k=c_1}^{m} k^{-2} \le 2/c_1$. A similar bound can be obtained for $k \in \mathcal{B}_2$.

For the remaining set $\mathcal{B}_3$, observe that

$$|\mathcal{B}_3| \le 2c_1.$$

This together with the crude upper bound $\left| (\boldsymbol{b}_l - \boldsymbol{b}_1)^{\mathsf{H}} \boldsymbol{b}_j \right| \le 2K/m$ gives

$$\begin{aligned}
\sum_{k \in \mathcal{B}_3} \sqrt{\sum_{j=1}^{\tau} \left| \boldsymbol{b}_1^{\mathsf{H}} \left(\boldsymbol{b}_{k\tau+j} - \boldsymbol{b}_{k\tau+1} \right) \right|^2} &\le |\mathcal{B}_3| \sqrt{\tau \max_j \left| \boldsymbol{b}_1^{\mathsf{H}} \left(\boldsymbol{b}_{k\tau+j} - \boldsymbol{b}_{k\tau+1} \right) \right|^2} \\
&\le |\mathcal{B}_3| \sqrt{\tau} \cdot \frac{2K}{m} \le \frac{4c_1 \sqrt{\tau} K}{m}.
\end{aligned}$$

The previous estimates taken collectively yield

$$\sum_{k=0}^{\lfloor m/\tau \rfloor} \sqrt{\sum_{j=1}^{\tau} \left| \boldsymbol{b}_1^{\mathsf{H}} \left(\boldsymbol{b}_{k\tau+j} - \boldsymbol{b}_{k\tau+1} \right) \right|^2} \leq 2 \left(8\sqrt{\tau} \frac{K \log m}{m} + \frac{16}{\pi} \frac{1}{\sqrt{\tau}} \frac{1}{c_1} \right) + \frac{4c_1 \sqrt{\tau} K}{m} \leq c \frac{1}{\sqrt{\tau}},$$

as long as $c_1 \gg 1/c$ and $m/(c_1 \tau K \log m) \gg 1/c$. □

D.3.3 Complex-Valued Alignment

Let $g_{\boldsymbol{h},\boldsymbol{x}}(\cdot) : \mathbb{C} \to \mathbb{R}$ be a real-valued function defined as

$$g_{\boldsymbol{h},\boldsymbol{x}}(\alpha) := \left\| \frac{1}{\overline{\alpha}} \boldsymbol{h} - \boldsymbol{h}^\star \right\|_2^2 + \left\| \alpha \boldsymbol{x} - \boldsymbol{x}^\star \right\|_2^2,$$

which is the key function in definition (34). Therefore, the alignment parameter of $(\boldsymbol{h}, \boldsymbol{x})$ to $(\boldsymbol{h}^\star, \boldsymbol{x}^\star)$ is the minimizer of $g_{\boldsymbol{h},\boldsymbol{x}}(\alpha)$. This section is devoted to studying various properties of $g_{\boldsymbol{h},\boldsymbol{x}}(\cdot)$. To begin with, the Wirtinger gradient and Hessian of $g_{\boldsymbol{h},\boldsymbol{x}}(\cdot)$ can be calculated as

$$\nabla g_{\boldsymbol{h},\boldsymbol{x}}(\alpha) = \begin{bmatrix} \frac{\partial g_{\boldsymbol{h},\boldsymbol{x}}(\alpha,\overline{\alpha})}{\partial \alpha} \\ \frac{\partial g_{\boldsymbol{h},\boldsymbol{x}}(\alpha,\overline{\alpha})}{\partial \overline{\alpha}} \end{bmatrix} = \begin{bmatrix} \alpha \|\boldsymbol{x}\|_2^2 - \boldsymbol{x}^{\mathsf{H}} \boldsymbol{x}^\star - \alpha^{-1} (\overline{\alpha})^{-2} \|\boldsymbol{h}\|_2^2 + (\overline{\alpha})^{-2} \boldsymbol{h}^{\star\mathsf{H}} \boldsymbol{h} \\ \overline{\alpha} \|\boldsymbol{x}\|_2^2 - \boldsymbol{x}^{\star\mathsf{H}} \boldsymbol{x} - (\overline{\alpha})^{-1} \alpha^{-2} \|\boldsymbol{h}\|_2^2 + \alpha^{-2} \boldsymbol{h}^{\mathsf{H}} \boldsymbol{h}^\star \end{bmatrix}; \tag{288}$$

$$\nabla^2 g_{\boldsymbol{h},\boldsymbol{x}}(\alpha) = \begin{bmatrix} \|\boldsymbol{x}\|_2^2 + |\alpha|^{-4} \|\boldsymbol{h}\|_2^2 & 2\alpha^{-1} (\overline{\alpha})^{-3} \|\boldsymbol{h}\|_2^2 - 2 (\overline{\alpha})^{-3} \boldsymbol{h}^{\star\mathsf{H}} \boldsymbol{h} \\ 2 (\overline{\alpha})^{-1} \alpha^{-3} \|\boldsymbol{h}\|_2^2 - 2\alpha^{-3} \boldsymbol{h}^{\mathsf{H}} \boldsymbol{h}^\star & \|\boldsymbol{x}\|_2^2 + |\alpha|^{-4} \|\boldsymbol{h}\|_2^2 \end{bmatrix}. \tag{289}$$

The first lemma reveals that, as long as $\left(\frac{1}{\beta} \boldsymbol{h}, \beta \boldsymbol{x} \right)$ is sufficiently close to $(\boldsymbol{h}^\star, \boldsymbol{x}^\star)$, the minimizer of $g_{\boldsymbol{h},\boldsymbol{x}}(\alpha)$ cannot be far away from β.

Lemma 52 *Assume there exists $\beta \in \mathbb{C}$ with $1/2 \leq |\beta| \leq 3/2$ such that $\max \left\{ \left\| \frac{1}{\beta} \boldsymbol{h} - \boldsymbol{h}^\star \right\|_2, \|\beta \boldsymbol{x} - \boldsymbol{x}^\star\|_2 \right\} \leq \delta \leq 1/4$. Denote by $\widehat{\alpha}$ the minimizer of $g_{\boldsymbol{h},\boldsymbol{x}}(\alpha)$, and then we necessarily have*

$$\left| |\widehat{\alpha}| - |\beta| \right| \leq |\widehat{\alpha} - \beta| \leq 18\delta.$$

Proof The first inequality is a direct consequence of the triangle inequality. Hence, we concentrate on the second one. Notice that by assumption,

$$g_{\boldsymbol{h},\boldsymbol{x}}(\beta) = \left\| \frac{1}{\overline{\beta}} \boldsymbol{h} - \boldsymbol{h}^\star \right\|_2^2 + \left\| \beta \boldsymbol{x} - \boldsymbol{x}^\star \right\|_2^2 \leq 2\delta^2, \tag{290}$$

which immediately implies that $g_{\boldsymbol{h},\boldsymbol{x}}(\widehat{\alpha}) \leq 2\delta^2$. It thus suffices to show that for any α obeying $|\alpha - \beta| > 18\delta$, one has $g_{\boldsymbol{h},\boldsymbol{x}}(\alpha) > 2\delta^2$, and hence, it cannot be the minimizer.

To this end, we lower bound $g_{\boldsymbol{h},\boldsymbol{x}}(\alpha)$ as follows:

$$\begin{aligned} g_{\boldsymbol{h},\boldsymbol{x}}(\alpha) &\geq \left\|\alpha\boldsymbol{x}-\boldsymbol{x}^\star\right\|_2^2 = \left\|(\alpha-\beta)\,\boldsymbol{x}+\left(\beta\boldsymbol{x}-\boldsymbol{x}^\star\right)\right\|_2^2 \\ &= |\alpha-\beta|^2\,\|\boldsymbol{x}\|_2^2+\left\|\beta\boldsymbol{x}-\boldsymbol{x}^\star\right\|_2^2+2\mathrm{Re}\left[(\alpha-\beta)\left(\beta\boldsymbol{x}-\boldsymbol{x}^\star\right)^{\mathsf{H}}\boldsymbol{x}\right] \\ &\geq |\alpha-\beta|^2\,\|\boldsymbol{x}\|_2^2-2\,|\alpha-\beta|\left|\left(\beta\boldsymbol{x}-\boldsymbol{x}^\star\right)^{\mathsf{H}}\boldsymbol{x}\right|. \end{aligned}$$

Given that $\|\beta\boldsymbol{x}-\boldsymbol{x}^\star\|_2\leq\delta\leq 1/4$ and $\|\boldsymbol{x}^\star\|_2=1$, we have

$$\|\beta\boldsymbol{x}\|_2\geq\|\boldsymbol{x}^\star\|_2-\left\|\beta\boldsymbol{x}-\boldsymbol{x}^\star\right\|_2\geq 1-\delta\geq 3/4,$$

which together with the fact that $1/2\leq|\beta|\leq 3/2$ implies

$$\|\boldsymbol{x}\|_2\geq 1/2\qquad\text{and}\qquad\|\boldsymbol{x}\|_2\leq 2$$

and

$$\left|\left(\beta\boldsymbol{x}-\boldsymbol{x}^\star\right)^{\mathsf{H}}\boldsymbol{x}\right|\leq\left\|\beta\boldsymbol{x}-\boldsymbol{x}^\star\right\|_2\|\boldsymbol{x}\|_2\leq 2\delta.$$

Taking the previous estimates collectively yields

$$g_{\boldsymbol{h},\boldsymbol{x}}(\alpha)\geq\frac{1}{4}|\alpha-\beta|^2-4\delta\,|\alpha-\beta|.$$

It is self-evident that once $|\alpha-\beta|>18\delta$, one gets $g_{\boldsymbol{h},\boldsymbol{x}}(\alpha)>2\delta^2$, and hence, α cannot be the minimizer as $g_{\boldsymbol{h},\boldsymbol{x}}(\alpha)>g_{\boldsymbol{h},\boldsymbol{x}}(\beta)$ according to (290). This concludes the proof. □

The next lemma reveals the local strong convexity of $g_{\boldsymbol{h},\boldsymbol{x}}(\alpha)$ when α is close to one.

Lemma 53 *Assume that* $\max\left\{\left\|\boldsymbol{h}-\boldsymbol{h}^\star\right\|_2,\|\boldsymbol{x}-\boldsymbol{x}^\star\|_2\right\}\leq\delta$ *for some sufficiently small constant* $\delta>0$. *Then, for any* α *satisfying* $|\alpha-1|\leq 18\delta$ *and any* $u,v\in\mathbb{C}$, *one has*

$$\begin{bmatrix}u^{\mathsf{H}},v^{\mathsf{H}}\end{bmatrix}\nabla^2 g_{\boldsymbol{h},\boldsymbol{x}}(\alpha)\begin{bmatrix}u\\v\end{bmatrix}\geq\frac{1}{2}\left(|u|^2+|v|^2\right),$$

where $\nabla^2 g_{\boldsymbol{h},\boldsymbol{x}}(\cdot)$ *stands for the Wirtinger Hessian of* $g_{\boldsymbol{h},\boldsymbol{x}}(\cdot)$.

Proof For simplicity of presentation, we use $g_{\boldsymbol{h},\boldsymbol{x}}(\alpha,\overline{\alpha})$ and $g_{\boldsymbol{h},\boldsymbol{x}}(\alpha)$ interchangeably. By (289), for any $u,v\in\mathbb{C}$, one has

$$\begin{bmatrix}u^{\mathsf{H}},v^{\mathsf{H}}\end{bmatrix}\nabla^2 g_{\boldsymbol{h},\boldsymbol{x}}(\alpha)\begin{bmatrix}u\\v\end{bmatrix}=\underbrace{\left(\|\boldsymbol{x}\|_2^2+|\alpha|^{-4}\,\|\boldsymbol{h}\|_2^2\right)}_{:=\beta_1}\left(|u|^2+|v|^2\right)$$

$$+2\underbrace{\operatorname{Re}\left[u^{\mathsf{H}}v\left(2\alpha^{-1}\left(\overline{\alpha}\right)^{-3}\|\boldsymbol{h}\|_2^2-2\left(\overline{\alpha}\right)^{-3}\boldsymbol{h}^{\star\mathsf{H}}\boldsymbol{h}\right)\right]}_{:=\beta_2}.$$

We would like to demonstrate that this is at least on the order of $|u|^2+|v|^2$. We first develop a lower bound on β_1. Given the assumption that $\max\left\{\left\|\boldsymbol{h}-\boldsymbol{h}^\star\right\|_2,\|\boldsymbol{x}-\boldsymbol{x}^\star\|_2\right\}\le\delta$, one necessarily has

$$1-\delta\le\|\boldsymbol{x}\|_2\le 1+\delta\qquad\text{and}\qquad 1-\delta\le\|\boldsymbol{h}\|_2\le 1+\delta.$$

Thus, for any α obeying $|\alpha-1|\le 18\delta$, one has

$$\beta_1\ge\left(1+|\alpha|^{-4}\right)(1-\delta)^2\ge\left(1+(1+18\delta)^{-4}\right)(1-\delta)^2\ge 1$$

as long as $\delta>0$ is sufficiently small. Regarding the second term β_2, we utilize the conditions $|\alpha-1|\le 18\delta$, $\|\boldsymbol{x}\|_2\le 1+\delta$ and $\|\boldsymbol{h}\|_2\le 1+\delta$ to get

$$\begin{aligned}
|\beta_2| &\le 2|u||v||\alpha|^{-3}\left|\alpha^{-1}\|\boldsymbol{h}\|_2^2-\boldsymbol{h}^{\star\mathsf{H}}\boldsymbol{h}\right|\\
&= 2|u||v||\alpha|^{-3}\left|\left(\alpha^{-1}-1\right)\|\boldsymbol{h}\|_2^2-(\boldsymbol{h}^\star-\boldsymbol{h})^{\mathsf{H}}\boldsymbol{h}\right|\\
&\le 2|u||v||\alpha|^{-3}\left(\left|\alpha^{-1}-1\right|\|\boldsymbol{h}\|_2^2+\left\|\boldsymbol{h}-\boldsymbol{h}^\star\right\|_2\|\boldsymbol{h}\|_2\right)\\
&\le 2|u||v|(1-18\delta)^{-3}\left(\frac{18\delta}{1-18\delta}(1+\delta)^2+\delta(1+\delta)\right)\\
&\lesssim \delta\left(|u|^2+|v|^2\right),
\end{aligned}$$

where the last relation holds since $2|u||v|\le|u|^2+|v|^2$ and $\delta>0$ is sufficiently small. Combining the previous bounds on β_1 and β_2, we arrive at

$$\left[u^{\mathsf{H}},v^{\mathsf{H}}\right]\nabla^2 g_{\boldsymbol{h},\boldsymbol{x}}(\alpha)\begin{bmatrix}u\\v\end{bmatrix}\ge(1-O(\delta))\left(|u|^2+|v|^2\right)\ge\frac{1}{2}\left(|u|^2+|v|^2\right)$$

as long as δ is sufficiently small. This completes the proof. □

Additionally, in a local region surrounding the optimizer, the alignment parameter is Lipschitz continuous; namely, the difference of the alignment parameters associated with two distinct vector pairs is at most proportional to the ℓ_2 distance between the two vector pairs involved, as demonstrated below.

Lemma 54 *Suppose that the vectors $\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{h}_1,\boldsymbol{h}_2\in\mathbb{C}^K$ satisfy*

$$\max\left\{\left\|\boldsymbol{x}_1-\boldsymbol{x}^\star\right\|_2,\left\|\boldsymbol{h}_1-\boldsymbol{h}^\star\right\|_2,\left\|\boldsymbol{x}_2-\boldsymbol{x}^\star\right\|_2,\left\|\boldsymbol{h}_2-\boldsymbol{h}^\star\right\|_2\right\}\le\delta\le 1/4\qquad(291)$$

for some sufficiently small constant $\delta>0$. Denote by α_1 and α_2 the minimizers of $g_{\boldsymbol{h}_1,\boldsymbol{x}_1}(\alpha)$ and $g_{\boldsymbol{h}_2,\boldsymbol{x}_2}(\alpha)$, respectively. Then we have

$$|\alpha_1-\alpha_2|\lesssim\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_2+\|\boldsymbol{h}_1-\boldsymbol{h}_2\|_2.$$

Proof Since α_1 minimizes $g_{\boldsymbol{h}_1,\boldsymbol{x}_1}(\alpha)$, the mean value form of Taylor's theorem (see Appendix D.3.1) gives

$$\begin{aligned} g_{\boldsymbol{h}_1,\boldsymbol{x}_1}(\alpha_2) &\geq g_{\boldsymbol{h}_1,\boldsymbol{x}_1}(\alpha_1) \\ &= g_{\boldsymbol{h}_1,\boldsymbol{x}_1}(\alpha_2) + \nabla g_{\boldsymbol{h}_1,\boldsymbol{x}_1}(\alpha_2)^{\mathsf{H}} \begin{bmatrix} \alpha_1-\alpha_2 \\ \alpha_1-\alpha_2 \end{bmatrix} \\ &\quad + \frac{1}{2}\left(\overline{\alpha_1-\alpha_2}, \alpha_1-\alpha_2\right) \nabla^2 g_{\boldsymbol{h}_1,\boldsymbol{x}_1}(\widetilde{\alpha}) \begin{bmatrix} \alpha_1-\alpha_2 \\ \alpha_1-\alpha_2 \end{bmatrix}, \end{aligned}$$

where $\widetilde{\alpha}$ is some complex number lying between α_1 and α_2, and $\nabla g_{\boldsymbol{h}_1,\boldsymbol{x}_1}$ and $\nabla^2 g_{\boldsymbol{h}_1,\boldsymbol{x}_1}$ are the Wirtinger gradient and Hessian of $g_{\boldsymbol{h}_1,\boldsymbol{x}_1}(\cdot)$, respectively. Rearrange the previous inequality to obtain

$$|\alpha_1-\alpha_2| \lesssim \frac{\left\|\nabla g_{\boldsymbol{h}_1,\boldsymbol{x}_1}(\alpha_2)\right\|_2}{\lambda_{\min}\left(\nabla^2 g_{\boldsymbol{h}_1,\boldsymbol{x}_1}(\widetilde{\alpha})\right)} \tag{292}$$

as long as $\lambda_{\min}\left(\nabla^2 g_{\boldsymbol{h}_1,\boldsymbol{x}_1}(\widetilde{\alpha})\right) > 0$. This calls for evaluation of the Wirtinger gradient and Hessian of $g_{\boldsymbol{h}_1,\boldsymbol{x}_1}(\cdot)$.

Regarding the Wirtinger Hessian, by assumption (291), we can invoke Lemma 52 with $\beta = 1$ to reach $\max\{|\alpha_1-1|, |\alpha_2-1|\} \leq 18\delta$. This together with Lemma 53 implies

$$\lambda_{\min}\left(\nabla^2 g_{\boldsymbol{h}_1,\boldsymbol{x}_1}(\widetilde{\alpha})\right) \geq 1/2,$$

since $\widetilde{\alpha}$ lies between α_1 and α_2.

For the Wirtinger gradient, since α_2 is the minimizer of $g_{\boldsymbol{h}_2,\boldsymbol{x}_2}(\alpha)$, the first-order optimality condition [69, equation (38)] requires $\nabla g_{\boldsymbol{h}_2,\boldsymbol{x}_2}(\alpha_2) = \boldsymbol{0}$, which gives

$$\left\|\nabla g_{\boldsymbol{h}_1,\boldsymbol{x}_1}(\alpha_2)\right\|_2 = \left\|\nabla g_{\boldsymbol{h}_1,\boldsymbol{x}_1}(\alpha_2) - \nabla g_{\boldsymbol{h}_2,\boldsymbol{x}_2}(\alpha_2)\right\|_2.$$

Plug in the gradient expression (288) to reach

$$\begin{aligned} &\left\|\nabla g_{\boldsymbol{h}_1,\boldsymbol{x}_1}(\alpha_2) - \nabla g_{\boldsymbol{h}_2,\boldsymbol{x}_2}(\alpha_2)\right\|_2 \\ &= \sqrt{2}\left|\left[\alpha_2\|\boldsymbol{x}_1\|_2^2 - \boldsymbol{x}_1^{\mathsf{H}}\boldsymbol{x}^\star - \alpha_2^{-1}(\overline{\alpha_2})^{-2}\|\boldsymbol{h}_1\|_2^2 + (\overline{\alpha_2})^{-2}\boldsymbol{h}^{\star\mathsf{H}}\boldsymbol{h}_1\right]\right. \\ &\quad \left. - \left[\alpha_2\|\boldsymbol{x}_2\|_2^2 - \boldsymbol{x}_2^{\mathsf{H}}\boldsymbol{x}^\star - \alpha_2^{-1}(\overline{\alpha_2})^{-2}\|\boldsymbol{h}_2\|_2^2 + (\overline{\alpha_2})^{-2}\boldsymbol{h}^{\star\mathsf{H}}\boldsymbol{h}_2\right]\right| \\ &\lesssim |\alpha_2|\left|\|\boldsymbol{x}_1\|_2^2 - \|\boldsymbol{x}_2\|_2^2\right| + \left|\boldsymbol{x}_1^{\mathsf{H}}\boldsymbol{x}^\star - \boldsymbol{x}_2^{\mathsf{H}}\boldsymbol{x}^\star\right| + \frac{1}{|\alpha_2|^3}\left|\|\boldsymbol{h}_1\|_2^2 - \|\boldsymbol{h}_2\|_2^2\right| + \frac{1}{|\alpha_2|^2}\left|\boldsymbol{h}^{\star\mathsf{H}}\boldsymbol{h}_1 - \boldsymbol{h}^{\star\mathsf{H}}\boldsymbol{h}_2\right| \\ &\lesssim |\alpha_2|\left|\|\boldsymbol{x}_1\|_2^2 - \|\boldsymbol{x}_2\|_2^2\right| + \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2 + \frac{1}{|\alpha_2|^3}\left|\|\boldsymbol{h}_1\|_2^2 - \|\boldsymbol{h}_2\|_2^2\right| + \frac{1}{|\alpha_2|^2}\|\boldsymbol{h}_1 - \boldsymbol{h}_2\|_2, \end{aligned}$$

where the last line follows from the triangle inequality. It is straightforward to see that

$$1/2 \leq |\alpha_2| \leq 2, \quad \left|\|\boldsymbol{x}_1\|_2^2 - \|\boldsymbol{x}_2\|_2^2\right| \lesssim \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2, \quad \left|\|\boldsymbol{h}_1\|_2^2 - \|\boldsymbol{h}_2\|_2^2\right| \lesssim \|\boldsymbol{h}_1 - \boldsymbol{h}_2\|_2$$

under condition (291) and assumption $\|\boldsymbol{x}^\star\|_2 = \|\boldsymbol{h}^\star\|_2 = 1$, where the first inequality follows from Lemma 52. Taking these estimates together reveals that

$$\left\|\nabla g_{\boldsymbol{h}_1,\boldsymbol{x}_1}(\alpha_2) - \nabla g_{\boldsymbol{h}_2,\boldsymbol{x}_2}(\alpha_2)\right\|_2 \lesssim \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2 + \|\boldsymbol{h}_1 - \boldsymbol{h}_2\|_2 .$$

The proof is accomplished by substituting the two bounds on the gradient and the Hessian into (292). □

Further, if two vector pairs are both close to the optimizer, then their distance after alignment (w.r.t. the optimizer) cannot be much larger than their distance without alignment, as revealed by the following lemma.

Lemma 55 *Suppose that the vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{h}_1, \boldsymbol{h}_2 \in \mathbb{C}^K$ satisfy*

$$\max\left\{\left\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\right\|_2, \left\|\boldsymbol{h}_1 - \boldsymbol{h}^\star\right\|_2, \left\|\boldsymbol{x}_2 - \boldsymbol{x}^\star\right\|_2, \left\|\boldsymbol{h}_2 - \boldsymbol{h}^\star\right\|_2\right\} \le \delta \le 1/4 \tag{293}$$

for some sufficiently small constant $\delta > 0$. Denote by α_1 and α_2 the minimizers of $g_{\boldsymbol{h}_1,\boldsymbol{x}_1}(\alpha)$ and $g_{\boldsymbol{h}_2,\boldsymbol{x}_2}(\alpha)$, respectively. Then we have

$$\|\alpha_1\boldsymbol{x}_1 - \alpha_2\boldsymbol{x}_2\|_2^2 + \left\|\frac{1}{\overline{\alpha_1}}\boldsymbol{h}_1 - \frac{1}{\overline{\alpha_2}}\boldsymbol{h}_2\right\|_2^2 \lesssim \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2^2 + \|\boldsymbol{h}_1 - \boldsymbol{h}_2\|_2^2 .$$

Proof To start with, we control the magnitudes of α_1 and α_2. Lemma 52 together with assumption (293) guarantees that

$$1/2 \le |\alpha_1| \le 2 \qquad \text{and} \qquad 1/2 \le |\alpha_2| \le 2.$$

Now we can prove the lemma. The triangle inequality gives

$$\begin{aligned}
\|\alpha_1\boldsymbol{x}_1 - \alpha_2\boldsymbol{x}_2\|_2 &= \|\alpha_1(\boldsymbol{x}_1 - \boldsymbol{x}_2) + (\alpha_1 - \alpha_2)\boldsymbol{x}_2\|_2 \\
&\le |\alpha_1|\,\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2 + |\alpha_1 - \alpha_2|\,\|\boldsymbol{x}_2\|_2 \\
&\overset{\text{(i)}}{\le} 2\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2 + 2|\alpha_1 - \alpha_2| \\
&\overset{\text{(ii)}}{\lesssim} \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2 + \|\boldsymbol{h}_1 - \boldsymbol{h}_2\|_2 ,
\end{aligned}$$

where (i) holds since $|\alpha_1| \le 2$ and $\|\boldsymbol{x}_2\|_2 \le 1 + \delta \le 2$, and (ii) arises from Lemma 54 that $|\alpha_1 - \alpha_2| \lesssim \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2 + \|\boldsymbol{h}_1 - \boldsymbol{h}_2\|_2$. Similarly,

$$\begin{aligned}
\left\|\frac{1}{\alpha_1}\boldsymbol{h}_1 - \frac{1}{\alpha_2}\boldsymbol{h}_2\right\|_2 &= \left\|\frac{1}{\alpha_1}(\boldsymbol{h}_1 - \boldsymbol{h}_2) + \left(\frac{1}{\alpha_1} - \frac{1}{\alpha_2}\right)\boldsymbol{h}_2\right\|_2 \\
&\le \left|\frac{1}{\alpha_1}\right|\|\boldsymbol{h}_1 - \boldsymbol{h}_2\|_2 + \left|\frac{1}{\alpha_1} - \frac{1}{\alpha_2}\right|\|\boldsymbol{h}_2\|_2 \\
&\le 2\|\boldsymbol{h}_1 - \boldsymbol{h}_2\|_2 + 2\frac{|\alpha_1 - \alpha_2|}{|\alpha_1\alpha_2|} \\
&\lesssim \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2 + \|\boldsymbol{h}_1 - \boldsymbol{h}_2\|_2 ,
\end{aligned}$$

where the last inequality comes from Lemma 54 as well as the facts that $|\alpha_1| \geq 1/2$ and $|\alpha_2| \geq 1/2$ as shown above. Combining all of the above bounds and recognizing that $\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2 + \|\boldsymbol{h}_1 - \boldsymbol{h}_2\|_2 \leq \sqrt{2\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2^2 + 2\|\boldsymbol{h}_1 - \boldsymbol{h}_2\|_2^2}$, we conclude the proof. □

Finally, there is a useful identity associated with the minimizer of $\widetilde{g}(\alpha)$ as defined below.

Lemma 56 *For any $\boldsymbol{h}_1, \boldsymbol{h}_2, \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{C}^K$, denote*

$$\alpha^\sharp := \arg\min_{\alpha} \widetilde{g}(\alpha), \quad \textit{where} \quad \widetilde{g}(\alpha) := \left\| \frac{1}{\overline{\alpha}} \boldsymbol{h}_1 - \boldsymbol{h}_2 \right\|_2^2 + \|\alpha \boldsymbol{x}_1 - \boldsymbol{x}_2\|_2^2 .$$

Let $\widetilde{\boldsymbol{x}}_1 = \alpha^\sharp \boldsymbol{x}_1$ and $\widetilde{\boldsymbol{h}}_1 = \frac{1}{\overline{\alpha^\sharp}} \boldsymbol{h}_1$, then we have

$$\left\|\widetilde{\boldsymbol{x}}_1 - \boldsymbol{x}_2\right\|_2^2 + \boldsymbol{x}_2^{\mathsf{H}} (\widetilde{\boldsymbol{x}}_1 - \boldsymbol{x}_2) = \left\|\widetilde{\boldsymbol{h}}_1 - \boldsymbol{h}_2\right\|_2^2 + \left(\widetilde{\boldsymbol{h}}_1 - \boldsymbol{h}_2\right)^{\mathsf{H}} \boldsymbol{h}_2.$$

Proof We can rewrite the function $\widetilde{g}(\alpha)$ as

$$\begin{aligned}
\widetilde{g}(\alpha) &= |\alpha|^2 \|\boldsymbol{x}_1\|_2^2 + \|\boldsymbol{x}_2\|_2^2 - (\alpha \boldsymbol{x}_1)^{\mathsf{H}} \boldsymbol{x}_2 - \boldsymbol{x}_2^{\mathsf{H}} (\alpha \boldsymbol{x}_1) + \left|\frac{1}{\overline{\alpha}}\right|^2 \|\boldsymbol{h}_1\|_2^2 + \|\boldsymbol{h}_2\|_2^2 \\
&\quad - \left(\frac{1}{\overline{\alpha}} \boldsymbol{h}_1\right)^{\mathsf{H}} \boldsymbol{h}_2 - \boldsymbol{h}_2^{\mathsf{H}} \left(\frac{1}{\overline{\alpha}} \boldsymbol{h}_1\right) \\
&= \overline{\alpha}\alpha \|\boldsymbol{x}_1\|_2^2 + \|\boldsymbol{x}_2\|_2^2 - \overline{\alpha} \boldsymbol{x}_1^{\mathsf{H}} \boldsymbol{x}_2 - \alpha \boldsymbol{x}_2^{\mathsf{H}} \boldsymbol{x}_1 + \frac{1}{\overline{\alpha}\alpha} \|\boldsymbol{h}_1\|_2^2 + \|\boldsymbol{h}_2\|_2^2 \\
&\quad - \frac{1}{\alpha} \boldsymbol{h}_1^{\mathsf{H}} \boldsymbol{h}_2 - \frac{1}{\overline{\alpha}} \boldsymbol{h}_2^{\mathsf{H}} \boldsymbol{h}_1.
\end{aligned}$$

The first-order optimality condition [69, equation (38)] requires

$$\left.\frac{\partial \widetilde{g}}{\partial \overline{\alpha}}\right|_{\alpha=\alpha^\sharp} = \alpha^\sharp \|\boldsymbol{x}_1\|_2^2 - \boldsymbol{x}_1^{\mathsf{H}} \boldsymbol{x}_2 + \frac{1}{\alpha^\sharp} \left(-\frac{1}{\overline{\alpha^\sharp}^2}\right) \|\boldsymbol{h}_1\|_2^2 - \left(-\frac{1}{\overline{\alpha^\sharp}^2}\right) \boldsymbol{h}_2^{\mathsf{H}} \boldsymbol{h}_1 = 0,$$

which further simplifies to

$$\|\widetilde{\boldsymbol{x}}_1\|_2^2 - \widetilde{\boldsymbol{x}}_1^{\mathsf{H}} \boldsymbol{x}_2 = \left\|\widetilde{\boldsymbol{h}}_1\right\|_2^2 - \boldsymbol{h}_2^{\mathsf{H}} \widetilde{\boldsymbol{h}}_1$$

since $\widetilde{\boldsymbol{x}}_1 = \alpha^\sharp \boldsymbol{x}_1$, $\widetilde{\boldsymbol{h}}_1 = \frac{1}{\overline{\alpha^\sharp}} \boldsymbol{h}_1$, and $\alpha^\sharp \neq 0$ (otherwise $\widetilde{g}(\alpha^\sharp) = \infty$ and cannot be the minimizer). Furthermore, this condition is equivalent to

$$\widetilde{\boldsymbol{x}}_1^{\mathsf{H}} (\widetilde{\boldsymbol{x}}_1 - \boldsymbol{x}_2) = \left(\widetilde{\boldsymbol{h}}_1 - \boldsymbol{h}_2\right)^{\mathsf{H}} \widetilde{\boldsymbol{h}}_1.$$

Recognizing that

$$\widetilde{\boldsymbol{x}}_1^{\mathsf{H}} (\widetilde{\boldsymbol{x}}_1 - \boldsymbol{x}_2) = \boldsymbol{x}_2^{\mathsf{H}} (\widetilde{\boldsymbol{x}}_1 - \boldsymbol{x}_2) + \left(\widetilde{\boldsymbol{x}}_1 - \boldsymbol{x}_2\right)^{\mathsf{H}} (\widetilde{\boldsymbol{x}}_1 - \boldsymbol{x}_2) = \boldsymbol{x}_2^{\mathsf{H}} (\widetilde{\boldsymbol{x}}_1 - \boldsymbol{x}_2) + \|\widetilde{\boldsymbol{x}}_1 - \boldsymbol{x}_2\|_2^2,$$

$$\widetilde{\boldsymbol{h}}_1^{\mathsf{H}}\big(\widetilde{\boldsymbol{h}}_1-\boldsymbol{h}_2\big)=\boldsymbol{h}_2^{\mathsf{H}}\big(\widetilde{\boldsymbol{h}}_1-\boldsymbol{h}_2\big)+\big(\widetilde{\boldsymbol{h}}_1-\boldsymbol{h}_2\big)^{\mathsf{H}}\big(\widetilde{\boldsymbol{h}}_1-\boldsymbol{h}_2\big)=\boldsymbol{h}_2^{\mathsf{H}}\big(\widetilde{\boldsymbol{h}}_1-\boldsymbol{h}_2\big)+\|\widetilde{\boldsymbol{h}}_1-\boldsymbol{h}_2\|_2^2,$$

we arrive at the desired identity. □

D.3.4 Matrix Concentration Inequalities

The proof for blind deconvolution is largely built upon the concentration of random matrices that are functions of $\left\{\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\right\}$. In this subsection, we collect the measure concentration results for various forms of random matrices that we encounter in the analysis.

Lemma 57 *Suppose* $\boldsymbol{a}_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\boldsymbol{0},\frac{1}{2}\boldsymbol{I}_K\right)+i\mathcal{N}\left(\boldsymbol{0},\frac{1}{2}\boldsymbol{I}_K\right)$ *for every* $1\le j\le m$, *and* $\{c_j\}_{1\le j\le m}$ *are a set of fixed numbers. Then there exist some universal constants* $\widetilde{C}_1,\widetilde{C}_2>0$ *such that for all* $t\ge 0$

$$\mathbb{P}\left(\left\|\sum_{j=1}^m c_j(\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}-\boldsymbol{I}_K)\right\|\ge t\right)\le 2\exp\left(\widetilde{C}_1K-\widetilde{C}_2\min\left\{\frac{t}{\max_j|c_j|},\frac{t^2}{\sum_{j=1}^m c_j^2}\right\}\right).$$

Proof This is a simple variant of [116, Theorem 5.39], which uses the Bernstein inequality and the standard covering argument. Hence, we omit its proof. □

Lemma 58 *Suppose* $\boldsymbol{a}_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\boldsymbol{0},\frac{1}{2}\boldsymbol{I}_K\right)+i\mathcal{N}\left(\boldsymbol{0},\frac{1}{2}\boldsymbol{I}_K\right)$ *for every* $1\le j\le m$. *Then there exist some absolute constants* $\widetilde{C}_1,\widetilde{C}_2,\widetilde{C}_3>0$ *such that for all* $\max\{1,3\widetilde{C}_1K/\widetilde{C}_2\}/m\le\varepsilon\le 1$, *one has*

$$\mathbb{P}\left(\sup_{|J|\le\varepsilon m}\left\|\sum_{j\in J}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\right\|\ge 4\widetilde{C}_3\varepsilon m\log\frac{e}{\varepsilon}\right)\le 2\exp\left(-\frac{\widetilde{C}_2\widetilde{C}_3}{3}\varepsilon m\log\frac{e}{\varepsilon}\right),$$

where $J\subseteq[m]$ *and* $|J|$ *denotes its cardinality.*

Proof The proof relies on Lemma 57 and the union bound. First, invoke Lemma 57 to see that for any fixed $J\subseteq[m]$ and for all $t\ge 0$, we have

$$\mathbb{P}\left(\left\|\sum_{j\in J}(\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}-\boldsymbol{I}_K)\right\|\ge|J|\,t\right)\le 2\exp\left(\widetilde{C}_1K-\widetilde{C}_2|J|\min\left\{t,t^2\right\}\right),\qquad(294)$$

for some constants $\widetilde{C}_1,\widetilde{C}_2>0$, and as a result,

$$\begin{aligned}&\mathbb{P}\left(\sup_{|J|\le\varepsilon m}\left\|\sum_{j\in J}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\right\|\ge\lceil\varepsilon m\rceil(1+t)\right)\\&\overset{\text{(i)}}{\le}\mathbb{P}\left(\sup_{|J|=\lceil\varepsilon m\rceil}\left\|\sum_{j\in J}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\right\|\ge\lceil\varepsilon m\rceil(1+t)\right)\end{aligned}$$

$$\leq \mathbb{P}\left(\sup_{|J|=\lceil \varepsilon m\rceil}\left\|\sum_{j\in J}(\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}-\boldsymbol{I}_K)\right\|\geq \lceil \varepsilon m\rceil t\right)$$
$$\overset{\text{(ii)}}{\leq}\binom{m}{\lceil \varepsilon m\rceil}\cdot 2\exp\left(\widetilde{C}_1K-\widetilde{C}_2\lceil \varepsilon m\rceil\min\left\{t,t^2\right\}\right),$$

where $\lceil c\rceil$ denotes the smallest integer that is no smaller than c. Here, (i) holds since we take the supremum over a larger set and (ii) results from (294) and the union bound. Apply the elementary inequality $\binom{n}{k}\leq (en/k)^k$ for any $0\leq k\leq n$ to obtain

$$\begin{aligned}
&\mathbb{P}\left(\sup_{|J|\leq \varepsilon m}\left\|\sum_{j\in J}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\right\|\geq \lceil \varepsilon m\rceil(1+t)\right)\\
&\leq 2\left(\frac{em}{\lceil \varepsilon m\rceil}\right)^{\lceil \varepsilon m\rceil}\exp\left(\widetilde{C}_1K-\widetilde{C}_2\lceil \varepsilon m\rceil\min\left\{t,t^2\right\}\right)\\
&\leq 2\left(\frac{e}{\varepsilon}\right)^{2\varepsilon m}\exp\left(\widetilde{C}_1K-\widetilde{C}_2\varepsilon m\min\left\{t,t^2\right\}\right)\\
&=2\exp\left[\widetilde{C}_1K-\varepsilon m\left(\widetilde{C}_2\min\left\{t,t^2\right\}-2\log(e/\varepsilon)\right)\right],
\end{aligned}\tag{295}$$

where the second inequality uses $\varepsilon m\leq \lceil \varepsilon m\rceil\leq 2\varepsilon m$ whenever $1/m\leq \varepsilon\leq 1$.

The proof is then completed by taking $\widetilde{C}_3\geq \max\{1,6/\widetilde{C}_2\}$ and $t=\widetilde{C}_3\log(e/\varepsilon)$. To see this, it is easy to check that $\min\{t,t^2\}=t$ since $t\geq 1$. In addition, one has $\widetilde{C}_1K\leq \widetilde{C}_2\varepsilon m/3\leq \widetilde{C}_2\varepsilon mt/3$, and $2\log(e/\varepsilon)\leq \widetilde{C}_2t/3$. Combine the estimates above with (295) to arrive at

$$\begin{aligned}
&\mathbb{P}\left(\sup_{|J|\leq \varepsilon m}\left\|\sum_{j\in J}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\right\|\geq 4\widetilde{C}_3\varepsilon m\log(e/\varepsilon)\right)\\
&\overset{\text{(i)}}{\leq}\mathbb{P}\left(\sup_{|J|\leq \varepsilon m}\left\|\sum_{j\in J}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\right\|\geq \lceil \varepsilon m\rceil(1+t)\right)\\
&\leq 2\exp\left[\widetilde{C}_1K-\varepsilon m\left(\widetilde{C}_2\min\left\{t,t^2\right\}-2\log(e/\varepsilon)\right)\right]\\
&\overset{\text{(ii)}}{\leq} 2\exp\left(-\varepsilon m\widetilde{C}_2t/3\right)=2\exp\left(-\frac{\widetilde{C}_2\widetilde{C}_3}{3}\varepsilon m\log(e/\varepsilon)\right)
\end{aligned}$$

as claimed. Here, (i) holds due to the facts that $\lceil \varepsilon m\rceil\leq 2\varepsilon m$ and $1+t\leq 2t\leq 2\widetilde{C}_3\log(e/\varepsilon)$. Inequality (ii) arises from the estimates listed above. □

Lemma 59 *Suppose $m\gg K\log^3 m$. With probability exceeding $1-O\left(m^{-10}\right)$, we have*

$$\left\|\sum_{j=1}^{m}\left|\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}^{\star}\right|^2\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}-\boldsymbol{I}_K\right\|\lesssim\sqrt{\frac{K}{m}\log m}.$$

Proof The identity $\sum_{j=1}^m \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} = \boldsymbol{I}_K$ allows us to rewrite the quantity on the left-hand side as

$$\left\| \sum_{j=1}^m \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} - \boldsymbol{I}_K \right\| = \left\| \sum_{j=1}^m \underbrace{\left(\left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 - 1 \right) \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}}}_{:=\boldsymbol{Z}_j} \right\|,$$

where the $\boldsymbol{Z}_j$'s are independent zero-mean random matrices. To control the above spectral norm, we resort to the matrix Bernstein inequality [66, Theorem 2.7]. To this end, we first need to upper bound the sub-exponential norm $\|\cdot\|_{\psi_1}$ (see definition in [116]) of each summand $\boldsymbol{Z}_j$, i.e.,

$$\left\| \|\boldsymbol{Z}_j\| \right\|_{\psi_1} = \|\boldsymbol{b}_j\|_2^2 \left\| \left| \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 - 1 \right| \right\|_{\psi_1} \lesssim \|\boldsymbol{b}_j\|_2^2 \left\| \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 \right\|_{\psi_1} \lesssim \frac{K}{m},$$

where we make use of the facts that

$$\|\boldsymbol{b}_j\|_2^2 = K/m \qquad \text{and} \qquad \left\| \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 \right\|_{\psi_1} \lesssim 1.$$

We further need to bound the variance parameter, that is,

$$\begin{aligned} \sigma_0^2 &:= \left\| \mathbb{E}\left[\sum_{j=1}^m \boldsymbol{Z}_j \boldsymbol{Z}_j^{\mathsf{H}} \right] \right\| = \left\| \mathbb{E}\left[\sum_{j=1}^m \left(\left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 - 1 \right)^2 \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \right] \right\| \\ &\lesssim \left\| \sum_{j=1}^m \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \right\| = \frac{K}{m} \left\| \sum_{j=1}^m \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \right\| = \frac{K}{m}, \end{aligned}$$

where the second line arises since $\mathbb{E}\left[\left(|\boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star|^2 - 1 \right)^2 \right] \asymp 1$, $\|\boldsymbol{b}_j\|_2^2 = K/m$, and $\sum_{j=1}^m \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} = \boldsymbol{I}_K$. A direct application of the matrix Bernstein inequality [66, Theorem 2.7] leads us to conclude that with probability exceeding $1 - O\left(m^{-10}\right)$,

$$\left\| \sum_{j=1}^m \boldsymbol{Z}_j \right\| \lesssim \max\left\{ \sqrt{\frac{K}{m} \log m}, \frac{K}{m} \log^2 m \right\} \asymp \sqrt{\frac{K}{m} \log m},$$

where the last relation holds under the assumption that $m \gg K \log^3 m$. □

D.3.5 Matrix Perturbation Bounds

We also need the following perturbation bound on the top singular vectors of a given matrix. The following lemma is parallel to Lemma 34.

Lemma 60 *Let $\sigma_1(\boldsymbol{A})$, $\boldsymbol{u}$, and $\boldsymbol{v}$ be the leading singular value, left and right singular vectors of $\boldsymbol{A}$, respectively, and let $\sigma_1(\widetilde{\boldsymbol{A}})$, $\widetilde{\boldsymbol{u}}$, and $\widetilde{\boldsymbol{v}}$ be the leading singular value, left and right singular vectors of $\widetilde{\boldsymbol{A}}$, respectively. Suppose $\sigma_1(\boldsymbol{A})$ and $\sigma_1(\widetilde{\boldsymbol{A}})$ are not identically zero, and then one has*

$$\begin{aligned}
&\left|\sigma_1(\boldsymbol{A}) - \sigma_1(\widetilde{\boldsymbol{A}})\right| \le \left\|\left(\boldsymbol{A} - \widetilde{\boldsymbol{A}}\right)\boldsymbol{v}\right\|_2 + \left(\|\boldsymbol{u} - \widetilde{\boldsymbol{u}}\|_2 + \|\boldsymbol{v} - \widetilde{\boldsymbol{v}}\|_2\right)\left\|\widetilde{\boldsymbol{A}}\right\|; \\
&\left\|\sqrt{\sigma_1(\boldsymbol{A})}\,\boldsymbol{u} - \sqrt{\sigma_1(\widetilde{\boldsymbol{A}})}\,\widetilde{\boldsymbol{u}}\right\|_2 + \left\|\sqrt{\sigma_1(\boldsymbol{A})}\,\boldsymbol{v} - \sqrt{\sigma_1(\widetilde{\boldsymbol{A}})}\,\widetilde{\boldsymbol{v}}\right\|_2 \\
&\quad \le \sqrt{\sigma_1(\boldsymbol{A})}\left(\|\boldsymbol{u} - \widetilde{\boldsymbol{u}}\|_2 + \|\boldsymbol{v} - \widetilde{\boldsymbol{v}}\|_2\right) + \frac{2\left|\sigma_1(\boldsymbol{A}) - \sigma_1(\widetilde{\boldsymbol{A}})\right|}{\sqrt{\sigma_1(\boldsymbol{A})} + \sqrt{\sigma_1(\widetilde{\boldsymbol{A}})}}.
\end{aligned}$$

Proof The first claim follows since

$$\begin{aligned}
\left|\sigma_1(\boldsymbol{A}) - \sigma_1(\widetilde{\boldsymbol{A}})\right| &= \left|\boldsymbol{u}^{\mathsf{H}}\boldsymbol{A}\boldsymbol{v} - \widetilde{\boldsymbol{u}}^{\mathsf{H}}\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{v}}\right| \\
&\le \left|\boldsymbol{u}^{\mathsf{H}}\left(\boldsymbol{A} - \widetilde{\boldsymbol{A}}\right)\boldsymbol{v}\right| + \left|\boldsymbol{u}^{\mathsf{H}}\widetilde{\boldsymbol{A}}\boldsymbol{v} - \widetilde{\boldsymbol{u}}^{\mathsf{H}}\widetilde{\boldsymbol{A}}\boldsymbol{v}\right| + \left|\widetilde{\boldsymbol{u}}^{\mathsf{H}}\widetilde{\boldsymbol{A}}\boldsymbol{v} - \widetilde{\boldsymbol{u}}^{\mathsf{H}}\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{v}}\right| \\
&\le \left\|\left(\boldsymbol{A} - \widetilde{\boldsymbol{A}}\right)\boldsymbol{v}\right\|_2 + \|\boldsymbol{u} - \widetilde{\boldsymbol{u}}\|_2\left\|\widetilde{\boldsymbol{A}}\right\| + \left\|\widetilde{\boldsymbol{A}}\right\|\|\boldsymbol{v} - \widetilde{\boldsymbol{v}}\|_2.
\end{aligned}$$

With regard to the second claim, we see that

$$\begin{aligned}
\left\|\sqrt{\sigma_1(\boldsymbol{A})}\,\boldsymbol{u} - \sqrt{\sigma_1(\widetilde{\boldsymbol{A}})}\,\widetilde{\boldsymbol{u}}\right\|_2 &\le \left\|\sqrt{\sigma_1(\boldsymbol{A})}\,\boldsymbol{u} - \sqrt{\sigma_1(\boldsymbol{A})}\,\widetilde{\boldsymbol{u}}\right\|_2 + \left\|\sqrt{\sigma_1(\boldsymbol{A})}\,\widetilde{\boldsymbol{u}} - \sqrt{\sigma_1(\widetilde{\boldsymbol{A}})}\,\widetilde{\boldsymbol{u}}\right\|_2 \\
&= \sqrt{\sigma_1(\boldsymbol{A})}\,\|\boldsymbol{u} - \widetilde{\boldsymbol{u}}\|_2 + \left|\sqrt{\sigma_1(\boldsymbol{A})} - \sqrt{\sigma_1(\widetilde{\boldsymbol{A}})}\right| \\
&= \sqrt{\sigma_1(\boldsymbol{A})}\,\|\boldsymbol{u} - \widetilde{\boldsymbol{u}}\|_2 + \frac{\left|\sigma_1(\boldsymbol{A}) - \sigma_1(\widetilde{\boldsymbol{A}})\right|}{\sqrt{\sigma_1(\boldsymbol{A})} + \sqrt{\sigma_1(\widetilde{\boldsymbol{A}})}}.
\end{aligned}$$

Similarly, one can obtain

$$\left\|\sqrt{\sigma_1(\boldsymbol{A})}\,\boldsymbol{v} - \sqrt{\sigma_1(\widetilde{\boldsymbol{A}})}\,\widetilde{\boldsymbol{v}}\right\|_2 \le \sqrt{\sigma_1(\boldsymbol{A})}\,\|\boldsymbol{v} - \widetilde{\boldsymbol{v}}\|_2 + \frac{\left|\sigma_1(\boldsymbol{A}) - \sigma_1(\widetilde{\boldsymbol{A}})\right|}{\sqrt{\sigma_1(\boldsymbol{A})} + \sqrt{\sigma_1(\widetilde{\boldsymbol{A}})}}.$$

Add these two inequalities to complete the proof. □

References

1. Abbe, E., Fan, J., Wang, K., Zhong, Y.: Entrywise eigenvector analysis of random matrices with low expected rank. arXiv preprint arXiv:1709.09565 (2017)
2. Aghasi, A., Ahmed, A., Hand, P., Joshi, B.: Branchhull: Convex bilinear inversion from the entrywise product of signals with known signs. Applied and Computational Harmonic Analysis (2019)
3. Ahmed, A., Recht, B., Romberg, J.: Blind deconvolution using convex programming. IEEE Transactions on Information Theory **60**(3), 1711–1732 (2014)
4. Alon, N., Spencer, J.H.: The Probabilistic Method (3rd Edition). Wiley (2008)
5. Bahmani, S., Romberg, J.: Phase retrieval meets statistical learning theory: A flexible convex relaxation. In: Artificial Intelligence and Statistics, pp. 252–260 (2017)
6. Bendory, T., Eldar, Y.C., Boumal, N.: Non-convex phase retrieval from STFT measurements. IEEE Transactions on Information Theory (2017)

7. Bhojanapalli, S., Neyshabur, B., Srebro, N.: Global optimality of local search for low rank matrix recovery. In: Advances in Neural Information Processing Systems, pp. 3873–3881 (2016)
8. Bousquet, O., Elisseeff, A.: Stability and generalization. Journal of Machine Learning Research **2**(Mar), 499–526 (2002)
9. Bubeck, S.: Convex optimization: Algorithms and complexity. Foundations and Trends® in Machine Learning **8**(3-4), 231–357 (2015)
10. Cai, J.F., Liu, H., Wang, Y.: Fast rank-one alternating minimization algorithm for phase retrieval. Journal of Scientific Computing **79**(1), 128–147 (2019)
11. Cai, T., Zhang, A.: ROP: Matrix recovery via rank-one projections. The Annals of Statistics **43**(1), 102–138 (2015)
12. Cai, T.T., Li, X., Ma, Z.: Optimal rates of convergence for noisy sparse phase retrieval via thresholded Wirtinger flow. The Annals of Statistics **44**(5), 2221–2251 (2016)
13. Candès, E., Plan, Y.: A probabilistic and RIPless theory of compressed sensing. IEEE Transactions on Information Theory **57**(11), 7235–7254 (2011). https://doi.org/10.1109/TIT.2011.2161794
14. Candès, E., Tao, T.: The power of convex relaxation: Near-optimal matrix completion. IEEE Transactions on Information Theory **56**(5), 2053 –2080 (2010)
15. Candès, E.J., Eldar, Y.C., Strohmer, T., Voroninski, V.: Phase retrieval via matrix completion. SIAM Journal on Imaging Sciences **6**(1), 199–225 (2013)
16. Candès, E.J., Li, X.: Solving quadratic equations via PhaseLift when there are about as many equations as unknowns. Foundations of Computational Mathematics **14**(5), 1017–1026 (2014)
17. Candès, E.J., Li, X., Ma, Y., Wright, J.: Robust principal component analysis? Journal of ACM **58**(3), 11:1–11:37 (2011)
18. Candès, E.J., Li, X., Soltanolkotabi, M.: Phase retrieval via Wirtinger flow: Theory and algorithms. IEEE Transactions on Information Theory **61**(4), 1985–2007 (2015)
19. Candès, E.J., Recht, B.: Exact matrix completion via convex optimization. Foundations of Computational Mathematics **9**(6), 717–772 (2009)
20. Candès, E.J., Strohmer, T., Voroninski, V.: Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. Communications on Pure and Applied Mathematics **66**(8), 1017–1026 (2013)
21. Chandrasekaran, V., Sanghavi, S., Parrilo, P.A., Willsky, A.S.: Rank-sparsity incoherence for matrix decomposition. SIAM Journal on Optimization **21**(2), 572–596 (2011)
22. Chen, P., Fannjiang, A., Liu, G.R.: Phase retrieval with one or two diffraction patterns by alternating projections with the null initialization. Journal of Fourier Analysis and Applications, pp. 1–40 (2015)
23. Chen, Y.: Incoherence-optimal matrix completion. IEEE Transactions on Information Theory **61**(5), 2909–2923 (2015)
24. Chen, Y., Candès, E.: The projected power method: An efficient algorithm for joint alignment from pairwise differences. Communications on Pure and Applied Mathematics **71**(8), 1648–1714 (2018)
25. Chen, Y., Candès, E.J.: Solving random quadratic systems of equations is nearly as easy as solving linear systems. Communications on Pure and Applied Mathematics **70**(5), 822–883 (2017). https://doi.org/10.1002/cpa.21638.
26. Chen, Y., Cheng, C., Fan, J.: Asymmetry helps: Eigenvalue and eigenvector analyses of asymmetrically perturbed low-rank matrices. arXiv preprint arXiv:1811.12804 (2018)
27. Chen, Y., Chi, Y., Fan, J., Ma, C.: Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. Mathematical Programming **176**(1-2), 5–37 (2019)
28. Chen, Y., Chi, Y., Fan, J., Ma, C., Yan, Y.: Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. arXiv preprint arXiv:1902.07698 (2019)
29. Chen, Y., Chi, Y., Goldsmith, A.J.: Exact and stable covariance estimation from quadratic sampling via convex programming. IEEE Transactions on Information Theory **61**(7), 4034–4059 (2015)
30. Chen, Y., Fan, J., Ma, C., Wang, K.: Spectral method and regularized MLE are both optimal for top-K ranking. Annals of Statistics **47**(4), 2204–2235 (2019)
31. Chen, Y., Fan, J., Ma, C., Yan, Y.: Inference and uncertainty quantification for noisy matrix completion. arXiv preprint arXiv:1906.04159 (2019)
32. Chen, Y., Wainwright, M.J.: Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. arXiv preprint arXiv:1509.03025 (2015)
33. Chen, Y., Yi, X., Caramanis, C.: A convex formulation for mixed regression with two components: Minimax optimal rates. In: Conference on Learning Theory, pp. 560–604 (2014)

34. Cherapanamjeri, Y., Jain, P., Netrapalli, P.: Thresholding based outlier robust PCA. In: Conference on Learning Theory, pp. 593–628 (2017)
35. Chi, Y.: Guaranteed blind sparse spikes deconvolution via lifting and convex optimization. IEEE Journal of Selected Topics in Signal Processing **10**(4), 782–794 (2016)
36. Chi, Y., Lu, Y.M.: Kaczmarz method for solving quadratic equations. IEEE Signal Processing Letters **23**(9), 1183–1187 (2016)
37. Chi, Y., Lu, Y.M., Chen, Y.: Nonconvex optimization meets low-rank matrix factorization: An overview. arXiv preprint arXiv:1809.09573 (2018)
38. Davenport, M.A., Romberg, J.: An overview of low-rank matrix recovery from incomplete observations. IEEE Journal of Selected Topics in Signal Processing **10**(4), 608–622 (2016)
39. Davis, C., Kahan, W.M.: The rotation of eigenvectors by a perturbation. iii. SIAM Journal on Numerical Analysis **7**(1), 1–46 (1970)
40. Davis, D., Drusvyatskiy, D., Paquette, C.: The nonsmooth landscape of phase retrieval. arXiv preprint arXiv:1711.03247 (2017)
41. Dhifallah, O., Thrampoulidis, C., Lu, Y.M.: Phase retrieval via linear programming: Fundamental limits and algorithmic improvements. arXiv preprint arXiv:1710.05234 (2017)
42. Dopico, F.M.: A note on sin Θ theorems for singular subspace variations. BIT **40**(2), 395–403 (2000). https://doi.org/10.1023/A:1022303426500.
43. Duchi, J.C., Ruan, F.: Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. Information and Inference (2018)
44. El Karoui, N.: On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. Probability Theory and Related Fields pp. 1–81 (2015)
45. El Karoui, N., Bean, D., Bickel, P.J., Lim, C., Yu, B.: On robust regression with high-dimensional predictors. Proceedings of the National Academy of Sciences **110**(36), 14557–14562 (2013)
46. Fan, J., Ma, C., Zhong, Y.: A selective overview of deep learning. arXiv preprint arXiv:1904.05526 (2019)
47. Gao, B., Xu, Z.: Phase retrieval using Gauss-Newton method. arXiv preprint arXiv:1606.08135 (2016)
48. Ge, R., Lee, J.D., Ma, T.: Matrix completion has no spurious local minimum. In: Advances in Neural Information Processing Systems, pp. 2973–2981 (2016)
49. Ge, R., Ma, T.: On the optimization landscape of tensor decompositions. In: Advances in Neural Information Processing Systems, pp. 3653–3663 (2017)
50. Goldstein, T., Studer, C.: Phasemax: Convex phase retrieval via basis pursuit. IEEE Transactions on Information Theory **64**(4), 2675–2689 (2018)
51. Gross, D.: Recovering low-rank matrices from few coefficients in any basis. IEEE Transactions on Information Theory **57**(3), 1548–1566 (2011)
52. Gunasekar, S., Woodworth, B.E., Bhojanapalli, S., Neyshabur, B., Srebro, N.: Implicit regularization in matrix factorization. In: Advances in Neural Information Processing Systems, pp. 6151–6159 (2017)
53. Hand, P., Voroninski, V.: An elementary proof of convex phase retrieval in the natural parameter space via the linear program phasemax. Communications in Mathematical Sciences **16**(7), 2047–2051 (2018)
54. Hardt, M., Wootters, M.: Fast matrix completion without the condition number. Conference on Learning Theory, pp. 638–678 (2014)
55. Hastie, T., Mazumder, R., Lee, J.D., Zadeh, R.: Matrix completion and low-rank SVD via fast alternating least squares. Journal of Machine Learning Research **16**, 3367–3402 (2015)
56. Higham, N.J.: Estimating the matrix p-norm. Numerische Mathematik **62**(1), 539–555 (1992)
57. Hsu, D., Kakade, S.M., Zhang, T.: A tail inequality for quadratic forms of subgaussian random vectors. Electron. Commun. Probab. **17**, no. 52, 6 (2012). https://doi.org/10.1214/ECP.v17-2079.
58. Huang, W., Hand, P.: Blind deconvolution by a steepest descent algorithm on a quotient manifold. SIAM Journal on Imaging Sciences **11**(4), 2757–2785 (2018)
59. Jaganathan, K., Eldar, Y.C., Hassibi, B.: Phase retrieval: An overview of recent developments. arXiv preprint arXiv:1510.07713 (2015)
60. Jain, P., Netrapalli, P.: Fast exact matrix completion with finite samples. In: Conference on Learning Theory, pp. 1007–1034 (2015)
61. Jain, P., Netrapalli, P., Sanghavi, S.: Low-rank matrix completion using alternating minimization. In: ACM symposium on Theory of computing, pp. 665–674 (2013)

62. Javanmard, A., Montanari, A., et al.: Debiasing the lasso: Optimal sample size for gaussian designs. The Annals of Statistics **46**(6A), 2593–2622 (2018)
63. Jin, C., Kakade, S.M., Netrapalli, P.: Provable efficient online matrix completion via non-convex stochastic gradient descent. In: Advances in Neural Information Processing Systems, pp. 4520–4528 (2016)
64. Keshavan, R.H., Montanari, A., Oh, S.: Matrix completion from a few entries. IEEE Transactions on Information Theory **56**(6), 2980–2998 (2010)
65. Keshavan, R.H., Montanari, A., Oh, S.: Matrix completion from noisy entries. J. Mach. Learn. Res. **11**, 2057–2078 (2010)
66. Koltchinskii, V.: Oracle inequalities in empirical risk minimization and sparse recovery problems, *Lecture Notes in Mathematics*, vol. 2033. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-22147-7
67. Koltchinskii, V., Lounici, K., Tsybakov, A.B.: Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. Ann. Statist. **39**(5), 2302–2329 (2011). https://doi.org/10.1214/11-AOS894.
68. Kolte, R., Özgür, A.: Phase retrieval via incremental truncated Wirtinger flow. arXiv preprint arXiv:1606.03196 (2016)
69. Kreutz-Delgado, K.: The complex gradient operator and the CR-calculus. arXiv preprint arXiv:0906.4835 (2009)
70. Lang, S.: Real and functional analysis. Springer-Verlag, New York, **10**, 11–13 (1993)
71. Lee, K., Bresler, Y.: Admira: Atomic decomposition for minimum rank approximation. IEEE Transactions on Information Theory **56**(9), 4402–4416 (2010)
72. Lee, K., Li, Y., Junge, M., Bresler, Y.: Blind recovery of sparse signals from subsampled convolution. IEEE Transactions on Information Theory **63**(2), 802–821 (2017)
73. Lee, K., Tian, N., Romberg, J.: Fast and guaranteed blind multichannel deconvolution under a bilinear system model. IEEE Transactions on Information Theory **64**(7), 4792–4818 (2018)
74. Lerman, G., Maunu, T.: Fast, robust and non-convex subspace recovery. Information and Inference: A Journal of the IMA **7**(2), 277–336 (2017)
75. Li, Q., Tang, G.: The nonconvex geometry of low-rank matrix optimizations with general objective functions. In: 2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP), pp. 1235–1239. IEEE (2017)
76. Li, X., Ling, S., Strohmer, T., Wei, K.: Rapid, robust, and reliable blind deconvolution via nonconvex optimization. Applied and computational harmonic analysis (2018)
77. Li, X., Wang, Z., Lu, J., Arora, R., Haupt, J., Liu, H., Zhao, T.: Symmetry, saddle points, and global geometry of nonconvex matrix factorization. arXiv preprint arXiv:1612.09296 (2016)
78. Li, Y., Lee, K., Bresler, Y.: Blind gain and phase calibration for low-dimensional or sparse signal sensing via power iteration. In: Sampling Theory and Applications (SampTA), 2017 International Conference on, pp. 119–123. IEEE (2017)
79. Li, Y., Ma, C., Chen, Y., Chi, Y.: Nonconvex matrix factorization from rank-one measurements. arXiv preprint arXiv:1802.06286 (2018)
80. Lin, J., Camoriano, R., Rosasco, L.: Generalization properties and implicit regularization for multiple passes SGM. In: International Conference on Machine Learning, pp. 2340–2348 (2016)
81. Ling, S., Strohmer, T.: Self-calibration and biconvex compressive sensing. Inverse Problems **31**(11), 115002 (2015)
82. Ling, S., Strohmer, T.: Regularized gradient descent: a non-convex recipe for fast joint blind deconvolution and demixing. Information and Inference: A Journal of the IMA **8**(1), 1–49 (2018)
83. Lu, Y.M., Li, G.: Phase transitions of spectral initialization for high-dimensional nonconvex estimation. arXiv preprint arXiv:1702.06435 (2017)
84. Mathias, R.: The spectral norm of a nonnegative matrix. Linear Algebra Appl. **139**, 269–284 (1990). https://doi.org/10.1016/0024-3795(90)90403-Y.
85. Mathias, R.: Perturbation bounds for the polar decomposition. SIAM Journal on Matrix Analysis and Applications **14**(2), 588–597 (1993)
86. Maunu, T., Zhang, T., Lerman, G.: A well-tempered landscape for non-convex robust subspace recovery. Journal of Machine Learning Research **20**(37), 1–59 (2019)
87. Mei, S., Bai, Y., Montanari, A.: The landscape of empirical risk for nonconvex losses. The Annals of Statistics **46**(6A), 2747–2774 (2018)

88. Mondelli, M., Montanari, A.: Fundamental limits of weak recovery with applications to phase retrieval. Foundations of Computational Mathematics, pp. 1–71 (2017)
89. Negahban, S., Wainwright, M.J.: Restricted strong convexity and weighted matrix completion: optimal bounds with noise. J. Mach. Learn. Res. **13**, 1665–1697 (2012)
90. Netrapalli, P., Jain, P., Sanghavi, S.: Phase retrieval using alternating minimization. Advances in Neural Information Processing Systems (NIPS) (2013)
91. Netrapalli, P., Niranjan, U., Sanghavi, S., Anandkumar, A., Jain, P.: Non-convex robust PCA. In: Advances in Neural Information Processing Systems, pp. 1107–1115 (2014)
92. Qing, Q., Zhang, Y., Eldar, Y., Wright, J.: Convolutional phase retrieval via gradient descent. Neural Information Processing Systems (2017)
93. Recht, B.: A simpler approach to matrix completion. Journal of Machine Learning Research **12**(Dec), 3413–3430 (2011)
94. Recht, B., Fazel, M., Parrilo, P.A.: Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. SIAM Review **52**(3), 471–501 (2010)
95. Rudelson, M., Vershynin, R., et al.: Hanson-Wright inequality and sub-Gaussian concentration. Electronic Communications in Probability **18** (2013)
96. Sanghavi, S., Ward, R., White, C.D.: The local convexity of solving systems of quadratic equations. Results in Mathematics **71**(3-4), 569–608 (2017)
97. Schmitt, B.A.: Perturbation bounds for matrix square roots and Pythagorean sums. Linear Algebra Appl. **174**, 215–227 (1992). https://doi.org/10.1016/0024-3795(92)90052-C.
98. Schniter, P., Rangan, S.: Compressive phase retrieval via generalized approximate message passing. IEEE Transactions on Signal Processing **63**(4), 1043–1055 (2015)
99. Schudy, W., Sviridenko, M.: Concentration and moment inequalities for polynomials of independent random variables. In: Symposium on Discrete Algorithms, pp. 437–446. ACM, New York (2012)
100. Shechtman, Y., Beck, A., Eldar, Y.C.: GESPAR: Efficient phase retrieval of sparse signals. IEEE Transactions on Signal Processing **62**(4), 928–938 (2014)
101. Soltanolkotabi, M.: Algorithms and theory for clustering and nonconvex quadratic programming. Ph.D. thesis, Stanford University (2014)
102. Soltanolkotabi, M.: Structured signal recovery from quadratic measurements: Breaking sample complexity barriers via nonconvex optimization. IEEE Transactions on Information Theory **65**(4), 2374–2400 (2019)
103. Soltanolkotabi, M., Javanmard, A., Lee, J.D.: Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. IEEE Transactions on Information Theory **65**(2), 742–769 (2019)
104. Soudry, D., Hoffer, E., Nacson, M.S., Gunasekar, S., Srebro, N.: The implicit bias of gradient descent on separable data. The Journal of Machine Learning Research **19**(1), 2822–2878 (2018)
105. Sun, J., Qu, Q., Wright, J.: A geometric analysis of phase retrieval. In: Information Theory (ISIT), 2016 IEEE International Symposium on, pp. 2379–2383. IEEE (2016)
106. Sun, J., Qu, Q., Wright, J.: Complete dictionary recovery over the sphere i: Overview and the geometric picture. IEEE Transactions on Information Theory **63**(2), 853–884 (2017)
107. Sun, R., Luo, Z.Q.: Guaranteed matrix completion via non-convex factorization. IEEE Transactions on Information Theory **62**(11), 6535–6579 (2016)
108. Sur, P., Chen, Y., Candès, E.J.: The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. arXiv preprint arXiv:1706.01191, accepted to Probability Theory and Related Fields (2017)
109. Tan, Y.S., Vershynin, R.: Phase retrieval via randomized kaczmarz: Theoretical guarantees. Information and Inference: A Journal of the IMA **8**(1), 97–123 (2018)
110. Tanner, J., Wei, K.: Low rank matrix completion by alternating steepest descent methods. Applied and Computational Harmonic Analysis **40**(2), 417–429 (2016)
111. Tao, T.: Topics in Random Matrix Theory. Graduate Studies in Mathematics. American Mathematical Society, Providence, Rhode Island (2012)
112. Ten Berge, J.M.: Orthogonal procrustes rotation for two or more matrices. Psychometrika **42**(2), 267–276 (1977)
113. Tropp, J.A.: Convex recovery of a structured signal from independent random linear measurements. In: Sampling Theory, a Renaissance, pp. 67–101. Springer (2015)
114. Tropp, J.A.: An introduction to matrix concentration inequalities. Found. Trends Mach. Learn. **8**(1-2), 1–230 (2015). https://doi.org/10.1561/2200000048.

115. Tu, S., Boczar, R., Simchowitz, M., Soltanolkotabi, M., Recht, B.: Low-rank solutions of linear matrix equations via procrustes flow. In: International Conference on Machine Learning, pp. 964–973. JMLR. org (2016)
116. Vershynin, R.: Introduction to the non-asymptotic analysis of random matrices. Compressed Sensing, Theory and Applications, pp. 210–268 (2012)
117. Wang, G., Giannakis, G., Saad, Y., Chen, J.: Solving most systems of random quadratic equations. In: Advances in Neural Information Processing Systems, pp. 1867–1877 (2017)
118. Wang, G., Giannakis, G.B., Eldar, Y.C.: Solving systems of random quadratic equations via truncated amplitude flow. IEEE Transactions on Information Theory (2017)
119. Wang, G., Zhang, L., Giannakis, G.B., Akçakaya, M., Chen, J.: Sparse phase retrieval via truncated amplitude flow. IEEE Transactions on Signal Processing **66**(2), 479–491 (2018)
120. Wang, L., Chi, Y.: Blind deconvolution from multiple sparse inputs. IEEE Signal Processing Letters **23**(10), 1384–1388 (2016)
121. Wedin, P.Å.: Perturbation bounds in connection with singular value decomposition. BIT Numerical Mathematics **12**(1), 99–111 (1972)
122. Wei, K.: Solving systems of phaseless equations via Kaczmarz methods: A proof of concept study. Inverse Problems **31**(12), 125008 (2015)
123. Wei, K., Cai, J.F., Chan, T.F., Leung, S.: Guarantees of riemannian optimization for low rank matrix recovery. SIAM Journal on Matrix Analysis and Applications **37**(3), 1198–1222 (2016)
124. Yu, Y., Wang, T., Samworth, R.J.: A useful variant of the Davis-Kahan theorem for statisticians. Biometrika **102**(2), 315–323 (2015). https://doi.org/10.1093/biomet/asv008.
125. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning requires rethinking generalization. International Conference on Learning Representations (2017)
126. Zhang, H., Chi, Y., Liang, Y.: Provable non-convex phase retrieval with outliers: Median truncated Wirtinger flow. In: International conference on machine learning, pp. 1022–1031 (2016)
127. Zhang, H., Zhou, Y., Liang, Y., Chi, Y.: A nonconvex approach for phase retrieval: Reshaped wirtinger flow and incremental algorithms. Journal of Machine Learning Research (2017)
128. Zhang, Y., Lau, Y., Kuo, H.w., Cheung, S., Pasupathy, A., Wright, J.: On the global geometry of sphere-constrained sparse blind deconvolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4894–4902 (2017)
129. Zhao, T., Wang, Z., Liu, H.: A nonconvex optimization framework for low rank matrix estimation. In: Advances in Neural Information Processing Systems, pp. 559–567 (2015)
130. Zheng, Q., Lafferty, J.: A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. In: Advances in Neural Information Processing Systems, pp. 109–117 (2015)
131. Zheng, Q., Lafferty, J.: Convergence analysis for rectangular matrix completion using Burer-Monteiro factorization and gradient descent. arXiv preprint arXiv:1605.07051 (2016)
132. Zhong, K., Song, Z., Jain, P., Bartlett, P.L., Dhillon, I.S.: Recovery guarantees for one-hidden-layer neural networks. In: International Conference on Machine Learning, pp. 4140–4149. JMLR. org (2017)
133. Zhong, Y., Boumal, N.: Near-optimal bounds for phase synchronization. SIAM Journal on Optimization **28**(2), 989–1016 (2018)

Affiliations

Cong Ma[1] · Kaizheng Wang[1] · Yuejie Chi[2] · Yuxin Chen[3]

✉ Yuxin Chen
yuxin.chen@princeton.edu

Cong Ma
congm@princeton.edu

Kaizheng Wang
kaizheng@princeton.edu

Yuejie Chi
yuejiechi@cmu.edu

1 Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA

2 Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA

3 Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA