Convex and Nonconvex Optimization Are Both Minimax-Optimal for Noisy Blind Deconvolution under Random Designs

Yuxin Chen*
Department of Electrical and Computer Engineering, Princeton University
and
Jianqing Fan†, Bingyan Wang‡, Yuling Yan
Department of Operations Research and Financial Engineering,
Princeton University

Abstract

We investigate the effectiveness of convex relaxation and nonconvex optimization in solving bilinear systems of equations under two different designs (i.e. a sort of random Fourier design and Gaussian design). Despite the wide applicability, the theoretical understanding about these two paradigms remains largely inadequate in the presence of random noise. The current paper makes two contributions by demonstrating that: (1) a two-stage nonconvex algorithm attains minimax-optimal accuracy within a logarithmic number of iterations, and (2) convex relaxation also achieves minimax-optimal statistical accuracy vis-à-vis random noise. Both results significantly improve upon the state-of-the-art theoretical guarantees.

Keywords: blind deconvolution, bilinear systems of equations, nonconvex optimization, convex relaxation, leave-one-out analysis

*Y. Chen is supported in part by the AFOSR YIP award FA9550-19-1-0030, by the ONR grant N00014-19-1-2120, by the ARO grants W911NF-20-1-0097 and W911NF-18-1-0303, by the NSF grants CCF-1907661, IIS-1900140, IIS-2100158 and DMS-2014279, and by the Princeton SEAS innovation award.

†J. Fan is supported in part by the ONR grant N00014-19-1-2120 and the NSF grants DMS-1662139, DMS-1712591, DMS-2052926, DMS-2053832, and the NIH grant 2R01-GM072611-15.

‡B. Wang is supported in part by the Gordon Y. S. Wu Fellowship in Engineering.

1 Introduction and motivation

Suppose we are interested in a pair of unknown objects $\boldsymbol{h}^{\star}, \boldsymbol{x}^{\star} \in \mathbb{C}^{K}$ and are given a collection of m nonlinear measurements taking the following form

$$y_j = \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^{\star} \boldsymbol{x}^{\star \mathsf{H}} \boldsymbol{a}_j + \xi_j, \qquad 1 \leq j \leq m. \tag{1}$$

Here, $\boldsymbol{z}^{\mathsf{H}}$ denotes the conjugate transpose of a vector $\boldsymbol{z}$, $\{\xi_j\}$ stands for the additive noise, whereas $\{\boldsymbol{a}_j\}$ and $\{\boldsymbol{b}_j\}$ are design vectors (or sampling vectors). The aim is to faithfully reconstruct both $\boldsymbol{h}^{\star}$ and $\boldsymbol{x}^{\star}$ from the above set of bilinear measurements.[1]

This problem of solving bilinear systems of equations spans multiple domains in science and engineering, including but not limited to astronomy, medical imaging, optics, and communication engineering [Campisi and Egiazarian, 2016, Jefferies and Christou, 1993, Wang and Poor, 1998, Wunder et al., 2015, Tong et al., 1994, Chan and Wong, 1998]. Particularly worth emphasizing is the application of blind deconvolution [Ahmed et al., 2013, Kundur and Hatzinakos, 1996, Ling and Strohmer, 2015, Ma et al., 2018], which involves recovering two unknown signals from their circular convolution. As has been made apparent in the seminal work Ahmed et al. [2013], deconvolving two signals can be reduced to solving bilinear equations, provided that the unknown signals lie within some *a priori* known subspaces; the interested reader is referred to Ahmed et al. [2013] for details. A variety of approaches have since been put forward for blind deconvolution, most notable of which are convex relaxation and nonconvex optimization [Ahmed et al., 2013, Ling and Strohmer, 2017, Li et al., 2019, Ma et al., 2018, Huang and Hand, 2018, Ling and

[1]This formulation is reminiscent of the problem of phase retrieval (or solving quadratic systems of equations). But the two problems turn out to be quite different due to the common assumptions imposed on the design vectors, as we shall elucidate in Section 3.

Strohmer, 2019]. Despite a large body of prior work tackling this problem, however, where these algorithms stand vis-à-vis random noise remains unsettled, which we seek to address in the current paper.

1.1 Convex and nonconvex algorithms

Among various algorithms that have been proposed for blind deconvolution, two paradigms have received much attention: (1) convex relaxation and (2) nonconvex optimization, both of which can be explained rather simply. The starting point for both paradigms is a natural least-squares formulation

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

which is, unfortunately, highly nonconvex due to the bilinear structure of the sampling mechanism. It then boils down to how to guarantee a reliable solution despite the intrinsic nonconvexity.

Convex relaxation. In order to tame nonconvexity, a popular strategy is to lift the problem into higher dimension followed by convex relaxation (namely, representing $\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}$ by a matrix variable $\boldsymbol{Z}$ and then dropping the rank-1 constraint) [Ahmed et al., 2013, Ling and Strohmer, 2015, 2017]. More concretely, we consider the following convex program:[2]

$$\underset{\boldsymbol{Z}\in\mathbb{C}^{K\times K}}{\text{minimize}} \quad g(\boldsymbol{Z}) = \sum_{j=1}^{m}\left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{Z}\boldsymbol{a}_j - y_j\right|^2 + 2\lambda\left\|\boldsymbol{Z}\right\|_*, \tag{3}$$

where $\lambda > 0$ denotes the regularization parameter, and $\|\boldsymbol{Z}\|_*$ is the nuclear norm of $\boldsymbol{Z}$ (i.e. the sum of singular values of $\boldsymbol{Z}$) and is known to be the convex surrogate for the rank

[2]As we shall see shortly, we keep a factor 2 here so as to better connect the convex and nonconvex algorithms; it does not affect our main theoretical guarantees at all.

function. The rationale is rather simple: given that we seek to recover a rank-1 matrix $\boldsymbol{Z}^{\star} = \boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}$, it is common to enforce nuclear norm penalization to encourage the rank-1 structure. In truth, this comes down to solving a nuclear-norm regularized least squares problem in the matrix domain $\mathbb{C}^{K\times K}$.

Nonconvex optimization. Another popular paradigm maintains all iterates in the original vector space (i.e. $\mathbb{C}^{K}$) and attempts solving the above nonconvex formulation or its variants directly. The crucial ingredient is to ensure fast and reliable convergence in spite of nonconvexity. While multiple variants of the nonconvex formulation (2) have been studied in the literature (e.g. Li et al. [2019], Ma et al. [2018], Charisopoulos et al. [2019, 2021], Huang and Hand [2018]), the present paper focuses attention on the following ridge-regularized least-squares problem:

$$\underset{\boldsymbol{h},\boldsymbol{x}\in\mathbb{C}^{K}}{\text{minimize}} \quad f\left(\boldsymbol{h},\boldsymbol{x}\right) = \sum_{j=1}^{m}\left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}\boldsymbol{a}_j - y_j\right|^2 + \lambda\left\|\boldsymbol{h}\right\|_2^2 + \lambda\left\|\boldsymbol{x}\right\|_2^2, \tag{4}$$

with $\lambda > 0$ the regularization parameter. This choice of objective function is crucial to the establishment of our main theorems as can be seen later. Owing to the nonconvexity of (4), one needs to also specify which algorithm to employ in attempt to solve this nonconvex problem. Our focal point is a two-stage optimization algorithm: it starts with a rough initial guess $(\boldsymbol{h}^0, \boldsymbol{x}^0)$ computed by means of a spectral method, followed by Wirtinger gradient descent (GD) that iteratively refines the estimates (to be made precise in (6a)). At the end of each gradient iteration, we further rescale the sizes of the two iterates $\boldsymbol{h}^t$ and $\boldsymbol{x}^t$, so as to ensure that they have identical ℓ_2 norm (see (6b)). In truth, this balancing step helps stabilize the algorithm, while facilitating analysis. The whole algorithm is summarized in Algorithm 1.

Algorithm 1 Nonconvex gradient descent with spectral initialization

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

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

$$\boldsymbol{M} := \sum_{j=1}^{m} y_j \boldsymbol{b}_j \boldsymbol{a}_j^{\mathsf{H}}. \tag{5}$$

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** $t = 0, 1, \ldots, t_0 - 1$ **do**

$$\begin{bmatrix} \boldsymbol{h}^{t+1/2} \\ \boldsymbol{x}^{t+1/2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{h}^{t} \\ \boldsymbol{x}^{t} \end{bmatrix} - \eta \begin{bmatrix} \nabla_{\boldsymbol{h}} f(\boldsymbol{h}^t, \boldsymbol{x}^t) \\ \nabla_{\boldsymbol{x}} f(\boldsymbol{h}^t, \boldsymbol{x}^t) \end{bmatrix}, \tag{6a}$$

$$\begin{bmatrix} \boldsymbol{h}^{t+1} \\ \boldsymbol{x}^{t+1} \end{bmatrix} = \begin{bmatrix} \sqrt{\frac{\|\boldsymbol{x}^{t+1/2}\|_2}{\|\boldsymbol{h}^{t+1/2}\|_2}}\boldsymbol{h}^{t+1/2} \\ \sqrt{\frac{\|\boldsymbol{h}^{t+1/2}\|_2}{\|\boldsymbol{x}^{t+1/2}\|_2}}\boldsymbol{x}^{t+1/2} \end{bmatrix}, \tag{6b}$$

where $\nabla_{\boldsymbol{h}} f(\cdot)$ and $\nabla_{\boldsymbol{x}} f(\cdot)$ represent the Wirtinger gradient (see Li et al. [2019, Section 3.3] and Section A.2.1 of supplementary materials) of $f(\cdot)$ w.r.t. $\boldsymbol{h}$ and $\boldsymbol{x}$, respectively.

1.2 Inadequacy of prior theory

The aforementioned two algorithms have found solid theoretical support under certain randomized sampling mechanisms. Informally, imagine that the $\boldsymbol{a}_j$'s and the $\boldsymbol{b}_j$'s follow standard Gaussian and partial Fourier designs, respectively, and that each noise component ξ_j is a zero-mean sub-Gaussian random variable with variance at most σ^2 (more precise descriptions are deferred to Assumption 1). Table 1 summarizes the performance guarantees established in prior theory.

Table 1: Comparison of our theoretical guarantees of blind deconvolution under Fourier design to prior theory, where we hide all logarithmic factors. Here, the Euclidean estimation error refers to $\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\|_{\mathrm{F}}$ for the convex case and $\|\boldsymbol{h}_{\mathsf{ncvx}}\boldsymbol{x}_{\mathsf{ncvx}}^{\mathsf{H}} - \boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\|_{\mathrm{F}}$ for the nonconvex case, respectively.

	Sample complexity	Algorithm	Euclidean error in the noisy case	Computational complexity
Ahmed et al. [2013]	$\mu^2 K$	convex relaxation	$\sigma\sqrt{Km}$	—
Ling and Strohmer [2017]	$\mu^2 K$	convex relaxation	$\sigma\sqrt{Km}$	—
This paper	$\mu^2 K$	convex relaxation	$\sigma\sqrt{K}$	—
Li et al. [2019]	$\mu^2 K$	nonconvex regularized GD	$\sigma\sqrt{K}$	mK^2
Huang and Hand [2018]	$\mu^2 K$	Riemannian steepest descent	$\sigma\sqrt{K}$	mK^2
Ma et al. [2018]	$\mu^2 K$	nonconvex vanilla GD	—	mK (noiseless)
This paper	$\mu^2 K$	nonconvex GD (with balancing operations)	$\sigma\sqrt{K}$	mK

- Convex relaxation is guaranteed to return an estimate of $\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}$ with an Euclidean estimation error bounded by $\sigma\sqrt{Km}$ (modulo some log factor) [Ahmed et al., 2013, Ling and Strohmer, 2017]. This, however, exceeds the minimax lower bound (to be presented in Theorem 5) by at least a factor of $\sqrt{m}$.

- In comparison, nonconvex algorithms are capable of achieving nearly minimax optimal statistical accuracy, with a computational complexity on the order of mK^2 (up to some log factor) [Li et al., 2019, Huang and Hand, 2018]. Here, the computational complexity encompasses the cost of spectral initialization in Algorithm 1 if implemented by power methods [Golub and Van Loan, 2013]. This computational cost, however, could be an order of K times larger than the cost taken to read the data.

See Table 1 for a more complete summary of existing theoretical results for this scenario.

These prior results, while offering rigorous theoretical underpinnings for the two popular algorithms, lead to several natural questions:

1. (Improving statistical guarantees) *Is the statistical accuracy of convex relaxation inherently suboptimal when coping with random noise?*

2. (Improving computational complexity) *Is it possible to further accelerate the nonconvex algorithm without compromising statistical accuracy?*

The present paper is devoted to addressing these two questions. Informally, we aim to demonstrate that (1) convex relaxation achieves minimax-optimal statistical accuracy in the face of random noise, and (2) nonconvex optimization converges to a nearly minimax-optimal solution in time proportional to that taken to read the data.

1.3 Paper organization and notation

The outline of the paper is as follows. Section 2 gives the formal statement of the model assumptions and presents our main results for two different designs. Section 3 reviews previous literature on blind deconvolution. Section 4 presents numerical experiments that corroborate our theoretical results. We conclude the paper in Section 5 by pointing out several future directions. All the proof details are deferred to the supplementary materials.

Throughout the paper, we shall often use the vector notation $\boldsymbol{y} := [y_1, \cdots, y_m]^\top$ and $\boldsymbol{\xi} := [\xi_1, \cdots, \xi_m]^\top \in \mathbb{C}^m$. For any vector $\boldsymbol{v}$ and any matrix $\boldsymbol{M}$, we denote by $\boldsymbol{v}^{\mathsf{H}}$ and $\boldsymbol{M}^{\mathsf{H}}$ their conjugate transpose, respectively. The notation $\|\boldsymbol{v}\|_2$ represents the ℓ_2 norm of an vector $\boldsymbol{v}$, and we let $\|\boldsymbol{M}\|$, $\|\boldsymbol{M}\|_{\mathrm{F}}$ and $\|\boldsymbol{M}\|_*$ represent the spectral norm, the Frobenius norm and the nuclear norm of $\boldsymbol{M}$, respectively. For a function $f(\boldsymbol{h}, \boldsymbol{x})$, we use $\nabla_{\boldsymbol{h}} f(\boldsymbol{h}, \boldsymbol{x})$ (resp. $\nabla_{\boldsymbol{x}} f(\boldsymbol{h}, \boldsymbol{x})$) to denote its Wirtinger gradient (see Li et al. [2019, Section 3.3] for detailed introduction) of $f(\cdot)$ with respect to $\boldsymbol{h}$ (resp. $\boldsymbol{x}$). Further, we define $\nabla f(\boldsymbol{h}, \boldsymbol{x}) = [\nabla_{\boldsymbol{h}} f(\boldsymbol{h}, \boldsymbol{x})^\top, \nabla_{\boldsymbol{x}} f(\boldsymbol{h}, \boldsymbol{x})^\top]^\top$. For any subspace T, we use $T^\perp$ to denote its orthogonal complement, and $\mathcal{P}_T(\boldsymbol{M})$ the Euclidean projection of a matrix $\boldsymbol{M}$ onto T. Moreover, we adopt $f_1(m, K) \lesssim f_2(m, K)$ or $f_1(m, K) = O(f_2(m, K))$ to indicate that there exists some constant $C_1 > 0$ such that $f_1(m, K) \leq C f_2(m, K)$ holds for all (m, K) that are sufficiently large, and use $f_1(m, K) \gtrsim f_2(m, K)$ to indicate that $f_1(m, K) \geq C_2 f_2(m, K)$ holds for some constant $C > 0$ whenever (m, K) are sufficiently large. The notation $f_1(m, K) \asymp f_2(m, K)$ means that $f_1(m, K) \lesssim f_2(m, K)$ and $f_1(m, K) \gtrsim f_2(m, K)$ hold simultaneously. In our proof, C serves as a universal constant whose value might change from line to line.

2 Main results

In this section, we present our theoretical guarantees for the above two algorithms for two types of random designs commonly studied in the blind deconvolution literature.

2.1 Blind deconvolution under random Fourier designs

Model and assumptions. We start by introducing a sort of random Fourier designs motivated by practical engineering applications (see Ahmed et al. [2013], Li et al. [2019]).

Assumption 1. *Let* $\boldsymbol{A} := [\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_m]^{\mathsf{H}} \in \mathbb{C}^{m \times K}$ *and* $\boldsymbol{B} := [\boldsymbol{b}_1, \boldsymbol{b}_2, \cdots, \boldsymbol{b}_m]^{\mathsf{H}} \in \mathbb{C}^{m \times K}$.

- *The entries of* $\boldsymbol{A}$ *are independently drawn from standard complex Gaussian distributions, namely,* $\boldsymbol{a}_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\mathbf{0}, \frac{1}{2}\boldsymbol{I}_K\right) + i\mathcal{N}\left(\mathbf{0}, \frac{1}{2}\boldsymbol{I}_K\right)$ *with* i *the imaginary unit;*

- *The design matrix* $\boldsymbol{B}$ *consists of the first* K *columns of the unitary discrete Fourier transform (DFT) matrix* $\boldsymbol{F} \in \mathbb{C}^{m\times m}$ *obeying* $\boldsymbol{F}\boldsymbol{F}^{\mathsf{H}} = \boldsymbol{I}_m$*;*

- *The noise components* $\{\xi_i\}$ *are independent zero-mean sub-Gaussian random variables with sub-Gaussian norm obeying* $\|\xi_i\|_{\psi_2} \le \sigma$ $(1 \le i \le m)$*. See Vershynin [2010, Definition 5.7] for the definition of* $\|\cdot\|_{\psi_2}$*.*

Remark 1. It is easy to show that $\|\boldsymbol{b}_j\|_2 = \sqrt{K/m}$ $(1 \le j \le m)$ under this model.

It is worth noting that the Fourier design is largely motivated by the duality relation between convolution in the time domain and multiplication in the frequency domain, which is closely related to practical scenarios; see Ahmed et al. [2013] for details. In fact, the model described in Assumption 1 has been the focus of a number of recent papers including Ahmed et al. [2013], Li et al. [2019], Ma et al. [2018], Huang and Hand [2018], Ling and Strohmer [2019, 2016, 2017], to name a few.

In addition, as pointed out by prior works Ahmed et al. [2013], Li et al. [2019], Ma et al. [2018], the following incoherence condition — which captures the interplay between the truth and the measurement mechanism — plays a crucial role in enabling tractable estimation schemes.

Definition 1 (Incoherence)**.** *Define the incoherence parameter* μ *as the smallest number obeying*

$$\left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\right| \le \frac{\mu}{\sqrt{K}}\left\|\boldsymbol{b}_j\right\|_2\left\|\boldsymbol{h}^\star\right\|_2 = \frac{\mu}{\sqrt{m}}\left\|\boldsymbol{h}^\star\right\|_2, \qquad 1 \le j \le m. \tag{7}$$

Remark 2. Comparing the Cauchy-Schwarz inequality $\left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\right| \leq \|\boldsymbol{b}_j\|_2 \|\boldsymbol{h}^\star\|_2$ with (7) reveals that $\mu \leq \sqrt{K}$. It is noteworthy that our theory does not require μ to be small constant; in fact, all of our theoretical findings allow μ to grow with the problem dimension. Informally, a small incoherence parameter indicates that the truth is not quite aligned with the sampling basis. As a concrete example, when $\boldsymbol{h}^\star$ is randomly generated (i.e. $\boldsymbol{h}^\star \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_K)$), it can be easily verified that the incoherence parameter μ is, with high probability, at most $O(\sqrt{\log m})$. In fact, this type of condition is widely proposed in statistical literature on various problem besides blind deconvolution, such as Candès and Recht [2009], Ma et al. [2018], Chen et al. [2020b] on matrix completion and Candès et al. [2011], Chandrasekaran et al. [2011], Chen et al. [2020c] on robust principal component analysis. The important role of this incoherence parameter will also be confirmed by our numerical simulations momentarily (cf. Figure 3).

Main theory. We are now positioned to state our main theory for this setting, followed by discussing the implications of our theory. Towards this end, we begin with the statistical guarantees for the convex formulation. Denote the minimizer of (3) by $\boldsymbol{Z}_{\mathsf{cvx}}$. Then our result is this:

Theorem 1 (Convex relaxation)**.** *Set $\lambda = C_\lambda \sigma \sqrt{K \log m}$ for some large enough constant $C_\lambda > 0$. Assume*

$$m \geq C\mu^2 K \log^9 m \qquad \textit{and} \qquad \sigma\sqrt{K \log^5 m} \leq c \left\|\boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}} \tag{8}$$

for some sufficiently large (resp. small) constant $C > 0$ (resp. $c > 0$). Then under Assumption 1 and the incoherence condition (7), one has with probability exceeding $1 -$

$O\left(m^{-3}+me^{-K}\right)$ that

$$\left\|\boldsymbol{Z}_{\mathsf{cvx}}-\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\right\|\leq\left\|\boldsymbol{Z}_{\mathsf{cvx}}-\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}}\lesssim\sigma\sqrt{K\log m}. \tag{9}$$

In addition, the bounds in (9) continue to hold if $\boldsymbol{Z}_{\mathsf{cvx}}$ is replaced by

$$\boldsymbol{Z}_{\mathsf{cvx},1}\coloneqq\arg\min_{\boldsymbol{Z}:\mathsf{rank}(\boldsymbol{Z})\leq 1}\left\|\boldsymbol{Z}-\boldsymbol{Z}_{\mathsf{cvx}}\right\|_{\mathrm{F}}$$

i.e. the best rank-1 approximation of $\boldsymbol{Z}_{\mathsf{cvx}}$.

Remark 3. In (8), $\log^9 m$ and $\log^5 m$ appear due to our decoupling arguments. We believe it would be difficult to get rid of the logarithmic factors completely using the current analyis framework, although it might be possible to reduce the power of the logarithmic factors slightly by means of more refined analysis.

Our proof for this theorem, whose details are postponed to Section B.1 of supplementary materials, is largely inspired by the idea of connecting convex and nonconvecx optimization as proposed by Chen et al. [2020b,c] for noisy matrix completion and robust principal component analysis respectively. Note, however, that implementing this high-level idea requires drastically different analysis from Chen et al. [2020b,c], primarily due to the absence of randomness in the highly structured Fourier design matrix $\boldsymbol{B}$. For instance, in contrast to prior works that were built upon a "leave-one-out" analysis framework to decouple statistical dependency, simply "leaving out" one row of $\boldsymbol{B}$ in the blind deconvolution analysis does not lead to immediate statistical benefits due to the deterministic nature of $\boldsymbol{B}$. Consequently, considerably more delicate analyses are needed in order to enable fine-grained statistical analysis.

Next, we turn to theoretical guarantees for the nonconvex algorithm described in

Algorithm 1. For notational convenience, we define

$$\boldsymbol{z}^t := \begin{bmatrix} \boldsymbol{h}^t \\ \boldsymbol{x}^t \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{z}^\star := \begin{bmatrix} \boldsymbol{h}^\star \\ \boldsymbol{x}^\star \end{bmatrix} \tag{10}$$

throughout this paper. Before presenting the results, we make note of an unavoidable scaling ambiguity issue underlying this model. Given that $\boldsymbol{h}^\star$ and $\boldsymbol{x}^\star$ are only identifiable up to global scaling (meaning that one cannot hope to distinguish $(\alpha\boldsymbol{h}^\star, \frac{1}{\overline{\alpha}}\boldsymbol{x}^\star)$ from $(\boldsymbol{h}^\star, \boldsymbol{x}^\star)$ given only bilinear measurements), we shall measure the discrepancy between $\boldsymbol{z}^\star$ and any point $\boldsymbol{z} := \begin{bmatrix} \boldsymbol{h} \\ \boldsymbol{x} \end{bmatrix}$ through the following metric:

$$\mathsf{dist}\left(\boldsymbol{z}, \boldsymbol{z}^\star\right) := \min_{\alpha \in \mathbb{C}} \sqrt{\left\| \frac{1}{\overline{\alpha}}\boldsymbol{h} - \boldsymbol{h}^\star \right\|_2^2 + \left\| \alpha\boldsymbol{x} - \boldsymbol{x}^\star \right\|_2^2}. \tag{11}$$

In words, this metric is an extension of the ℓ_2 distance modulo global scaling. Our result is this:

Theorem 2 (Nonconvex optimization)**.** *Set $\lambda = C_\lambda \sigma \sqrt{K \log m}$ for some large enough constant $C_\lambda > 0$. Take $\eta = c_\eta$ for some sufficiently small constant $c_\eta > 0$. Suppose that Assumption 1, the incoherence condition (7) and the condition (8) hold. Then with probability at least $1 - O\left(m^{-5} + me^{-K}\right)$, the iterates $\{\boldsymbol{h}^t, \boldsymbol{x}^t\}_{0 \le t \le t_0}$ of the spectrally initialized nonconvex algorithm (see Algorithm 1) obey*

$$\mathsf{dist}\left(\boldsymbol{z}^0, \boldsymbol{z}^\star\right) \lesssim \sqrt{\frac{\mu^2 K \log m}{m}} \left\|\boldsymbol{z}^\star\right\|_2 + \frac{\sigma\sqrt{K\log m}}{\left\|\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}}^{1/2}}, \tag{12a}$$

$$\mathsf{dist}\left(\boldsymbol{z}^t, \boldsymbol{z}^\star\right) \le \rho^t \mathsf{dist}\left(\boldsymbol{z}^0, \boldsymbol{z}^\star\right) + \frac{C_1\left(\lambda + \sigma\sqrt{K\log m}\right)}{c_\rho \left\|\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}}^{1/2}}, \tag{12b}$$

$$\left\|\boldsymbol{h}^t\left(\boldsymbol{x}^t\right)^{\mathsf{H}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}} \le 2\rho^t \mathsf{dist}\left(\boldsymbol{z}^0, \boldsymbol{z}^\star\right)\left\|\boldsymbol{z}^\star\right\|_2 + \frac{2C_1\left(\lambda + \sigma\sqrt{K\log m}\right)}{c_\rho}, \tag{12c}$$

simultaneously for all $0 \leq t \leq t_0 \leq m^{20}$*. Here, we take* $C_1 > 0$ *to be some sufficiently large constant and* $0 < \rho = 1 - c_\rho \eta < 1$ *for some sufficiently small constant* $c_\rho > 0$*.*

Remark 4. It is noteworthy that the quantity m^{-5} in the probability term $1 - O\left(m^{-5} + me^{-K}\right)$ in this theorem can actually be replaced by m^{-C} for any positive integer C.

Informally, this theorem guarantees that the estimation error of the iterates $\{\boldsymbol{h}^t, \boldsymbol{x}^t\}_{0 \leq t \leq t_0}$ generated by Algorithm 1 decays geometrically fast until some error floor is hit. As we shall demonstrate momentarily in Theorem 5, this error floor matches the minimax-optimal statistical error up to some logarithmic term.

Compared with one of the most relevant papers to us — Ma et al. [2018] — on blind deconvolution under Fourier designs, this theorem generalizes the noiseless case studied in Ma et al. [2018] to the noisy case. This generalization needs a lot of efforts since it calls for delicate and careful control of the noise effect, as detailed in the proof in Section A of supplementary materials.

2.2 Blind deconvolution under Gaussian designs

In addition to the above-mentioned random Fourier design, our results also extend to the scenario under Gaussian design, as formalized in the assumption below.

Assumption 2.
- *The entries of* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *are independently drawn from standard complex Gaussian distributions, namely,* $\boldsymbol{a}_j, \boldsymbol{b}_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\boldsymbol{0}, \frac{1}{2}\boldsymbol{I}_K\right) + i\mathcal{N}\left(\boldsymbol{0}, \frac{1}{2}\boldsymbol{I}_K\right)$*;*
- *The noise components* $\{\xi_i\}$ *are independent zero-mean sub-Gaussian random variables with sub-Gaussian norm obeying* $\|\xi_i\|_{\psi_2} \leq \sigma$ $(1 \leq i \leq m)$*.*

Akin to Theorems 1 and 2, we consider the loss functions (3) and (4). The main results under the Gaussian design are summarized in the following theorems.

Theorem 3 (Convex relaxation)**.** *Let* $\lambda = C_\lambda \sigma \sqrt{mK \log m}$ *for some sufficiently large constant* $C_\lambda > 0$. *Assume the sample complexity and the noise level satisfy*

$$m \geq CK \log^6 m \qquad \textit{and} \qquad \sigma \sqrt{\frac{K \log^5 m}{m}} \leq c \left\| \boldsymbol{h}^\star \boldsymbol{x}^{\star \mathsf{H}} \right\|_{\mathrm{F}} \tag{13}$$

for some sufficiently large (resp. small) constant $C > 0$ *(resp.* $c > 0$*). Then*

$$\left\| \boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{h}^\star \boldsymbol{x}^{\star \mathsf{H}} \right\| \leq \left\| \boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{h}^\star \boldsymbol{x}^{\star \mathsf{H}} \right\|_{\mathrm{F}} \lesssim \sigma \sqrt{\frac{K \log m}{m}} \tag{14}$$

holds with probability at least $1 - O(m^{-5} + m \exp(-c_1 K))$ *for some constant* $c_1 > 0$. *In addition, the bounds in (14) continue to hold if* $\boldsymbol{Z}_{\mathsf{cvx}}$ *is replaced by*

$$\boldsymbol{Z}_{\mathsf{cvx},1} := \arg \min_{\boldsymbol{Z}:\mathsf{rank}(\boldsymbol{Z}) \leq 1} \left\| \boldsymbol{Z} - \boldsymbol{Z}_{\mathsf{cvx}} \right\|_{\mathrm{F}},$$

i.e. the best rank-1 approximation of $\boldsymbol{Z}_{\mathsf{cvx}}$.

This theorem, which is in parallel to Theorem 1 for Fourier designs, confirms the appealing statistical guarantees of convex relaxation under Gaussian designs. The minimax optimality of this result will be discussed in Section 2.3 in detail.

Theorem 4 (Nonconvex optimization)**.** *Set* $\lambda = C_\lambda \sigma \sqrt{mK \log m}$ *for some large enough constant* $C_\lambda > 0$. *Take* $\eta = c_\eta / m$ *for some sufficiently small constant* $c_\eta > 0$. *Suppose that Assumption 2 and Condition (13) hold. Then with probability at least* $1 - O\left(m^{-5} + m e^{-K}\right)$, *the iterates* $\{\boldsymbol{h}^t, \boldsymbol{x}^t\}_{0 \leq t \leq t_0}$ *of Algorithm 1 obey*

$$\mathsf{dist}\left(\boldsymbol{z}^0, \boldsymbol{z}^\star\right) \lesssim \sqrt{\frac{K \log^2 m}{m}} \left\| \boldsymbol{z}^\star \right\|_2 + \sigma \sqrt{\frac{K \log m}{m \left\| \boldsymbol{h}^\star \boldsymbol{x}^{\star \mathsf{H}} \right\|_{\mathrm{F}}}}, \tag{15a}$$

$$\mathsf{dist}\left(\boldsymbol{z}^t, \boldsymbol{z}^\star\right) \leq \rho^t \mathsf{dist}\left(\boldsymbol{z}^0, \boldsymbol{z}^\star\right) + \frac{C_{11}\left(\lambda + \sigma\sqrt{mK \log m}\right)}{c_\rho m \left\| \boldsymbol{h}^\star \boldsymbol{x}^{\star \mathsf{H}} \right\|_{\mathrm{F}}^{1/2}} \tag{15b}$$

$$\left\| \boldsymbol{h}^t \left(\boldsymbol{x}^t\right)^{\mathsf{H}} - \boldsymbol{h}^\star \boldsymbol{x}^{\star \mathsf{H}} \right\|_{\mathrm{F}} \leq 2\rho^t \mathsf{dist}\left(\boldsymbol{z}^0, \boldsymbol{z}^\star\right) \left\| \boldsymbol{z}^\star \right\|_2 + \frac{2C_{11}\left(\lambda + \sigma\sqrt{mK \log m}\right)}{c_\rho m} \tag{15c}$$

simultaneously for all $0 \le t \le t_0 \le m^{20}$*. Here, we take* $C_{11} > 0$ *to be some sufficiently large constant and* $0 < \rho = 1 - c_\rho c_\eta < 1$ *for some sufficiently small constant* $c_\rho > 0$*.*

Similar to the Fourier designs studied in Section 2.1, our theory asserts that the estimation error of $\{\boldsymbol{h}^t, \boldsymbol{x}^t\}_{0 \le t \le t_0}$ produced by Algorithm 1 decreases geometrically fast before reaching an error floor on the order of the minimax-optimal statistical limit modulo some logarithmic factor (cf. Theorem 5).

2.3 Insights

The above theorems strengthen our understanding about the performance of both convex and nonconvex algorithms in the presence of random noise. In what follows, we elaborate on the tightness of our results as well as other important algorithmic implications.

- *Minimax optimality of both convex relaxation and nonconvex optimization.* Theorems 1-2 (resp. Theorems 3-4) reveal that both convex and nonconvex optimization estimate $\boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}}$ to within an Euclidean error at most $\sigma\sqrt{K}$ (resp. $\sigma\sqrt{K/m}$) up to some log factor for random Fourier design (resp. Gaussian design), provided that the regularization parameter is taken to be $\lambda \asymp \sigma\sqrt{K \log m}$ (resp. $\lambda \asymp \sigma\sqrt{mK \log m}$). This closes the gap between the statistical guarantees for convex and nonconvex optimization, confirming that convex relaxation is no less statistically efficient than nonconvex optimization. Further, in order to assess the statistical optimality of our results, it is instrumental to understand the statistical limit one can hope for. This is provided in the following claim, whose proof is postponed to Section E of supplementary materials.

  Theorem 5. *Suppose that the noise components obey* $\xi_j \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2/2) + i\mathcal{N}(0, \sigma^2/2)$*.*

Define

$$\mathcal{M}^\star := \left\{ \boldsymbol{Z} = \boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} \,\middle|\, \boldsymbol{h}, \boldsymbol{x} \in \mathbb{C}^K \right\}.$$

Then under Assumption 1, there exists some universal constant $c_{\mathsf{lb}}^{(1)} > 0$ *such that, with probability exceeding* $1 - O(K^{-10})$*,*

$$\inf_{\widehat{\boldsymbol{Z}}} \sup_{\boldsymbol{Z}^\star \in \mathcal{M}^\star} \mathbb{E}\left[\left\| \widehat{\boldsymbol{Z}} - \boldsymbol{Z}^\star \right\|_{\mathrm{F}}^2 \mid \boldsymbol{A} \right] \geq c_{\mathsf{lb}}^{(1)} \frac{\sigma^2 K}{\log m}, \tag{16}$$

where the infimum is taken over all estimator $\widehat{\boldsymbol{Z}}$*. Furthermore, under Assumption 2, there exists another universal constant* $c_{\mathsf{lb}}^{(2)} > 0$ *such that*

$$\inf_{\widehat{\boldsymbol{Z}}} \sup_{\boldsymbol{Z}^\star \in \mathcal{M}^\star} \mathbb{E}\left[\left\| \widehat{\boldsymbol{Z}} - \boldsymbol{Z}^\star \right\|_{\mathrm{F}}^2 \mid \boldsymbol{A}, \boldsymbol{B} \right] \geq c_{\mathsf{lb}}^{(2)} \frac{\sigma^2 K}{m \log m} \tag{17}$$

holds with probability exceeding $1 - O(K^{-10})$*.*

Encouragingly, the minimax lower bound (16) (resp. (17)) matches the statistical error bounds in Theorems 1-2 (resp. Theorems 3-4) up to some logarithmic factor, thus confirming the near minimaxity of both convex relaxation and nonconvex optimization for blind deconvolution under both designs.

- *Fast convergence of nonconvex algorithms.* From the computational perspective, Theorem 2 guarantees linear convergence (or geometric convergence) of the nonconvex algorithm with a contraction rate ρ. Given that $1-\rho$ is a constant bounded away from 1 (as long as the stepsize is taken to be a sufficiently small constant), the iteration complexity of the algorithm scales at most logarithmically with the model parameters. As a result, the total computational complexity is proportional to the per-iteration cost $O(mK)$ (up to some log factor), which scales nearly linearly with the time taken to read the data. Compared with past work on nonconvex algorithms [Li et al., 2019, Huang and

Hand, 2018], our theory reveals considerably faster convergence and hence improved computational cost, without compromising statistical efficiency. A key enabler of the improved theory lies in fine-grained understanding of the part of optimization lanscape visited by the nonconvex algorithm, thus allowing for the use of more aggressive constant step sizes instead of diminishing step sizes. See Table 1 for details.

The careful reader might immediately remark that the validity of the above results requires the assumptions (8) on both the sample size and the noise level. Fortunately, a closer inspection of these conditions reveals the broad applicability of these conditions.

- *Sample complexity.* The sample size requirement in our theory of blind deconvolution under Fourier design (resp. Gaussian design), as stated in Condition (8) (resp. Condition (13)), scales as

$$m \gtrsim K \operatorname{poly}\log(m),$$

which matches the information-theoretical lower limit even in the absence of noise (modulo some logarithmic factor) as proved in Kech and Krahmer [2017] (resp. Cai et al. [2015]).

- *Signal-to-noise ratio (SNR).* The noise level required for our theory to work under Fourier design (see Condition (8)) is given by $\sigma\sqrt{K\log^5 m} \lesssim \left\|\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}}$. If we define the sample-wise signal-to-noise ratio as follows

$$\mathsf{SNR} := \frac{\frac{1}{m}\sum_{k=1}^m \mathbb{E}\left[\left|\boldsymbol{b}_k^{\mathsf{H}}\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}\right|^2\right]}{\sigma^2}, \tag{18}$$

then our noise requirement can be equivalently phrased as

$$\mathsf{SNR} = \frac{\|\boldsymbol{h}^\star\|_2^2\|\boldsymbol{x}^\star\|_2^2}{m\sigma^2} \gtrsim \frac{K\log^5 m}{m},$$

where the right-hand side of the above relation is vanishingly small in light of our sample complexity constraint $m \gtrsim \mu^2 K \log^9 m$. In other words, our theory works even in the low-SNR regime. Furthermore, for the Gaussian design, the noise level required in our theory is $\sigma\sqrt{K \log^5 m/m} \lesssim \left\|\boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}}$. We can introduce the following SNR that allows us to rewrite this requirement as

$$\mathsf{SNR} = \frac{\frac{1}{m}\sum_{k=1}^m \mathbb{E}\big[\big|\boldsymbol{b}_k^{\mathsf{H}}\boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}\big|^2\big]}{\sigma^2} = \frac{\|\boldsymbol{h}^\star\|_2^2\|\boldsymbol{x}^\star\|_2^2}{\sigma^2} \gtrsim \frac{K\log^5 m}{m},$$

which resembles the one for Fourier designs.

3 Prior art

Before embarking on our discussion on the prior art for blind deconvolution, it is noteworthy that the model (1) might remind readers of the famous problem of phase retrieval [Candes et al., 2013, Shechtman et al., 2015, Chi et al., 2019], which is concerned with solving random quadratic systems of equations and clearly related to the problem of solving bilinear systems. Despite the similarity between these two problems at first glance, the majority of prior phase retrieval theory focuses on either i.i.d. Gaussian designs or randomized coded diffraction patterns, which are drastically different from the kind of random Fourier designs commonly assumed in blind deconvolution. In fact, the presence of Fourier designs in blind deconvolution is a consequence of the duality relation between convolution in the time domain and multiplication in the frequency domain [Ahmed et al., 2013, Li et al., 2019]. The deterministic nature of the Fourier design matrix $\boldsymbol{B}$ under the Fourier model, however, presents a substantial challenge in the analysis of both convex and nonconvex optimization algorithms; in contrast, the Gaussian design matrix in prior phase retrieval theory is assumed to be highly random, which remarkably simplifies analysis.

We now turn attention to the blind deconvolution literature. As mentioned previously, recent years have witnessed much progress towards understanding convex and nonconvex optimization for solving bilinear systems of equations. First, we give a brief review on previous literature of blind deconvolution under Fourier design. Regarding the convex programming approach, Ahmed et al. [2013] was the first to apply the lifting idea to transform bilinear system of equations into linear measurements about a rank-one matrix — an idea that has proved effective in a number of nonconvex problems [Candes et al., 2013, Waldspurger et al., 2015, Chen and Chi, 2014, Tang et al., 2013, Chi, 2016, Chen et al., 2014, Goemans and Williamson, 1994, Shechtman et al., 2014, Oymak et al., 2015]. Focusing on convex relaxing in the lifted domain, Ahmed et al. [2013] showed that exact recovery is possible from a near-optimal number of measurements in the noiseless case, and developed the first statistical guarantees for the noisy case (which are, as alluded to previously, highly suboptimal). Several other works have also been devoted to understanding convex relaxation under possibly different assumptions. Another paper Aghasi et al. [2019] proposed an effective convex algorithm for bilinear inversion, assuming that the signs of the signals are known *a priori*. Moving beyond blind deconvolution, the convex approach has been extended to accommodate the blind demixing problem [Ling and Strohmer, 2017, Jung et al., 2017], which is more general than blind deconvolution.

$$\underset{\boldsymbol{Z}\in\mathbb{C}^{K\times K}}{\text{minimize}} \quad \|\boldsymbol{Z}\|_* \qquad \text{subject to} \quad \boldsymbol{y} = \mathcal{A}(\boldsymbol{Z}).$$

Another line of works has focused on the development of fast nonconvex algorithms [Li et al., 2019, Lee et al., 2018, Ma et al., 2018, Huang and Hand, 2018, Ling and Strohmer, 2019, Charisopoulos et al., 2019, 2021], which was largely motivated by recent advances in efficient nonconvex optimization for tackling statistical estimation problems [Candes et al., 2015, Chen and Candès, 2017, Charisopoulos et al., 2021, Keshavan et al., 2009, Jain et al., 2013, Zhang et al., 2016, Chen and Wainwright, 2015, Sun and Luo, 2016, Zheng and Lafferty, 2016, Wang et al., 2017a, Cai et al., 2021b, Wang et al., 2017b, Qu et al., 2017, Duchi and Ruan, 2019, Ma et al., 2019] (see Chi et al. [2019] for an overview). Li et al. [2019] proposed a feasible nonconvex recipe by attempting to optimize a regularized squared loss (which includes extra penalty term to promote incoherence), and showed that in conjunction with proper initialization, nonconvex gradient descent converges to the ground truth in the absence of noise. Another work Huang and Hand [2018] proposed a Riemannian steepest descent method by exploiting the quotient structure, which is also guaranteed to work in the noise-free setting with nearly minimal sample complexity. Further, Ling and Strohmer [2019], Dong and Shi [2018] extended the nonconvex paradigm to accommodate the blind demixing problem, which subsumes blind deconvolution a special case.

Going beyond algorithm designs, the past works Li et al. [2016, 2015], Kech and Krahmer [2017] investigated how many samples are needed to ensure the identifiability of blind deconvolution under the subspace model. Furthermore, it is worth noting that another line of recent works Wang and Chi [2016], Lee et al. [2016], Zhang et al. [2017, 2019, 2020], Li and Bresler [2019], Shi and Chi [2021], Qu et al. [2019] studied a different yet fundamentally important model of blind deconvolution, assuming that one of the two signals is sparse instead of lying within a known subspace. These are, however, beyond the scope of the current paper.

In addition, as far as we know, previous works on blind deconvolution under Gaussian design is not as extensive as the case with Fourier designs, the latter of which is closer to practical blind deconvolution applications. Among the most relevant works: Cai et al. [2015] proposed a constrained convex optimization problem under the same setting as Assumption 2 and establishes that the estimation error is bounded by $\sigma \min\{K\sqrt{\log m}/m+$

$\sqrt{K/m}, 1\}$, which is on the same order (up to logarithmic factors) as our bound in Theorem 3 when $m \gg K \log m$ and matches the minimax optimal estimation error lower bound; Zhong et al. [2015] studied the noiseless case in terms of both convex and nonconvex formulations; Charisopoulos et al. [2019] analyzed the nonsmooth nonconvex formulation of the problem for bilinear measurements with corruption frequency less than $1/2$, and proved that the subgradient algorithms proposed there converges linearly, while the specific prox-linear method converges quadratically albeit with higher per-iteration cost. Compared with these works, our paper studies the unconstrained version of convex relaxation and establishes an estimation error upper bound that nearly matches the minimax lower bound. When it comes to nonconvex formulation, the current paper is, as far as we know, the first to justify the optimality of its estimation accuracy in the noisy setting.

At the technical level, the pivotal idea of our paper lies in bridging convex and nonconvex estimators, which is motivated by prior works Chen et al. [2020b, 2019c, 2020c] on matrix completion and robust principal component analysis. Such crucial connections have been established with the assistance of the leave-one-out analysis framework, which has already proved effective in analyzing a variety of nonconvex statistical problems [El Karoui, 2018, Chen et al., 2019a,b, Ding and Chen, 2020, Cai et al., 2020, Dong and Shi, 2018, Xu et al., 2019, Cai et al., 2021a, Chen et al., 2020a, Zhong and Boumal, 2018].

4 Numerical experiments

In this subsection, we carry out a series of numerical experiments to confirm the validity of our theory. Throughout the experiments, the signals of interest $\boldsymbol{h}^\star, \boldsymbol{x}^\star \in \mathbb{C}^K$ are drawn from $\mathcal{N}\left(\mathbf{0}, \frac{1}{2K}\boldsymbol{I}_K\right) + \mathrm{i}\mathcal{N}\left(\mathbf{0}, \frac{1}{2K}\boldsymbol{I}_K\right)$ (so that they have approximately unit ℓ_2 norm). Under the

Assumption 1 (resp. Assumption 2), the stepsize η is set to be 0.05 (resp. $0.05/m$), whereas the regularization parameter is taken to be $\lambda = 5\sigma\sqrt{K \log m}$ (resp. $\lambda = 5\sigma\sqrt{mK \log m}$). The convex problem is solved by means of the proximal gradient method [Parikh and Boyd, 2014].

In the first series of experiments, we report the statistical estimation errors of both convex and nonconvex approaches as the noise level σ varies from 10^{-6} to 10^{-3} for blind deconvolution under Fourier design, while the noise level for blind deconvolution under Gaussian design is from 10^{-5} to 10^{-2}; here, we set $K = 100$ and $m = 10K$. Let $\boldsymbol{Z}_{\mathsf{ncvx}} = \boldsymbol{h}_{\mathsf{ncvx}}\boldsymbol{x}_{\mathsf{ncvx}}^{\mathsf{H}}$ be the nonconvex solution and $\boldsymbol{Z}_{\mathsf{cvx}}$ be the convex solution. Figure 1 depicts the relative Euclidean estimation errors ($\|\boldsymbol{Z}_{\mathsf{ncvx}} - \boldsymbol{Z}^{\star}\|_{\mathrm{F}} / \|\boldsymbol{Z}^{\star}\|_{\mathrm{F}}$ and $\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{Z}^{\star}\|_{\mathrm{F}} / \|\boldsymbol{Z}^{\star}\|_{\mathrm{F}}$) vs. the noise level, where the results are averaged from 20 independent trials. Clearly, both approaches enjoy almost identical statistical accuracy, thus confirming the optimality of convex relaxation as well. Another interesting observation revealed by Figure 1 is the closeness of the solutions of these two approaches, which, as we shall elucidate momentarily, forms the basis of our analysis idea.

In the second series of experiments, we report the numerical convergence of gradient descent (cf. Algorithm 1). We choose $K \in \{30, 100, 300, 1000\}$ and let $m = 10K$, with the noise level fixed at $\sigma = 10^{-4}$. Figure 2 plots the relative Euclidean estimation error $\left\|\boldsymbol{h}^{t}\boldsymbol{x}^{t\mathsf{H}} - \boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}} / \left\|\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}}$ vs. the iteration count. As can be seen from the plots, the nonconvex gradient algorithm studied here converges linearly (in fact, within around 200-300 iterations) before it hits an error floor. In addition, the relative error of blind deconvolution under Fourier design increases as the dimension K increases, which is consistent with Theorem 2. While the relative error of blind deconvolution under Gaussian design remains generally the same across different choices of K, this can be explained by

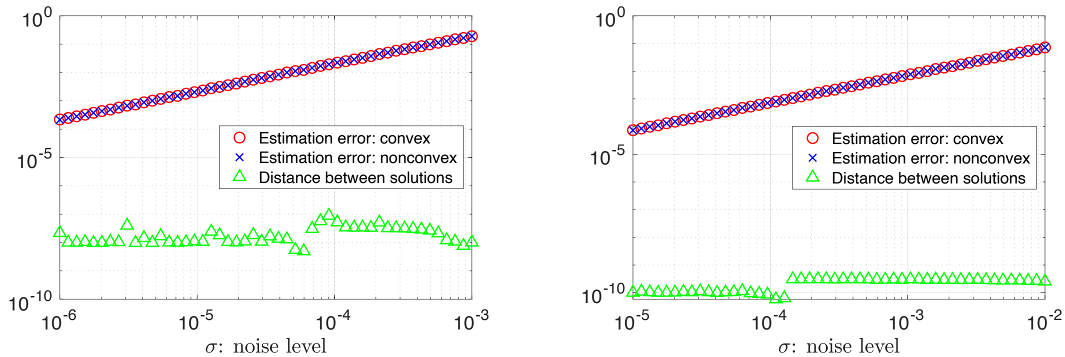

Figure 1: Left: blind deconvolution under Fourier design. Right: blind deconvolution under Gaussian design. Relative estimation errors of both $\boldsymbol{Z}_{\mathsf{cvx}}$ and $\boldsymbol{Z}_{\mathsf{ncvx}}$ and the relative distance between them vs. the noise level σ. The results are averaged over 20 independent trials.

Theorem 4 since the ratio between m and K is kept to be 10.

In the last series of experiments, we examine the necessity of the incoherence condition (7) empirically. The experiments are conducted with μ^2 taking on 10 equidistant values from 3 to 30. For each choice of μ, $\boldsymbol{h}^\star$ is generated by first setting the first μ^2 entries to be 1 and the others 0 , and then normalizing it to have unit norm; $\boldsymbol{x}^\star$ is generated randomly from Gaussian distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_K)$ and then normalized to have unit norm. This way we guarantee that $\max_{1\leq j\leq m} |\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star| = \mu/\sqrt{m}$. We fix $K = 100$ and the noise level $\sigma = 10^{-4}$ throughout. For each μ^2 and m, 20 random trials are conducted. In each trial, we run convex and nonconvex algorithms until convergence or the maximum number of iterations is reached, and then report the relative Euclidean error $\left\|\boldsymbol{h}^t\boldsymbol{x}^{t\mathsf{H}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}}$. If the relative error is less than 0.1, the trial is declared as successful. The proportion of successful recovery for convex and nonconvex problems are plotted in Figure 3, which suggests that sample complexity m does scale linearly with μ^2 for both problems and hence

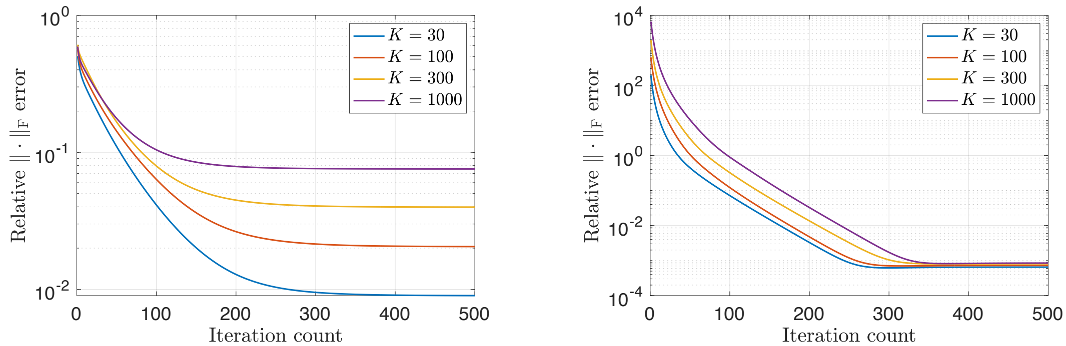

Figure 2: Left: blind deconvolution. Right: Gaussian design. Relative Euclidean error $\left\|\boldsymbol{h}^t\boldsymbol{x}^{t\mathsf{H}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}}$ vs. iteration count.

corroborates the theoretical results provided in Theorems 1 and 2.

5 Discussion

This paper has investigated the effectiveness of both convex relaxation and nonconvex optimization in solving bilinear systems of equations in the presence of random noise. We have demonstrated that a simple two-stage nonconvex algorithm solves the problem to optimal statistical accuracy within nearly linear time. Further, by establishing an intimate connection between convex programming and nonconvex optimization, we have established — for the first time — optimal statistical guarantees of convex relaxation when applied to blind deconvolution. Our results are established for two different types of design mechanisms: the random Fourier design and the Gaussian design. Our results considerably improve upon the state-of-the-art theory for blind deconvolution, and contribute towards demystifying the efficacy of optimization-based methods in solving this fundamental nonconvex problem.

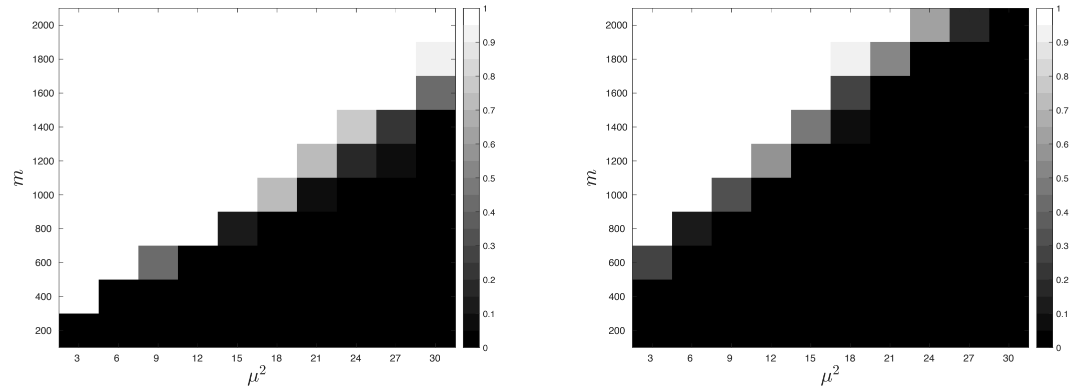

Figure 3: Left: nonconvex problem. Right: convex problem. Sample size m vs. squared incoherence μ^2. The scaled colormap represents the proportion of successful recovery out of 20 random trials.

Moving forward, the findings of this paper suggest multiple directions that merit further investigations. For instance, while the current paper adopts a balancing operation in each iteration of the nonconvex algorithm (cf. Algorithm 1), it might not be necessary in practice; in fact, numerical experiments suggest that the size of the scaling parameter $|\alpha^t|$ stays close to 1 even without proper balancing. It would be interesting to investigate whether vanilla GD without rescaling is able to achieve comparable performance. In addition, the estimation guarantees provided in this paper might serve as a starting point for conducting uncertainty quantification for noisy blind deconvolution — namely, how to use it to construct valid and short confidence intervals for the unknowns. Going beyond blind deconvolution, it would be of interest to extend the current analysis to handle blind demixing — a problem that can be viewed as an extension of blind deconvolution beyond the rank-one setting [Ling and Strohmer, 2017, 2019, Dong and Shi, 2018]. As can be expected, existing statistical guarantees for convex programming remain highly suboptimal for noisy blind demixing, and the analysis developed in the current paper suggests a feasible path towards closing the gap.

References

A. Aghasi, A. Ahmed, P. Hand, and B. Joshi. Branchhull: Convex bilinear inversion from the entrywise product of signals with known signs. *Applied and Computational Harmonic Analysis*, 2019.

A. Ahmed, B. Recht, and J. Romberg. Blind deconvolution using convex programming. *IEEE Transactions on Information Theory*, 60(3):1711–1732, 2013.

C. Cai, H. V. Poor, and Y. Chen. Uncertainty quantification for nonconvex tensor completion: Confidence intervals, heteroscedasticity and optimality. In *International Conference on Machine Learning*, pages 1271–1282. PMLR, 2020.

C. Cai, G. Li, Y. Chi, H. V. Poor, and Y. Chen. Subspace estimation from unbalanced and incomplete data matrices: $\ell_{2,\infty}$ statistical guarantees. *The Annals of Statistics*, 49 (2):944–967, 2021a.

C. Cai, G. Li, H. V. Poor, and Y. Chen. Nonconvex low-rank tensor completion from noisy data. *Operations Research*, 2021b.

T. T. Cai, A. Zhang, et al. Rop: Matrix recovery via rank-one projections. *Annals of Statistics*, 43(1):102–138, 2015.

P. Campisi and K. Egiazarian. *Blind image deconvolution: theory and applications*. CRC press, 2016.

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

E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

E. J. Candes, T. Strohmer, and V. Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

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

T. F. Chan and C.-K. Wong. Total variation blind deconvolution. *IEEE transactions on Image Processing*, 7(3):370–375, 1998.

V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

V. Charisopoulos, D. Davis, M. Díaz, and D. Drusvyatskiy. Composite optimization for robust blind deconvolution. *arXiv preprint arXiv:1901.01624*, 2019.

V. Charisopoulos, Y. Chen, D. Davis, M. Díaz, L. Ding, and D. Drusvyatskiy. Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence. *Foundations of Computational Mathematics*, pages 1–89, 2021.

P. Chen, C. Gao, and A. Y. Zhang. Partial recovery for top-k ranking: Optimality of mle and sub-optimality of spectral method. *arXiv preprint arXiv:2006.16485*, 2020a.

Y. Chen and E. J. Candès. 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.

Y. Chen and Y. Chi. Robust spectral compressed sensing via structured matrix completion. *IEEE Transactions on Information Theory*, 10(60):6576–6601, 2014.

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

Y. Chen, L. Guibas, and Q. Huang. Near-optimal joint object matching via convex relaxation. In *International Conference on International Conference on Machine Learning*, 2014.

Y. Chen, Y. Chi, J. Fan, and C. Ma. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176(1-2): 5–37, 2019a.

Y. Chen, J. Fan, C. Ma, and K. Wang. Spectral method and regularized mle are both optimal for top-k ranking. *Annals of statistics*, 47(4):2204, 2019b.

Y. Chen, J. Fan, C. Ma, and Y. Yan. Inference and uncertainty quantification for noisy matrix completion. *Proceedings of the National Academy of Sciences*, 116(46):22931–22937, 2019c.

Y. Chen, Y. Chi, J. Fan, C. Ma, and Y. Yan. Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *SIAM Journal on Optimization*, 30(4):3098–3121, 2020b.

Y. Chen, J. Fan, C. Ma, and Y. Yan. Bridging convex and nonconvex optimization in robust pca: Noise, outliers, and missing data. *arXiv preprint arXiv:2001.05484, accepted to Annals of Statistics*, 2020c.

Y. Chi. Guaranteed blind sparse spikes deconvolution via lifting and convex optimization. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):782–794, 2016.

Y. Chi, Y. M. Lu, and Y. Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.

L. Ding and Y. Chen. Leave-one-out approach for matrix completion: Primal and dual analysis. *IEEE Transactions on Information Theory*, 2020.

J. Dong and Y. Shi. Nonconvex demixing from bilinear measurements. *IEEE Transactions on Signal Processing*, 66(19):5152–5166, 2018.

J. C. Duchi and F. Ruan. Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. *Information and Inference: A Journal of the IMA*, 8(3):471–529, 2019.

N. El Karoui. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, 170(1-2):95–175, 2018.

M. X. Goemans and D. P. Williamson. . 879-approximation algorithms for max cut and max 2sat. In *Proceedings of the twenty-sixth annual ACM symposium on Theory of computing*, pages 422–431, 1994.

G. H. Golub and C. F. Van Loan. *Matrix computations*, volume 3. JHU press, 2013.

W. Huang and P. Hand. Blind deconvolution by a steepest descent algorithm on a quotient manifold. *SIAM Journal on Imaging Sciences*, 11(4):2757–2785, 2018.

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

S. M. Jefferies and J. C. Christou. Restoration of astronomical images by iterative blind deconvolution. *The Astrophysical Journal*, 415:862, 1993.

P. Jung, F. Krahmer, and D. Stöger. Blind demixing and deconvolution at near-optimal rate. *IEEE Transactions on Information Theory*, 64(2):704–727, 2017.

M. Kech and F. Krahmer. Optimal injectivity conditions for bilinear inverse problems with applications to identifiability of deconvolution problems. *SIAM Journal on Applied Algebra and Geometry*, 1(1):20–37, 2017.

R. Keshavan, A. Montanari, and S. Oh. Matrix completion from noisy entries. In *Advances in neural information processing systems*, pages 952–960, 2009.

D. Kundur and D. Hatzinakos. Blind image deconvolution. *IEEE signal processing magazine*, 13(3):43–64, 1996.

K. Lee, Y. Li, M. Junge, and Y. Bresler. Blind recovery of sparse signals from subsampled convolution. *IEEE Transactions on Information Theory*, 63(2):802–821, 2016.

K. Lee, N. Tian, and J. Romberg. Fast and guaranteed blind multichannel deconvolution under a bilinear system model. *IEEE Transactions on Information Theory*, 64(7):4792–4818, 2018.

X. Li, S. Ling, T. Strohmer, and K. Wei. Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *Applied and computational harmonic analysis*, 47(3):893–934, 2019.

Y. Li and Y. Bresler. Multichannel sparse blind deconvolution on the sphere. *IEEE Transactions on Information Theory*, 65(11):7415–7436, 2019.

Y. Li, K. Lee, and Y. Bresler. A unified framework for identifiability analysis in bilinear inverse problems with applications to subspace and sparsity models. *arXiv preprint arXiv:1501.06120*, 2015.

Y. Li, K. Lee, and Y. Bresler. Identifiability in blind deconvolution with subspace or sparsity constraints. *IEEE Transactions on information Theory*, 62(7):4266–4275, 2016.

S. Ling and T. Strohmer. Self-calibration and biconvex compressive sensing. *Inverse Problems*, 31(11):115002, 2015.

S. Ling and T. Strohmer. Simultaneous blind deconvolution and blind demixing via convex programming. In *2016 50th Asilomar Conference on Signals, Systems and Computers*, pages 1223–1227. IEEE, 2016.

S. Ling and T. Strohmer. Blind deconvolution meets blind demixing: Algorithms and performance bounds. *IEEE Transactions on Information Theory*, 63(7):4497–4520, 2017.

S. Ling and T. Strohmer. 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, 2019.

C. Ma, K. Wang, Y. Chi, and Y. Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In *International Conference on Machine Learning*, pages 3345–3354. PMLR, 2018.

J. Ma, J. Xu, and A. Maleki. Optimization-based amp for phase retrieval: The impact of initialization and ℓ_2 regularization. *IEEE Transactions on Information Theory*, 65(6): 3600–3629, 2019.

S. Oymak, A. Jalali, M. Fazel, Y. C. Eldar, and B. Hassibi. Simultaneously structured models with application to sparse and low-rank matrices. *IEEE Transactions on Information Theory*, 61(5):2886–2908, 2015.

N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1 (3):127–239, 2014.

Q. Qu, Y. Zhang, Y. Eldar, and J. Wright. Convolutional phase retrieval. In *Advances in Neural Information Processing Systems*, pages 6086–6096, 2017.

Q. Qu, X. Li, and Z. Zhu. A nonconvex approach for exact and efficient multichannel sparse blind deconvolution. In *Advances in Neural Information Processing Systems*, pages 4015–4026, 2019.

Y. Shechtman, A. Beck, and Y. C. Eldar. Gespar: Efficient phase retrieval of sparse signals. *IEEE transactions on signal processing*, 62(4):928–938, 2014.

Y. Shechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev. Phase retrieval with application to optical imaging: a contemporary overview. *IEEE signal processing magazine*, 32(3):87–109, 2015.

L. Shi and Y. Chi. Manifold gradient descent solves multi-channel sparse blind deconvolution provably and efficiently. *IEEE Transactions on Information Theory*, 2021.

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

G. Tang, B. N. Bhaskar, P. Shah, and B. Recht. Compressed sensing off the grid. *IEEE transactions on information theory*, 59(11):7465–7490, 2013.

L. Tong, G. Xu, and T. Kailath. Blind identification and equalization based on second-order statistics: A time domain approach. *IEEE Transactions on information Theory*, 40(2):340–349, 1994.

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

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

G. Wang, G. B. Giannakis, and Y. C. Eldar. Solving systems of random quadratic equations via truncated amplitude flow. *IEEE Transactions on Information Theory*, 64(2):773–794, 2017a.

G. Wang, L. Zhang, G. B. Giannakis, M. Akçakaya, and J. Chen. Sparse phase retrieval via truncated amplitude flow. *IEEE Transactions on Signal Processing*, 66(2):479–491, 2017b.

L. Wang and Y. Chi. Blind deconvolution from multiple sparse inputs. *IEEE Signal Processing Letters*, 23(10):1384–1388, 2016.

X. Wang and H. V. Poor. Blind equalization and multiuser detection in dispersive cdma channels. *IEEE Transactions on Communications*, 46(1):91–103, 1998.

G. Wunder, H. Boche, T. Strohmer, and P. Jung. Sparse signal processing concepts for efficient 5g system design. *IEEE Access*, 3:195–208, 2015.

J. Xu, A. Maleki, and K. R. Rad. Consistent risk estimation in high-dimensional linear regression. *arXiv preprint arXiv:1902.01753*, 2019.

H. Zhang, Y. Chi, and Y. Liang. Provable non-convex phase retrieval with outliers: Median truncatedwirtinger flow. In *International conference on machine learning*, pages 1022–1031, 2016.

Y. Zhang, Y. Lau, H.-w. Kuo, S. Cheung, A. Pasupathy, and J. Wright. On the global geometry of sphere-constrained sparse blind deconvolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4894–4902, 2017.

Y. Zhang, H.-W. Kuo, and J. Wright. Structured local optima in sparse blind deconvolution. *IEEE Transactions on Information Theory*, 66(1):419–452, 2019.

Y. Zhang, Q. Qu, and J. Wright. From symmetry to geometry: Tractable nonconvex problems. *arXiv preprint arXiv:2007.06753*, 2020.

Q. Zheng and J. Lafferty. Convergence analysis for rectangular matrix completion using burer-monteiro factorization and gradient descent. *arXiv preprint arXiv:1605.07051*, 2016.

K. Zhong, P. Jain, and I. S. Dhillon. Efficient matrix sensing using rank-1 gaussian measurements. In *International conference on algorithmic learning theory*, pages 3–18. Springer, 2015.

Y. Zhong and N. Boumal. Near-optimal bounds for phase synchronization. *SIAM Journal on Optimization*, 28(2):989–1016, 2018.

Proofs for the Paper "Convex and Nonconvex Optimization Are Both Minimax-Optimal for Noisy Blind Deconvolution under Random Designs"

Yuxin Chen* Jianqing Fan† Bingyan Wang† Yuling Yan†

July 12, 2021

Contents

*Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08544, USA; Email: yuxin.chen@princeton.edu.

†Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA; Email: {jqfan, bingyanw, yulingy}@princeton.edu.

Appendix structure

Appendix A and B analyze the Fourier designs. In Appendix A, we present the analysis of the nonconvex gradient method and the proof of Theorem 2. Appendix B gives the complete proof of Theorem 1. In addition, Appendix C and D and provide proofs for the Gaussian designs, while Appendix C proves Theorem 4 and Appendix D proves Theorem 3. Appendix E justifies two minimax lower bounds in Theorem 5. Appendix F lists several useful lemmas and their proofs.

A Analysis: Nonconvex gradient method under Fourier design

Since the proof of Theorem 1 is built upon Theorem 2, we shall first present the proof of the nonconvex part. Without loss of generality, we assume that

$$\|\boldsymbol{h}^\star\|_2 = \|\boldsymbol{x}^\star\|_2 = 1 \tag{A.1}$$

throughout the proof. For the sake of notational convenience, for each iterate $(\boldsymbol{h}^t, \boldsymbol{x}^t)$ we define the following alignment parameters

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

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

which lead to the following simple relations

$$\alpha^{t+1} = \sqrt{\frac{\left\|\boldsymbol{x}^{t+1/2}\right\|_2}{\left\|\boldsymbol{h}^{t+1/2}\right\|_2}}\,\alpha^{t+1/2} \qquad \text{and} \qquad \mathsf{dist}\big(\boldsymbol{z}^{t+1/2}, \boldsymbol{z}^\star\big) = \mathsf{dist}\big(\boldsymbol{z}^{t+1}, \boldsymbol{z}^\star\big). \tag{A.3}$$

With these in place, attention should be directed to the properly rescaled iterate

$$\widetilde{\boldsymbol{z}}^{t+1/2} = \big(\widetilde{\boldsymbol{h}}^{t+1/2}, \widetilde{\boldsymbol{x}}^{t+1/2}\big) := \big(\tfrac{1}{\overline{\alpha^{t+1/2}}}\boldsymbol{h}^{t+1/2}, \alpha^{t+1/2}\boldsymbol{x}^{t+1/2}\big), \tag{A.4a}$$

$$\widetilde{\boldsymbol{z}}^{t} = \big(\widetilde{\boldsymbol{h}}^{t}, \widetilde{\boldsymbol{x}}^{t}\big) := \big(\tfrac{1}{\overline{\alpha^{t}}}\boldsymbol{h}^{t}, \alpha^{t}\boldsymbol{x}^{t}\big). \tag{A.4b}$$

Additionally, we shall also define

$$\widehat{\boldsymbol{z}}^{t+1/2} = \big(\widehat{\boldsymbol{h}}^{t+1/2}, \widehat{\boldsymbol{x}}^{t+1/2}\big) := \big(\tfrac{1}{\overline{\alpha^{t}}}\boldsymbol{h}^{t+1/2}, \alpha^{t}\boldsymbol{x}^{t+1/2}\big) \tag{A.5a}$$

$$\widehat{\boldsymbol{z}}^{t+1} = \big(\widehat{\boldsymbol{h}}^{t+1}, \widehat{\boldsymbol{x}}^{t+1}\big) := \big(\tfrac{1}{\overline{\alpha^{t}}}\boldsymbol{h}^{t+1}, \alpha^{t}\boldsymbol{x}^{t+1}\big) \tag{A.5b}$$

that are rescaled in a different way, which will appear often in the analysis.

A.1 Induction hypotheses

Our analysis is inductive in nature; more concretely, we aim to justify the following set of hypotheses by induction:

$$\mathsf{dist}\left(\boldsymbol{z}^{t}, \boldsymbol{z}^{\star}\right) \le \left\|\widetilde{\boldsymbol{z}}^{t-1/2} - \boldsymbol{z}^{\star}\right\|_2 \le \rho\,\mathsf{dist}\left(\boldsymbol{z}^{t-1}, \boldsymbol{z}^{\star}\right) + C_1\eta\left(\lambda + \sigma\sqrt{K\log m}\right), \tag{A.6a}$$

$$\max_{1\le l\le m}\left|\boldsymbol{a}_l^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^{t} - \boldsymbol{x}^{\star}\right)\right| \le C_3\left(\sqrt{\frac{\mu^2 K\log^2 m}{m}} + \sqrt{\log m}\left(\lambda + \sigma\sqrt{K\log m}\right)\right), \tag{A.6b}$$

$$\max_{1\le l\le m}\left|\boldsymbol{b}_l^{\mathsf{H}}\widetilde{\boldsymbol{h}}^{t}\right| \le C_4\left(\frac{\mu\log^2 m}{\sqrt{m}} + \sigma\right), \tag{A.6c}$$

where $\rho = 1-\eta/16$ and $C_1, C_3, C_4 > 0$ are some universal constants. Here, the hypothesis (A.6a) is made for all $0 < t \le t_0$, while the hypotheses (A.6b) and (A.6c) are made for all $0 \le t \le t_0$. Clearly, if the hypotheses (A.6a) can be established, then simple recursion yields

$$\begin{aligned}\mathsf{dist}\left(\boldsymbol{z}^{t}, \boldsymbol{z}^{\star}\right) &\lesssim \rho^{t}\mathsf{dist}\left(\boldsymbol{z}^{0}, \boldsymbol{z}^{\star}\right) + \frac{C_1\eta\left(\lambda + \sigma\sqrt{K\log m}\right)}{1-\rho} \\ &= \rho^{t}\mathsf{dist}\left(\boldsymbol{z}^{0}, \boldsymbol{z}^{\star}\right) + \frac{C_1\left(\lambda + \sigma\sqrt{K\log m}\right)}{c_\rho}, \qquad 0 \le t \le t_0\end{aligned} \tag{A.6d}$$

as claimed. Moreover, one might naturally wonder why we are in need of the additional hypotheses (A.6b) and (A.6c) that might seem irrelevant at first glance. As it turns out, these two hypotheses — which characterize certain incoherence conditions of the iterates w.r.t. the design vectors — play a pivotal role in the analysis, as they enable some sort of "restricted strong convexity" that proves crucial for guaranteeing linear convergence.

In addition, the analysis also relies upon the following important properties of the initialization, which we shall establish momentarily:

$$\mathsf{dist}\left(\boldsymbol{z}^{0}, \boldsymbol{z}^{\star}\right) \lesssim \sqrt{\frac{\mu^2 K\log m}{m}} + \sigma\sqrt{K\log m}, \tag{A.6e}$$

$$\max_{1\le j\le m}\left|\boldsymbol{a}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^{0} - \boldsymbol{x}^{\star}\right)\right| \lesssim \sqrt{\frac{\mu^2 K\log^2 m}{m}} + \sigma\sqrt{K}\log m, \tag{A.6f}$$

$$\max_{1\le l\le m}\left|\boldsymbol{b}_l^{\mathsf{H}}\widetilde{\boldsymbol{h}}^{0}\right| \lesssim \frac{\mu\log^2 m}{\sqrt{m}} + \sigma, \tag{A.6g}$$

$$\big||\alpha^0| - 1\big| \le 1/4. \tag{A.6h}$$

A.2 Preliminaries

Before proceeding to the proof, we gather several preliminary facts that will be useful throughout.

A.2.1 Wirtinger calculus and notation

Given that this problem concerns complex-valued vectors/matrices, we find it convenient to work with Wirtinger calculus; see Candes et al. [2015, Section 6] and Ma et al. [2018, Section D.3.1] for a brief introduction. Here, we shall simply record below the expressions for the Wirtinger gradient and the Wirtinger Hessian w.r.t. the objective function $f(\cdot)$ defined in (4):

$$\nabla_{\boldsymbol{h}} f\left(\boldsymbol{h}, \boldsymbol{x}\right) = \sum_{j=1}^{m} \left(\boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h} \boldsymbol{x}^{\mathsf{H}} \boldsymbol{a}_j - y_j\right) \boldsymbol{b}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x} + \lambda \boldsymbol{h}, \tag{A.7a}$$

$$\nabla_{\boldsymbol{x}} f\left(\boldsymbol{h}, \boldsymbol{x}\right) = \sum_{j=1}^{m} \overline{\left(\boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h} \boldsymbol{x}^{\mathsf{H}} \boldsymbol{a}_j - y_j\right)} \boldsymbol{a}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h} + \lambda \boldsymbol{x}, \tag{A.7b}$$

$$\nabla^2 f\left(\boldsymbol{h}, \boldsymbol{x}\right) = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}^{\mathsf{H}} & \overline{\boldsymbol{A}} \end{bmatrix}, \tag{A.7c}$$

where

$$\boldsymbol{A} := \begin{bmatrix} \sum_{j=1}^{m} \left|\boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}\right|^2 \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} + \lambda & \sum_{j=1}^{m} \left(\boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h} \boldsymbol{x}^{\mathsf{H}} \boldsymbol{a}_j - y_j\right) \boldsymbol{b}_j \boldsymbol{a}_j^{\mathsf{H}} \\ \sum_{j=1}^{m} \left[\left(\boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h} \boldsymbol{x}^{\mathsf{H}} \boldsymbol{a}_j - y_j\right) \boldsymbol{b}_j \boldsymbol{a}_j^{\mathsf{H}}\right]^{\mathsf{H}} & \sum_{j=1}^{m} \left|\boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}\right|^2 \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} + \lambda \end{bmatrix} \in \mathbb{C}^{2K \times 2K},$$
$$\boldsymbol{B} := \begin{bmatrix} \boldsymbol{0} & \sum_{j=1}^{m} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h} \left(\boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}\right)^{\mathsf{H}} \\ \sum_{j=1}^{m} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x} \left(\boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}\right)^{\mathsf{H}} & \boldsymbol{0} \end{bmatrix} \in \mathbb{C}^{2K \times 2K}.$$

Throughout this paper, we shall often use $f\left(\boldsymbol{h}, \boldsymbol{x}\right)$ and $f\left(\boldsymbol{z}\right)$ interchangeably for any $\boldsymbol{z} = \begin{bmatrix} \boldsymbol{h} \\ \boldsymbol{x} \end{bmatrix}$, whenever it is clear from the context.

Before proceeding, we present two useful properties of the operator $\mathcal{A}$ and the design vectors $\{\boldsymbol{b}_j\}_{j=1}^m$.

Lemma 1. *For $\mathcal{A}$ defined in (B.3), with probability at least $1 - m^{-\gamma}$,*

$$\|\mathcal{A}\| \leq \sqrt{2K \log K + \gamma \log m}.$$

Proof. See Li et al. [2019, Lemma 5.12]. □

Lemma 2. *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. See Ma et al. [2018, Lemma 48]. □

A.2.2 Leave-one-out auxiliary sequences

The key to establishing the incoherence hypotheses (A.6b) and (A.6c) is to introduce a collection of auxiliary leave-one-out sequences — an approach first introduced by Ma et al. [2018]. Specifically, for each $1 \leq l \leq m$, define the leave-one-out loss function as follows

$$f^{(l)}\left(\boldsymbol{h}, \boldsymbol{x}\right) := \sum_{j: j \neq l} \left|\boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h} \boldsymbol{x}^{\mathsf{H}} \boldsymbol{a}_j - y_j\right|^2 + \lambda \left\|\boldsymbol{h}\right\|_2^2 + \lambda \left\|\boldsymbol{x}\right\|_2^2,$$

which is obtained by discarding the lth sample. We then generate the auxiliary sequence $\{\boldsymbol{h}^{(t),l}, \boldsymbol{x}^{(t),l}\}_{t \geq 0}$ by running the same nonconvex algorithm w.r.t. $f^{(l)}(\cdot, \cdot)$, as summarized in Algorithm 1. In a nutshell, the resulting leave-one-out sequence $\{\boldsymbol{h}^{(t),l}, \boldsymbol{x}^{(t),l}\}_{t \geq 0}$ is statistically independent from the design vector $\boldsymbol{a}_l$ and is expected to stay exceedingly close to the original sequence (given that only a single sample is dropped), which in turn facilitate the analysis of the correlation of $\boldsymbol{a}_l$ and $\boldsymbol{x}^t$ as claimed in (A.6b). In the mean time, this strategy also proves useful in controlling the correlation of $\boldsymbol{b}_l$ and $\boldsymbol{h}^t$ as in (A.6c), albeit with more delicate arguments.

Algorithm 1 The lth leave-one-out sequence for nonconvex blind deconvolution

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

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

$$\boldsymbol{M}^{(l)} := \sum_{j:j\ne l} y_j \boldsymbol{b}_j \boldsymbol{a}_j^{\mathsf{H}}, \tag{A.8}$$

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

Gradient updates: **for** $t = 0, 1, \ldots, t_0 - 1$ **do**

$$\begin{aligned} \begin{bmatrix} \boldsymbol{h}^{t+1/2,(l)} \\ \boldsymbol{x}^{t+1/2,(l)} \end{bmatrix} &= \begin{bmatrix} \boldsymbol{h}^{t,(l)} \\ \boldsymbol{x}^{t,(l)} \end{bmatrix} - \eta \begin{bmatrix} \nabla_{\boldsymbol{h}} f^{(l)}\left(\boldsymbol{h}^t, \boldsymbol{x}^t\right) \\ \nabla_{\boldsymbol{x}} f^{(l)}\left(\boldsymbol{h}^t, \boldsymbol{x}^t\right) \end{bmatrix}, \\ \begin{bmatrix} \boldsymbol{h}^{t+1,(l)} \\ \boldsymbol{x}^{t+1,(l)} \end{bmatrix} &= \begin{bmatrix} \sqrt{\frac{\left\|\boldsymbol{x}^{t+1/2,(l)}\right\|_2}{\left\|\boldsymbol{h}^{t+1/2,(l)}\right\|_2}}\boldsymbol{h}^{t+1/2,(l)} \\ \sqrt{\frac{\left\|\boldsymbol{h}^{t+1/2,(l)}\right\|_2}{\left\|\boldsymbol{x}^{t+1/2,(l)}\right\|_2}}\boldsymbol{x}^{t+1/2,(l)} \end{bmatrix}. \end{aligned} \tag{A.9a}$$

Similar to the notation adopted for the original sequence, we shall define the alignment parameter for the leave-one-out sequence as follows

$$\alpha^{t,(l)} := \arg\min_{\alpha\in\mathbb{C}} \left\{ \left\| \tfrac{1}{\overline{\alpha}}\boldsymbol{h}^{t,(l)} - \boldsymbol{h}^\star \right\|_2^2 + \left\| \alpha\boldsymbol{x}^{t,(l)} - \boldsymbol{x}^\star \right\|_2^2 \right\}, \tag{A.10a}$$

$$\alpha^{t+1/2,(l)} := \arg\min_{\alpha\in\mathbb{C}} \left\{ \left\| \tfrac{1}{\overline{\alpha}}\boldsymbol{h}^{t+1/2,(l)} - \boldsymbol{h}^\star \right\|_2^2 + \left\| \alpha\boldsymbol{x}^{t+1/2,(l)} - \boldsymbol{x}^\star \right\|_2^2 \right\}, \tag{A.10b}$$

along with the properly rescaled iterates

$$\widetilde{\boldsymbol{z}}^{t,(l)} = \begin{bmatrix} \widetilde{\boldsymbol{h}}^{t,(l)} \\ \widetilde{\boldsymbol{x}}^{t,(l)} \end{bmatrix} := \begin{bmatrix} \frac{1}{\overline{\alpha^{t,(l)}}}\boldsymbol{h}^{t,(l)} \\ \alpha^{t,(l)}\boldsymbol{x}^{t,(l)} \end{bmatrix}, \tag{A.11a}$$

$$\widetilde{\boldsymbol{z}}^{t+1/2,(l)} = \begin{bmatrix} \widetilde{\boldsymbol{h}}^{t+1/2,(l)} \\ \widetilde{\boldsymbol{x}}^{t+1/2,(l)} \end{bmatrix} := \begin{bmatrix} \frac{1}{\overline{\alpha^{t+1/2,(l)}}}\boldsymbol{h}^{t+1/2,(l)} \\ \alpha^{t+1/2,(l)}\boldsymbol{x}^{t+1/2,(l)} \end{bmatrix}. \tag{A.11b}$$

Further we define the alignment parameter between $\boldsymbol{z}^{t,(l)}$ and $\widetilde{\boldsymbol{z}}^t$ as

$$\alpha_{\text{mutual}}^{t,(l)} := \arg\min_{\alpha\in\mathbb{C}} \left\{ \left\| \tfrac{1}{\overline{\alpha}}\boldsymbol{h}^{t,(l)} - \tfrac{1}{\overline{\alpha^t}}\boldsymbol{h}^t \right\|_2^2 + \left\| \alpha\boldsymbol{x}^{t,(l)} - \alpha^t\boldsymbol{x}^t \right\|_2^2 \right\}, \tag{A.12a}$$

$$\alpha_{\text{mutual}}^{t+1/2,(l)} := \arg\min_{\alpha\in\mathbb{C}} \left\{ \left\| \tfrac{1}{\overline{\alpha}}\boldsymbol{h}^{t+1/2,(l)} - \tfrac{1}{\overline{\alpha^{t+1/2}}}\boldsymbol{h}^{t+1/2} \right\|_2^2 + \left\| \alpha\boldsymbol{x}^{t+1/2,(l)} - \alpha^{t+1/2}\boldsymbol{x}^{t+1/2} \right\|_2^2 \right\}. \tag{A.12b}$$

Hereafter, we shall also denote

$$\widehat{\boldsymbol{z}}^{t,(l)} := \begin{bmatrix} \widehat{\boldsymbol{h}}^{t,(l)} \\ \widehat{\boldsymbol{x}}^{t,(l)} \end{bmatrix} = \begin{bmatrix} \frac{1}{\overline{\alpha_{\text{mutual}}^{t,(l)}}}\boldsymbol{h}^{t,(l)} \\ \alpha_{\text{mutual}}^{t,(l)}\boldsymbol{x}^{t,(l)} \end{bmatrix}, \tag{A.13a}$$

$$\widehat{\boldsymbol{z}}^{t+1/2,(l)} := \begin{bmatrix} \widehat{\boldsymbol{h}}^{t+1/2,(l)} \\ \widehat{\boldsymbol{x}}^{t+1/2,(l)} \end{bmatrix} = \begin{bmatrix} \frac{1}{\overline{\alpha_{\text{mutual}}^{t+1/2,(l)}}}\boldsymbol{h}^{t+1/2,(l)} \\ \alpha_{\text{mutual}}^{t+1/2,(l)}\boldsymbol{x}^{t+1/2,(l)} \end{bmatrix}. \tag{A.13b}$$

A.2.3 Additional induction hypotheses

In addition to the set of induction hypotheses already listed in (A.6), we find it convenient to include the following hypotheses concerning the leave-one-out sequences. Specifically, for any $0 < t \le t_0$ and any

$1 \le l \le m$, the hypotheses claim that

$$\mathsf{dist}\left(\boldsymbol{z}^{t,(l)}, \widetilde{\boldsymbol{z}}^{t}\right) \le C_2 \left(\frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^9 m}{m}} + \frac{\sigma}{\log^2 m} \right) \tag{A.14a}$$

$$\left\| \widetilde{\boldsymbol{z}}^{t,(l)} - \widetilde{\boldsymbol{z}}^{t} \right\|_2 \lesssim C_2 \left(\frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^9 m}{m}} + \frac{\sigma}{\log^2 m} \right) \tag{A.14b}$$

$$\mathsf{dist}\left(\boldsymbol{z}^{0,(l)}, \boldsymbol{z}^{\star}\right) \lesssim \sqrt{\frac{\mu^2 K \log m}{m}} + \sigma \sqrt{K \log m} \tag{A.14c}$$

$$\mathsf{dist}\left(\boldsymbol{z}^{0,(l)}, \widetilde{\boldsymbol{z}}^{0}\right) \lesssim \frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^5 m}{m}} + \frac{\sigma}{\log^2 m} \tag{A.14d}$$

for some constant $C_2 \gg C_4^2$. Furthermore, there are several immediate consequences of the hypotheses (A.6) and (A.14) that are also useful in the analysis, which we gather as follows. Note that the notation $(\widetilde{\boldsymbol{h}}^t, \widetilde{\boldsymbol{x}}^t)$, $(\widehat{\boldsymbol{h}}^t, \widehat{\boldsymbol{x}}^t)$, $(\widehat{\boldsymbol{h}}^{t,(l)}, \widehat{\boldsymbol{x}}^{t,(l)})$ and α^t has been defined in (A.4b), (A.5b), (A.13a) and (A.2a), respectively.

Lemma 3. *Instate the notation and assumptions in Theorem 2. For $t \ge 0$, suppose that the hypotheses (A.6) and (A.14) hold in the first t iterations. Then there exist some constants $C_1, C > 0$ such that for any $1 \le l \le m$,*

$$\mathsf{dist}\left(\boldsymbol{z}^{t}, \boldsymbol{z}^{\star}\right) \le C_1 \left(\sqrt{\frac{\mu^2 K \log m}{m}} + \lambda + \sigma \sqrt{K \log m} \right), \tag{A.15a}$$

$$\left\| \boldsymbol{h}^t \left(\boldsymbol{x}^t\right)^{\mathsf{H}} - \boldsymbol{h}^{\star} \boldsymbol{x}^{\star \mathsf{H}} \right\| \le C \left(\sqrt{\frac{\mu^2 K \log m}{m}} + \lambda + \sigma \sqrt{K \log m} \right), \tag{A.15b}$$

$$\left\| \widetilde{\boldsymbol{z}}^{t,(l)} - \boldsymbol{z}^{\star} \right\|_2 \le 2C_1 \left(\sqrt{\frac{\mu^2 K \log m}{m}} + \lambda + \sigma \sqrt{K \log m} \right), \tag{A.15c}$$

$$\frac{1}{2} \le \left\| \widetilde{\boldsymbol{x}}^{t} \right\|_2 \le \frac{3}{2}, \qquad \frac{1}{2} \le \left\| \widetilde{\boldsymbol{h}}^{t} \right\|_2 \le \frac{3}{2}, \tag{A.15d}$$

$$\frac{1}{2} \le \left\| \widetilde{\boldsymbol{x}}^{t,(l)} \right\|_2 \le \frac{3}{2}, \qquad \frac{1}{2} \le \left\| \widetilde{\boldsymbol{h}}^{t,(l)} \right\|_2 \le \frac{3}{2}, \tag{A.15e}$$

$$\frac{1}{2} \le \left\| \widehat{\boldsymbol{x}}^{t,(l)} \right\|_2 \le \frac{3}{2}, \qquad \frac{1}{2} \le \left\| \widehat{\boldsymbol{h}}^{t,(l)} \right\|_2 \le \frac{3}{2}. \tag{A.15f}$$

In addition, if $t > 0$, then one also has

$$\left\| \widehat{\boldsymbol{z}}^{t-1/2} - \boldsymbol{z}^{\star} \right\|_2 \le C \left(\sqrt{\frac{\mu^2 K \log m}{m}} + \lambda + \sigma \sqrt{K \log m} \right). \tag{A.15g}$$

Proof. See Appendix A.4. □

A.3 Inductive analysis

In this subsection, we carry out the analysis by induction.

A.3.1 Step 1: Characterizing local geometry

Similar to Ma et al. [2018, Lemma 14], local linear convergence is made possible when some sort of restricted strong convexity and smoothness are present simultaneously. To be specific, define the following squared loss that excludes the regularization term

$$f_{\mathsf{reg\text{-}free}}\left(\boldsymbol{z}\right) = f_{\mathsf{reg\text{-}free}}\left(\boldsymbol{h}, \boldsymbol{x}\right) := \sum_{j=1}^{m} \left| \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h} \boldsymbol{x}^{\mathsf{H}} \boldsymbol{a}_j - y_j \right|^2. \tag{A.16}$$

Our result is this:

Lemma 4. *Let $\delta := c/\log^2 m$ for some sufficiently small constant $c > 0$. Suppose that $m \geq C\mu^2 K \log^9 m$ for some sufficiently large constant $C > 0$ and that $\sigma\sqrt{K \log^5 m} \leq c_1$ for some sufficiently small constant $c_1 > 0$. Then with probability $1 - O\left(m^{-10} + e^{-K} \log m\right)$, one has*

$$\boldsymbol{u}^{\mathsf{H}} \left[\boldsymbol{D}\nabla^2 f\left(\boldsymbol{z}\right) + \nabla^2 f\left(\boldsymbol{z}\right)\boldsymbol{D}\right]\boldsymbol{u} \geq \left\|\boldsymbol{u}\right\|_2^2 / 8 \quad \text{and}$$
$$\left\|\nabla^2 f\left(\boldsymbol{z}\right)\right\| \leq 4$$

simultaneously for all points

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

obeying the following properties:

- $\boldsymbol{z}$ *satisfies*

$$\max\left\{\left\|\boldsymbol{h} - \boldsymbol{h}^\star\right\|_2, \left\|\boldsymbol{x} - \boldsymbol{x}^\star\right\|_2\right\} \leq \delta,$$
$$\max_{1 \leq j \leq m} \left|\boldsymbol{a}_j^{\mathsf{H}}\left(\boldsymbol{x} - \boldsymbol{x}^\star\right)\right| \leq 2C_3 \frac{1}{\log^{3/2} m},$$
$$\max_{1 \leq j \leq m} \left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}\right| \leq 2C_4 \left(\frac{\mu \log^2 m}{\sqrt{m}} + \sigma\right);$$

- $\boldsymbol{z}_1 := (\boldsymbol{h}_1, \boldsymbol{x}_1)$ *is aligned with* $\boldsymbol{z}_2 := (\boldsymbol{h}_2, \boldsymbol{x}_2)$ *in the sense that* $\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|_2 = \mathsf{dist}(\boldsymbol{z}_1, \boldsymbol{z}_2)$*; in addition, they satisfy*

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

- $\gamma_1, \gamma_2 \in \mathbb{R}$ *and obey*

$$\max\left\{\left|\gamma_1 - 1\right|, \left|\gamma_2 - 1\right|\right\} \leq \delta.$$

Proof. See Appendix A.6. □

In words, the function $f(\cdot)$ resembles a strongly convex and smooth function when we restrict attention to (i) a highly restricted set of points $\boldsymbol{z}$ and (ii) a highly special set of directions $\boldsymbol{u}$.

A.3.2 Step 2: ℓ_2 error contraction

Next, we demonstrate that under the hypotheses (A.6) for the tth iteration, the next iterate will undergo ℓ_2 error contraction, as long as the stepsize is properly chosen. The proof is largely based on the restricted strong convexity and smoothness established in Lemma 4.

Lemma 5. *Set $\lambda = C_\lambda \sigma\sqrt{K \log m}$ for some large constant $C_\lambda > 0$. The stepsize parameter $\eta > 0$ in Algorithm 1 is taken to be some sufficiently small constant. There exists some constant $C > 0$ such that with probability at least $1 - O\left(m^{-100} + e^{-CK} \log m\right)$, if the hypotheses (A.6) hold true at the tth iteration, then*

$$\mathsf{dist}\left(\boldsymbol{z}^{t+1}, \boldsymbol{z}^\star\right) \leq \left\|\widehat{\boldsymbol{z}}^{t+1/2} - \boldsymbol{z}^\star\right\|_2 \leq \rho\,\mathsf{dist}\left(\boldsymbol{z}^t, \boldsymbol{z}^\star\right) + C_1 \eta \left(\lambda + \sigma\sqrt{K \log m}\right) \tag{A.17}$$

for some constants $\rho = 1 - \eta/16$ and $C_1 > 0$.

Proof. See Appendix A.7. □

To establish this lemma and many other results, we need to ensure that the alignment parameters and the sizes of the iterates do not change much, as stated below.

Corollary 1. *Instate the notation and assumptions in Theorem 2. For an integer $t > 0$, suppose that the hypotheses (A.6) and (A.14) hold in the first $t-1$ iterations. Then there exists some constant $C > 0$ such that for any $1 \leq l \leq m$, one has*

$$\left| |\alpha^t| - 1 \right| \lesssim \mathsf{dist}\left(\widetilde{\boldsymbol{z}}^t, \boldsymbol{z}^\star\right) \lesssim \sqrt{\frac{\mu^2 K \log m}{m}} + \lambda + \sigma \sqrt{K \log m}, \tag{A.18a}$$

$$\left| \frac{\alpha^{t-1/2}}{\alpha^{t-1}} - 1 \right| \lesssim \eta \left(\sqrt{\frac{\mu^2 K \log m}{m}} + \lambda + \sigma \sqrt{K \log m} \right), \tag{A.18b}$$

$$\left| \left| \alpha_{\text{mutual}}^{t,(l)} \right| - 1 \right| \lesssim \left\| \widetilde{\boldsymbol{z}}^{t,(l)} - \boldsymbol{z}^\star \right\|_2 \lesssim \sqrt{\frac{\mu^2 K \log m}{m}} + \lambda + \sigma \sqrt{K \log m}, \tag{A.18c}$$

$$\frac{1}{2} \leq \left\| \boldsymbol{x}^t \right\|_2 \leq \frac{3}{2}, \qquad \frac{1}{2} \leq \left\| \boldsymbol{h}^t \right\|_2 \leq \frac{3}{2}, \tag{A.18d}$$

$$\frac{1}{2} \leq \left\| \boldsymbol{x}^{t,(l)} \right\|_2 \leq \frac{3}{2}, \qquad \frac{1}{2} \leq \left\| \boldsymbol{h}^{t,(l)} \right\|_2 \leq \frac{3}{2} \tag{A.18e}$$

with probability at least $1 - O\left(m^{-100} + e^{-CK} \log m\right)$.

Proof. See Appendix A.5. ∎

A.3.3 Step 3: Leave-one-out proximity

We then move on to justifying the close proximity of the leave-one-out sequences and the original sequences, as stated in the hypothesis (A.14a).

Lemma 6. *Suppose the sample complexity obeys $m \geq C\mu^2 K \log^9 m$ for some sufficiently large constant $C > 0$. If the hypotheses (A.6a)-(A.6c) hold for the tth iteration, then with probability at least $1 - O\left(m^{-100} + me^{-cK}\right)$ for some constant $c > 0$, one has*

$$\max_{1 \leq l \leq m} \mathsf{dist}\left(\boldsymbol{z}^{t+1,(l)}, \widetilde{\boldsymbol{z}}^{t+1}\right) \leq C_2 \left(\frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^9 m}{m}} + \frac{\sigma}{\log^2 m} \right) \tag{A.19a}$$

and

$$\max_{1 \leq l \leq m} \left\| \widetilde{\boldsymbol{z}}^{t+1,(l)} - \widetilde{\boldsymbol{z}}^{t+1} \right\|_2 \lesssim C_2 \left(\frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^9 m}{m}} + \frac{\sigma}{\log^2 m} \right), \tag{A.19b}$$

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

Proof. See Appendix A.8. ∎

A.3.4 Step 4: Establishing incoherence

The next step is to establish the hypotheses concerning incoherence, namely, (A.6b) and (A.6c) for the $(t+1)$-th iteration.

We start with the incoherence of $\boldsymbol{a}_l$ and $\boldsymbol{x}^{t+1}$, which is much easier to handle. The standard Gaussian concentration inequality gives

$$\max_{1 \leq l \leq m} \left| \boldsymbol{a}_l^{\mathsf{H}} \left(\widetilde{\boldsymbol{x}}^{t+1,(l)} - \boldsymbol{x}^\star \right) \right| \leq 20 \sqrt{\log m} \max_{1 \leq l \leq m} \left\| \widetilde{\boldsymbol{x}}^{t+1,(l)} - \boldsymbol{x}^\star \right\|_2 \tag{A.20}$$

with probability exceeding $1 - O\left(m^{-100}\right)$. Then the triangle inequality and Cauchy-Schwarz inequality yield

$$\begin{aligned}
\left| \boldsymbol{a}_l^{\mathsf{H}} \left(\widetilde{\boldsymbol{x}}^{t+1} - \boldsymbol{x}^\star \right) \right| &\leq \left| \boldsymbol{a}_l^{\mathsf{H}} \left(\widetilde{\boldsymbol{x}}^{t+1} - \widetilde{\boldsymbol{x}}^{t+1,(l)} \right) \right| + \left| \boldsymbol{a}_l^{\mathsf{H}} \left(\widetilde{\boldsymbol{x}}^{t+1,(l)} - \boldsymbol{x}^\star \right) \right| \\
&\leq \left\| \boldsymbol{a}_l \right\|_2 \left\| \widetilde{\boldsymbol{x}}^{t+1} - \widetilde{\boldsymbol{x}}^{t+1,(l)} \right\|_2 + \left| \boldsymbol{a}_l^{\mathsf{H}} \left(\widetilde{\boldsymbol{x}}^{t+1,(l)} - \boldsymbol{x}^\star \right) \right|
\end{aligned}$$

$$\leq 10\sqrt{K}C_2\left(\frac{\mu}{\sqrt{m}}\sqrt{\frac{\mu^2K\log^9 m}{m}}+\frac{\sigma}{\log^2 m}\right)$$
$$+20\sqrt{\log m}\cdot 2C_1\left(\sqrt{\frac{\mu^2K\log m}{m}}+\lambda+\sigma\sqrt{K\log m}\right)$$
$$\leq C_3\left(\sqrt{\frac{\mu^2K\log^2 m}{m}}+\lambda+\sigma\sqrt{K}\log m\right), \tag{A.21}$$

where $C_3 \gg C_1$, the penultimate inequality follows from (F.2), (A.19b), (A.20) and (A.15c). This establishes the hypothesis (A.6b) for the $(t+1)$-th iteration.

Regarding the incoherence of $\boldsymbol{b}_l$ and $\boldsymbol{h}^{t+1}$ (as stated in the hypothesis (A.6c)), we have the following lemma.

Lemma 7. *Suppose the sample complexity obeys $m \geq C\mu^2K\log^9 m$ for some sufficiently large constant $C > 0$ and $\lambda = C_\lambda\sigma\sqrt{K\log m}$ for some absolute constant $C_\lambda > 0$. If the hypotheses (A.6a)-(A.6c) hold for the tth iteration, then with probability exceeding $1 - O\left(m^{-100} + me^{-CK}\right)$ for some constant $C > 0$, one has*

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

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

Proof. See Appendix A.9. ☐

A.3.5 The base case: Spectral initialization

To finish the induction analysis, it remains to justify the induction hypotheses for the base case. Recall that $\sigma(\boldsymbol{M}), \check{\boldsymbol{h}}^0$ and $\check{\boldsymbol{x}}^0$ denote respectively the leading singular value, the left and the right singular vectors of

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

The spectral initialization procedure sets $\boldsymbol{h}^0 = \sqrt{\sigma_1(\boldsymbol{M})}\check{\boldsymbol{h}}^0$ and $\boldsymbol{x}^0 = \sqrt{\sigma_1(\boldsymbol{M})}\check{\boldsymbol{x}}^0$.

To begin with, the following lemma guarantees that $\left(\boldsymbol{h}^0, \boldsymbol{x}^0\right)$ satisfies the desired conditions (A.6e) and (A.6h).

Lemma 8. *Suppose the sample size obeys $m \geq C\mu^2K\log^4 m$ for some sufficiently large constant $C > 0$. Then with probability at least $1 - O\left(m^{-100}\right)$, we have*

$$\min_{\alpha\in\mathbb{C},|\alpha|=1}\left\{\left\|\alpha\boldsymbol{h}^0-\boldsymbol{h}^\star\right\|_2+\left\|\alpha\boldsymbol{x}^0-\boldsymbol{x}^\star\right\|_2\right\} \lesssim \sqrt{\frac{\mu^2K\log m}{m}}+\sigma\sqrt{K\log m}$$

and $\left|\left|\alpha^0\right|-1\right| \leq 1/4$.

In view of the definition of $\mathsf{dist}(\cdot,\cdot)$, we can invoke Lemma 8 to reach

$$\mathsf{dist}\left(\boldsymbol{z}^0,\boldsymbol{z}^\star\right) = \min_{\alpha\in\mathbb{C}}\sqrt{\left\|\tfrac{1}{\overline{\alpha}}\boldsymbol{h}^0-\boldsymbol{h}^\star\right\|_2^2+\left\|\alpha\boldsymbol{x}^0-\boldsymbol{x}^\star\right\|_2^2} \leq \min_{\alpha\in\mathbb{C}}\left\{\left\|\tfrac{1}{\overline{\alpha}}\boldsymbol{h}^0-\boldsymbol{h}^\star\right\|_2+\left\|\alpha\boldsymbol{x}^0-\boldsymbol{x}^\star\right\|_2\right\}$$
$$\leq \min_{\alpha\in\mathbb{C},|\alpha|=1}\left\{\left\|\alpha\boldsymbol{h}^0-\boldsymbol{h}^\star\right\|_2+\left\|\alpha\boldsymbol{x}^0-\boldsymbol{x}^\star\right\|_2\right\} \leq C_1\left(\sqrt{\frac{\mu^2K\log m}{m}}+\sigma\sqrt{K\log m}\right). \tag{A.22}$$

Repeating the same arguments yields that, with probability exceeding $1 - O(m^{-20})$,

$$\mathsf{dist}\left(\boldsymbol{z}^{0,(l)},\boldsymbol{z}^\star\right) \leq C_1\left(\sqrt{\frac{\mu^2K\log m}{m}}+\sigma\sqrt{K\log m}\right), \qquad 1\leq l\leq m, \tag{A.23}$$

and $\left|\left|\alpha^{0,(l)}\right|-1\right| \leq 1/4$, as asserted in the hypothesis (A.14c).

The following lemma justifies (A.14d) as well as (A.6c) for the base case.

Lemma 9. *Suppose the sample size obeys $m \geq C\mu^2 K \log^9 m$ for some sufficiently large constant $C > 0$ and the noise satisfies $\sigma\sqrt{K \log m} \leq c/\log^2 m$ for some sufficiently small constant $c > 0$. Let $\tau = C_\tau \log^4 m$ for some sufficiently large constant $C_\tau > 0$ such that τ is an integer. Then with probability at least $1 - O\left(m^{-100} + me^{-cK}\right)$ for some constant $c > 0$, we have*

$$\max_{1\leq l\leq m} \mathsf{dist}\big(\boldsymbol{z}^{0,(l)}, \widetilde{\boldsymbol{z}}^0\big) \lesssim \frac{\mu}{\sqrt{m}}\sqrt{\frac{\mu^2 K \log^5 m}{m}} + \frac{\sigma}{\log^2 m}, \tag{A.24a}$$

$$\max_{1\leq l\leq m} \left|\boldsymbol{b}_l^{\mathsf{H}}\widetilde{\boldsymbol{h}}^0\right| \lesssim \frac{\mu \log^2 m}{\sqrt{m}} + \sigma, \tag{A.24b}$$

$$\max_{1\leq j\leq \tau} \left|\left(\boldsymbol{b}_j - \boldsymbol{b}_1\right)^{\mathsf{H}}\widetilde{\boldsymbol{h}}^0\right| \lesssim \frac{\mu}{\sqrt{m}}\frac{1}{\log m} + \frac{\sigma}{\log m}. \tag{A.24c}$$

Finally, we establish the hypothesis (A.6b) for the base case, which concerns the incoherence of $\boldsymbol{x}^0$ with respect to the design vectors $\{\boldsymbol{a}_l\}$.

Lemma 10. *Suppose the sample size obeys $m \geq C\mu^2 K \log^6 m$ for some sufficiently large constant $C > 0$ and $\sigma\sqrt{K \log^5 m} \leq c$ for some small constant $c > 0$. Then with probability at least $1 - O\left(m^{-100+} me^{-c_2 K}\right)$ for some constant $c_2 > 0$, we have*

$$\max_{1\leq j\leq m} \left|\boldsymbol{a}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^0 - \boldsymbol{x}^\star\right)\right| \lesssim \sqrt{\frac{\mu^2 K \log^2 m}{m}} + \sigma\sqrt{K}\log m.$$

The proof of these three lemmas can be easily obtained via straightforward modifications to Ma et al. [2018, Lemmas 19,20,21]; we omit the details here for the sake of brevity.

A.3.6 Proof of Theorem 2

With the above results in place, it is straightforward to prove Theorem 2. The first two claims follows respectively from (A.22) and (A.6d). Regarding (12c), it follows that

$$\begin{aligned}
\left\|\boldsymbol{h}^t\left(\boldsymbol{x}^t\right)^{\mathsf{H}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}} &\leq \left\|\boldsymbol{h}^t\left(\boldsymbol{x}^t\right)^{\mathsf{H}} - \boldsymbol{h}^\star\left(\boldsymbol{x}^t\right)^{\mathsf{H}}\right\|_{\mathrm{F}} + \left\|\boldsymbol{h}^\star\left(\boldsymbol{x}^t\right)^{\mathsf{H}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}} \\
&\leq \left\|\boldsymbol{h}^t - \boldsymbol{h}^\star\right\|_2\left\|\boldsymbol{x}^t\right\|_2 + \left\|\boldsymbol{h}^\star\right\|_2\left\|\boldsymbol{x}^t - \boldsymbol{x}^\star\right\|_2 \\
&\leq 2\left\|\boldsymbol{z}^\star\right\|_2\left(\rho^t \mathsf{dist}\left(\boldsymbol{z}^0, \boldsymbol{z}^\star\right) + \frac{C_1\left(\lambda + \sigma\sqrt{K\log m}\right)}{c_\rho\left\|\boldsymbol{z}^\star\right\|_2}\right)
\end{aligned}$$

where the last inequality follows from (A.6d) and the fact that

$$\left\|\boldsymbol{x}^t\right\|_2 \leq \left\|\boldsymbol{x}^\star\right\|_2 + \left\|\boldsymbol{x}^t - \boldsymbol{x}^\star\right\|_2 \leq \left\|\boldsymbol{z}^\star\right\|_2 + \rho^t \mathsf{dist}\left(\boldsymbol{z}^0, \boldsymbol{z}^\star\right) + \frac{C_1\left(\lambda + \sigma\sqrt{K\log m}\right)}{c_\rho\left\|\boldsymbol{z}^\star\right\|_2} \leq 2\left\|\boldsymbol{z}^\star\right\|_2.$$

This concludes the proof.

A.4 Proof of Lemma 3

1. Condition (A.15a) follows directly from the ℓ_2 contraction (A.6a) and the bound (A.6e) for the base case.

2. (A.15b) is direct consequence of (A.15a) and triangle inequality. We have

$$\begin{aligned}
\left\|\boldsymbol{h}^t\boldsymbol{x}^{t\mathsf{H}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}} &= \left\|\widetilde{\boldsymbol{h}}^t\widetilde{\boldsymbol{x}}^{t\mathsf{H}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}} \\
&\leq \left\|\widetilde{\boldsymbol{h}}^t\widetilde{\boldsymbol{x}}^{t\mathsf{H}} - \widetilde{\boldsymbol{h}}^t\boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}} + \left\|\widetilde{\boldsymbol{h}}^t\boldsymbol{x}^{\star\mathsf{H}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}} \\
&\leq \left\|\widetilde{\boldsymbol{h}}^t\right\|_2\left\|\widetilde{\boldsymbol{x}}^t - \boldsymbol{x}^\star\right\|_2 + \left\|\widetilde{\boldsymbol{h}}^t - \boldsymbol{h}^\star\right\|_2\left\|\boldsymbol{x}^\star\right\|_2
\end{aligned}$$

$$\leq \left(1 + \mathsf{dist}\left(\boldsymbol{z}^t, \boldsymbol{z}^\star\right)\right) \mathsf{dist}\left(\boldsymbol{z}^t, \boldsymbol{z}^\star\right) + \mathsf{dist}\left(\boldsymbol{z}^t, \boldsymbol{z}^\star\right)$$
$$\leq C \left(\sqrt{\frac{\mu^2 K \log m}{m}} + \lambda + \sigma \sqrt{K \log m} \right),$$

where the first equality follows from the definitions of $\widetilde{\boldsymbol{h}}^t$ and $\widetilde{\boldsymbol{x}}^t$ (cf. (A.4b)) and $C > 0$ is some sufficiently large constant.

3. Regarding (A.15c), it follows from the triangle inequality that

$$\max_{1 \leq l \leq m} \left\| \widetilde{\boldsymbol{z}}^{t,(l)} - \boldsymbol{z}^\star \right\|_2 \leq \max_{1 \leq l \leq m} \left\{ \left\| \widetilde{\boldsymbol{z}}^{t,(l)} - \widetilde{\boldsymbol{z}}^t \right\|_2 + \left\| \widetilde{\boldsymbol{z}}^t - \boldsymbol{z}^\star \right\|_2 \right\}$$
$$\leq \widetilde{C} C_2 \left(\frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^9 m}{m}} + \frac{\sigma}{\log^2 m} \right) + C_1 \left(\sqrt{\frac{\mu^2 K \log m}{m}} + \lambda + \sigma \sqrt{K \log m} \right)$$
$$\leq 2 C_1 \left(\sqrt{\frac{\mu^2 K \log m}{m}} + \lambda + \sigma \sqrt{K \log m} \right)$$

for $t > 0$. Here, the penultimate inequality follows from the distance bounds (A.14b) and (A.15a), while the last inequality holds as long as $m \geq C \mu^2 \log^8 m$ for some sufficiently large constant $C > 0$. The base case follows from (A.14c).

4. Condition (A.15d) immediately results from (A.15a), the assumption $\|\boldsymbol{x}^\star\|_2 = \|\boldsymbol{h}^\star\|_2 = 1$, the definition of $\mathsf{dist}(\cdot, \cdot)$, and the triangle inequality.

5. With regards to (A.15e) and (A.15f), we shall only provide the proof for the result concerning $\boldsymbol{h}$; the result concerning $\boldsymbol{x}$ can be derived analogously. In terms of (A.15f), one has

$$\left\| \widehat{\boldsymbol{h}}^{t,(l)} \right\|_2 \leq \left\| \widetilde{\boldsymbol{h}}^t \right\|_2 + \left\| \widehat{\boldsymbol{h}}^{t,(l)} - \widetilde{\boldsymbol{h}}^t \right\|_2 = \left\| \widetilde{\boldsymbol{h}}^t \right\|_2 + \mathsf{dist}\left(\boldsymbol{h}^{t,(l)}, \widetilde{\boldsymbol{h}}^t\right)$$
$$\lesssim 1 + C_2 \left(\sqrt{\frac{\mu^4 K \log^9 m}{m^2}} + \frac{\sigma}{\log^2 m} \right) \asymp 1.$$

Here, the first line comes from triangle inequality as well as the definitions of $\widehat{\boldsymbol{h}}^{t,(l)}$ and $\widetilde{\boldsymbol{h}}^t$, whereas the last inequality comes from (A.14a). A lower bound can be derived in a similar manner:

$$\left\| \widehat{\boldsymbol{h}}^{t,(l)} \right\|_2 \geq \left\| \widetilde{\boldsymbol{h}}^t \right\|_2 - \left\| \widehat{\boldsymbol{h}}^{t,(l)} - \widetilde{\boldsymbol{h}}^t \right\|_2 \gtrsim 1 - C_2 \left(\sqrt{\frac{\mu^4 K \log^9 m}{m^2}} + \frac{\sigma}{\log^2 m} \right) \asymp 1.$$

Regarding (A.15e), apply (A.14b) and (A.15d) to obtain

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

and, similarly,

$$\left\| \widetilde{\boldsymbol{h}}^{t,(l)} \right\|_2 \geq \left\| \widetilde{\boldsymbol{h}}^t \right\|_2 - \left\| \widetilde{\boldsymbol{h}}^{t,(l)} - \widetilde{\boldsymbol{h}}^t \right\|_2 \gtrsim 1 - C_2 \left(\frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^9 m}{m}} + \frac{\sigma}{\log^2 m} \right) \asymp 1.$$

The base case follows from similar deduction using (A.14d), (A.15d) and triangle inequality.

6. When it comes to Condition (A.15g), it is seen from (A.6a) and the choice $\rho = 1 - c_\rho \eta$ that

$$\left\| \widetilde{\boldsymbol{z}}^{t-1/2} - \boldsymbol{z}^\star \right\|_2 \leq \rho^t \mathsf{dist}\left(\boldsymbol{z}^0, \boldsymbol{z}^\star\right) + \frac{C_1}{1 - \rho} \eta \left(\lambda + \sigma \sqrt{K \log m} \right)$$

$$= \rho^t \mathsf{dist}\left(\boldsymbol{z}^0, \boldsymbol{z}^\star\right) + \frac{C_1}{c_\rho}\left(\lambda + \sigma\sqrt{K\log m}\right).$$

Combining this with (A.6e) guarantees the existence of some sufficiently large constant $\widetilde{C} > 0$ such that

$$\left\|\widehat{\boldsymbol{z}}^{t-1/2} - \boldsymbol{z}^\star\right\|_2 \le \rho^t \cdot \widetilde{C}\left(\sqrt{\frac{\mu^2 K\log m}{m}} + \sigma\sqrt{K\log m}\right) + \frac{C_1}{c_\rho}\left(\lambda + \sigma\sqrt{K\log m}\right)$$
$$\le C\left(\sqrt{\frac{\mu^2 K\log m}{m}} + \lambda + \sigma\sqrt{K\log m}\right),$$

provided that the constant $C > 0$ is large enough.

A.5 Proof of Corollary 1

1. To establish (A.18a), we recall that the balancing operation (6b) guarantees $\|\boldsymbol{h}^t\|_2 = \|\boldsymbol{x}^t\|_2$. Hence, in view of the definitions of $\widetilde{\boldsymbol{h}}^t$ and $\widetilde{\boldsymbol{x}}^t$ in (A.4b), we have

$$0 = \left\|\boldsymbol{h}^t\right\|_2^2 - \left\|\boldsymbol{x}^t\right\|_2^2 = \left|\alpha^t\right|^2\left\|\widetilde{\boldsymbol{h}}^t\right\|_2^2 - \frac{1}{\left|\alpha^t\right|^2}\left\|\widetilde{\boldsymbol{x}}^t\right\|_2^2.$$

It then follows from the triangle inequality and the assumption $\|\boldsymbol{x}^\star\|_2 = \|\boldsymbol{h}^\star\|_2$ that

$$0 = \left|\alpha^t\right|^2\left\|\widetilde{\boldsymbol{h}}^t\right\|_2^2 - \frac{1}{\left|\alpha^t\right|^2}\left\|\widetilde{\boldsymbol{x}}^t\right\|_2^2 \le \left|\alpha^t\right|^2\left(1 + \left\|\widetilde{\boldsymbol{h}}^t - \boldsymbol{h}^\star\right\|_2\right)^2 - \frac{\left(1 - \left\|\widetilde{\boldsymbol{x}}^t - \boldsymbol{x}^\star\right\|_2\right)^2}{\left|\alpha^t\right|^2};$$
$$0 = \left|\alpha^t\right|^2\left\|\widetilde{\boldsymbol{h}}^t\right\|_2^2 - \frac{1}{\left|\alpha^t\right|^2}\left\|\widetilde{\boldsymbol{x}}^t\right\|_2^2 \ge \left|\alpha^t\right|^2\left(1 - \left\|\widetilde{\boldsymbol{h}}^t - \boldsymbol{h}^\star\right\|_2\right)^2 - \frac{\left(1 + \left\|\widetilde{\boldsymbol{x}}^t - \boldsymbol{x}^\star\right\|_2\right)^2}{\left|\alpha^t\right|^2}.$$

Rearranging terms, we are left with

$$\sqrt{\frac{1 - \left\|\widetilde{\boldsymbol{x}}^t - \boldsymbol{x}^\star\right\|_2}{1 + \left\|\widetilde{\boldsymbol{h}}^t - \boldsymbol{h}^\star\right\|_2}} \le \left|\alpha^t\right| \le \sqrt{\frac{1 + \left\|\widetilde{\boldsymbol{x}}^t - \boldsymbol{x}^\star\right\|_2}{1 - \left\|\widetilde{\boldsymbol{h}}^t - \boldsymbol{h}^\star\right\|_2}}.$$

Combining this with (A.15a), we arrive at

$$\left|\left|\alpha^t\right| - 1\right| \lesssim \left\|\widetilde{\boldsymbol{x}}^t - \boldsymbol{x}^\star\right\|_2 + \left\|\widetilde{\boldsymbol{h}}^t - \boldsymbol{h}^\star\right\|_2 \lesssim \mathsf{dist}\big(\widetilde{\boldsymbol{z}}^t, \boldsymbol{z}^\star\big) \le C_1\left(\sqrt{\frac{\mu^2 K\log m}{m}} + \lambda + \sigma\sqrt{K\log m}\right).$$

2. Regarding (A.18a), take $\boldsymbol{x}_1 = \alpha^{t-1}\boldsymbol{x}^{t-1/2}$, $\boldsymbol{h}_1 = \boldsymbol{h}^{t-1/2}/\overline{\alpha^{t-1}}$, $\boldsymbol{x}_2 = \alpha^{t-1}\boldsymbol{x}^{t-1}$ and $\boldsymbol{h}_2 = \boldsymbol{h}^{t-1}/\overline{\alpha^{t-1}}$. Then we check that these vectors satisfy the conditions of Ma et al. [2018, Lemma 54]. Towards this, observe that

$$\max\left\{\left\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\right\|_2, \left\|\boldsymbol{h}_1 - \boldsymbol{h}^\star\right\|_2, \left\|\boldsymbol{x}_2 - \boldsymbol{x}^\star\right\|_2, \left\|\boldsymbol{h}_2 - \boldsymbol{h}^\star\right\|_2\right\}$$
$$\le \max\left\{\left\|\widehat{\boldsymbol{z}}^{t-1/2} - \boldsymbol{z}^\star\right\|_2, \mathsf{dist}\left(\boldsymbol{z}^{t-1}, \boldsymbol{z}^\star\right)\right\}$$
$$\lesssim \sqrt{\frac{\mu^2 K\log m}{m}} + \lambda + \sigma\sqrt{K\log m}$$

holds with probability over $1 - O(m^{-100} + e^{-CK}\log m)$ for some constant $C > 0$. Here, the first inequality comes from the definitions of $\widehat{\boldsymbol{z}}^{t-1/2}$ (cf. (A.5a)), and the last inequality follows from (A.15a) and (A.17). Hence, the condition of Ma et al. [2018, Lemma 54] is satisfied. Note that the statement of Ma et al. [2018, Lemma 54] involves two quantities α_1 and α_2, which in our case are given by $\alpha_1 = \alpha^{t-1/2}/\alpha^{t-1}$ and $\alpha_2 = 1$. Ma et al. [2018, Lemma 54] tells us that

$$\left|\alpha_1 - \alpha_2\right| = \left|\frac{\alpha^{t-1/2}}{\alpha^{t-1}} - 1\right| \lesssim \left\|\alpha^{t-1}\boldsymbol{x}^{t-1/2} - \alpha^{t-1}\boldsymbol{x}^{t-1}\right\|_2 + \left\|\frac{\boldsymbol{h}^{t-1/2}}{\overline{\alpha^{t-1}}} - \frac{\boldsymbol{h}^{t-1}}{\overline{\alpha^{t-1}}}\right\|_2.$$

Additionally, the gradient update rule (6a) reveals that

$$\begin{aligned}
&\left\| \begin{bmatrix} \frac{\boldsymbol{h}^{t-1/2}}{\overline{\alpha^{t-1}}} - \frac{\boldsymbol{h}^{t-1}}{\overline{\alpha^{t-1}}} \\ \alpha^{t-1}\boldsymbol{x}^{t-1/2} - \alpha^{t-1}\boldsymbol{x}^{t-1} \end{bmatrix} \right\|_2 \\
&= \left\| \begin{bmatrix} -\frac{\eta}{|\alpha^{t-1}|^2}\nabla_{\boldsymbol{h}} f_{\text{reg-free}}\left(\widetilde{\boldsymbol{z}}^{t-1}\right) - \eta\lambda\widetilde{\boldsymbol{h}}^{t-1} \\ -\eta\left|\alpha^{t-1}\right|^2\nabla_{\boldsymbol{x}} f_{\text{reg-free}}\left(\widetilde{\boldsymbol{z}}^{t-1}\right) - \eta\lambda\widetilde{\boldsymbol{x}}^{t-1} \end{bmatrix} \right\|_2 \\
&= \left\| \begin{bmatrix} -\frac{\eta}{|\alpha^{t-1}|^2}\left(\nabla_{\boldsymbol{h}} f_{\text{reg-free}}\left(\widetilde{\boldsymbol{z}}^{t-1}\right) - \nabla_{\boldsymbol{h}} f_{\text{reg-free}}\left(\boldsymbol{z}^{\star}\right)\right) - \eta\lambda\widetilde{\boldsymbol{h}}^{t-1} - \frac{\eta}{|\alpha^{t-1}|^2}\nabla_{\boldsymbol{h}} f_{\text{reg-free}}\left(\boldsymbol{z}^{\star}\right) \\ -\eta\left|\alpha^{t-1}\right|^2\left(\nabla_{\boldsymbol{x}} f_{\text{reg-free}}\left(\widetilde{\boldsymbol{z}}^{t-1}\right) - \nabla_{\boldsymbol{x}} f_{\text{reg-free}}\left(\boldsymbol{z}^{\star}\right)\right) - \eta\lambda\widetilde{\boldsymbol{x}}^{t-1} - \eta\left|\alpha^{t-1}\right|^2\nabla_{\boldsymbol{x}} f_{\text{reg-free}}\left(\boldsymbol{z}^{\star}\right) \end{bmatrix} \right\|_2 \\
&\le \left\| \begin{bmatrix} \frac{\eta}{|\alpha^{t-1}|^2}\left(\nabla_{\boldsymbol{h}} f_{\text{reg-free}}\left(\widetilde{\boldsymbol{z}}^{t-1}\right) - \nabla_{\boldsymbol{h}} f_{\text{reg-free}}\left(\boldsymbol{z}^{\star}\right)\right) \\ \eta\left|\alpha^{t-1}\right|^2\left(\nabla_{\boldsymbol{x}} f_{\text{reg-free}}\left(\widetilde{\boldsymbol{z}}^{t-1}\right) - \nabla_{\boldsymbol{x}} f_{\text{reg-free}}\left(\boldsymbol{z}^{\star}\right)\right) \end{bmatrix} \right\|_2 + \left\| \begin{bmatrix} \eta\lambda\widetilde{\boldsymbol{h}}^{t-1} \\ \eta\lambda\widetilde{\boldsymbol{x}}^{t-1} \end{bmatrix} \right\|_2 \\
&\quad + \left\| \begin{bmatrix} \frac{\eta}{|\alpha^{t-1}|^2}\nabla_{\boldsymbol{h}} f_{\text{reg-free}}\left(\boldsymbol{z}^{\star}\right) \\ \eta\left|\alpha^{t-1}\right|^2\nabla_{\boldsymbol{x}} f_{\text{reg-free}}\left(\boldsymbol{z}^{\star}\right) \end{bmatrix} \right\|_2 \\
&\le 4\eta\left\|\nabla f_{\text{reg}-\text{free}}\left(\widetilde{\boldsymbol{z}}^{t-1}\right) - \nabla f_{\text{reg}-\text{free}}\left(\boldsymbol{z}^{\star}\right)\right\|_2 + \eta\lambda\left\|\widetilde{\boldsymbol{z}}^{t-1}\right\|_2 + 4\eta\left\|\nabla f_{\text{reg}-\text{free}}\left(\boldsymbol{z}^{\star}\right)\right\|_2,
\end{aligned}$$

where the last inequality utilizes the consequence of (A.18a) that

$$\frac{1}{2} \le 1 - \left|\left|\alpha^{t-1}\right| - 1\right| \le \left|\alpha^{t-1}\right| \le 1 + \left|\left|\alpha^{t-1}\right| - 1\right| \le 2.$$

Then, one has

$$\begin{bmatrix} \nabla f_{\text{reg-free}}\left(\widetilde{\boldsymbol{z}}^{t-1}\right) - \nabla f_{\text{reg-free}}\left(\boldsymbol{z}^{\star}\right) \\ \overline{\nabla f_{\text{reg-free}}\left(\widetilde{\boldsymbol{z}}^{t-1}\right) - \nabla f_{\text{reg-free}}\left(\boldsymbol{z}^{\star}\right)} \end{bmatrix} = \int_0^1 \nabla^2 f_{\text{reg-free}}\left(\boldsymbol{z}\left(s\right)\right) \mathrm{d}s \begin{bmatrix} \widetilde{\boldsymbol{z}}^t - \boldsymbol{z}^{\star} \\ \overline{\widetilde{\boldsymbol{z}}^t - \boldsymbol{z}^{\star}} \end{bmatrix},$$

where $\boldsymbol{z}\left(s\right) = \boldsymbol{z}^{\star} + s\left(\widetilde{\boldsymbol{z}}^t - \boldsymbol{z}^{\star}\right)$. Therefore, for all $0 \le s \le 1$ we have

$$\begin{aligned}
\max\left\{\left\|\boldsymbol{h}\left(s\right) - \boldsymbol{h}^{\star}\right\|_2, \left\|\boldsymbol{x}\left(s\right) - \boldsymbol{x}^{\star}\right\|_2\right\} &\le \frac{c}{\log^2 m}, \\
\max_{1\le j\le m}\left|\boldsymbol{a}_j^{\mathsf{H}}\left(\boldsymbol{x}\left(s\right) - \boldsymbol{x}^{\star}\right)\right| &\le 2C_3\tfrac{1}{\log^{3/2} m}, \\
\max_{1\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}\left(s\right)\right| &\le 2C_4\left(\tfrac{\mu\log^2 m}{\sqrt{m}} + \sigma\right),
\end{aligned}$$

which are guaranteed by the induction hypotheses (A.6). The conditions of Lemma (4) are satisfied, allowing us to obtain

$$\left\|\int_0^1 \nabla^2 f_{\text{reg-free}}\left(\boldsymbol{z}\left(s\right)\right)\mathrm{d}s\right\| \le \left\|\int_0^1 \nabla^2 f\left(\boldsymbol{z}\left(s\right)\right)\mathrm{d}s\right\| + \lambda \le 4 + \lambda \le 5.$$

Consequently, it follows that

$$\begin{aligned}
\left\| \begin{bmatrix} \frac{\boldsymbol{h}^{t-1/2}}{\overline{\alpha^{t-1}}} - \frac{\boldsymbol{h}^{t-1}}{\overline{\alpha^{t-1}}} \\ \alpha^{t-1}\boldsymbol{x}^{t-1/2} - \alpha^{t-1}\boldsymbol{x}^{t-1} \end{bmatrix} \right\|_2 &\le 20\eta\left\|\widetilde{\boldsymbol{z}}^{t-1} - \boldsymbol{z}^{\star}\right\|_2 + \eta\lambda\left\|\widetilde{\boldsymbol{z}}^{t-1}\right\|_2 + 4\eta\left\|\nabla f_{\text{reg-free}}\left(\boldsymbol{z}^{\star}\right)\right\|_2 \\
&\le C\eta\left(\sqrt{\frac{\mu^2 K\log m}{m}} + \lambda + \sigma\sqrt{K\log m}\right),
\end{aligned}$$

where the last inequality results from (A.15a), (A.15d), and (A.31). Hence, we arrive at

$$\left|\frac{\alpha^{t-1/2}}{\alpha^{t-1}} - 1\right| \lesssim \eta\left(\sqrt{\frac{\mu^2 K\log m}{m}} + \lambda + \sigma\sqrt{K\log m}\right).$$

3. Similarly, the balancing step (A.9a) implies $\|\boldsymbol{h}^{t,(l)}\|_2^2 = \|\boldsymbol{x}^{t,(l)}\|_2^2$. From the definitions of $\alpha_{\text{mutual}}^{t,(l)}$ (cf. (A.12a)), $\widehat{\boldsymbol{h}}^{t,(l)}$ and $\widehat{\boldsymbol{x}}^{t,(l)}$ (cf. (A.13a)), we have

$$0 = \|\boldsymbol{h}^{t,(l)}\|_2^2 - \|\boldsymbol{x}^{t,(l)}\|_2^2 = |\alpha_{\text{mutual}}^{t,(l)}|^2 \|\widehat{\boldsymbol{h}}^{t,(l)}\|_2^2 - |\alpha_{\text{mutual}}^{t,(l)}|^{-2} \|\widehat{\boldsymbol{x}}^{t,(l)}\|_2^2.$$

Then the triangle inequality together with the assumption $\|\boldsymbol{x}^\star\|_2 = \|\boldsymbol{h}^\star\|_2$ gives

$$0 = |\alpha_{\text{mutual}}^{t,(l)}|^2 \|\widehat{\boldsymbol{h}}^{t,(l)}\|_2^2 - \frac{1}{|\alpha_{\text{mutual}}^{t,(l)}|^2} \|\widehat{\boldsymbol{x}}^{t,(l)}\|_2^2 \leq |\alpha_{\text{mutual}}^{t,(l)}|^2 \left(1 + \|\widehat{\boldsymbol{h}}^{t,(l)} - \boldsymbol{h}^\star\|_2\right)^2 - \frac{\left(1 - \|\widehat{\boldsymbol{x}}^{t,(l)} - \boldsymbol{x}^\star\|_2\right)^2}{|\alpha_{\text{mutual}}^{t,(l)}|^2},$$

$$0 = |\alpha_{\text{mutual}}^{t,(l)}|^2 \|\widehat{\boldsymbol{h}}^{t,(l)}\|_2^2 - \frac{1}{|\alpha_{\text{mutual}}^{t,(l)}|^2} \|\widehat{\boldsymbol{x}}^{t,(l)}\|_2^2 \geq |\alpha_{\text{mutual}}^{t,(l)}|^2 \left(1 - \|\widehat{\boldsymbol{h}}^{t,(l)} - \boldsymbol{h}^\star\|_2\right)^2 - \frac{\left(1 + \|\widehat{\boldsymbol{x}}^{t,(l)} - \boldsymbol{x}^\star\|_2\right)^2}{|\alpha_{\text{mutual}}^{t,(l)}|^2},$$

which in turn lead to

$$\sqrt{\frac{1 - \|\widehat{\boldsymbol{x}}^{t,(l)} - \boldsymbol{x}^\star\|_2}{1 + \|\widehat{\boldsymbol{h}}^{t,(l)} - \boldsymbol{h}^\star\|_2}} \leq |\alpha_{\text{mutual}}^{t,(l)}| \leq \sqrt{\frac{1 + \|\widehat{\boldsymbol{x}}^{t,(l)} - \boldsymbol{x}^\star\|_2}{1 - \|\widehat{\boldsymbol{h}}^{t,(l)} - \boldsymbol{h}^\star\|_2}}.$$

Taking this together with (A.14a) and (A.15a), we reach

$$\begin{aligned}
\left| |\alpha_{\text{mutual}}^{t,(l)}| - 1 \right| &\lesssim \|\widehat{\boldsymbol{z}}^{t,(l)} - \boldsymbol{z}^\star\|_2 \leq \|\widehat{\boldsymbol{z}}^{t,(l)} - \widetilde{\boldsymbol{z}}^t\|_2 + \|\widetilde{\boldsymbol{z}}^t - \boldsymbol{z}^\star\|_2 \\
&\leq C_2 \left(\sqrt{\frac{\mu^4 K \log^9 m}{m^2}} + \frac{\sigma}{\log^2 m} \right) + C_1 \left(\sqrt{\frac{\mu^2 K \log m}{m}} + \lambda + \sigma \sqrt{K \log m} \right) \\
&\leq (C_1 + C_2) \left(\sqrt{\frac{\mu^2 K \log m}{m}} + \lambda + \sigma \sqrt{K \log m} \right),
\end{aligned}$$

where the second line follows from the distance bounds (A.14a) and (A.15a), and the last line holds with the proviso that $m \geq \mu^2 K \log^8 m$. This establishes the claim (A.18c).

4. Finally, (A.18d) and (A.18e) are direct consequences of (A.18a), (A.18c) as well as the fact that $\|\boldsymbol{h}^\star\|_2 = \|\boldsymbol{x}^\star\|_2 = 1$. We omit the details for the sake of brevity.

A.6 Proof of Lemma 4

Define another loss function as follows

$$f_{\mathsf{clean}}(\boldsymbol{z}) := \sum_{j=1}^{m} \left| \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h} \boldsymbol{x}^{\mathsf{H}} \boldsymbol{a}_j - \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}} \boldsymbol{a}_j \right|^2,$$

which excludes both the noise $\boldsymbol{\xi}$ and the regularization term from consideration when compared with the original loss $f(\cdot)$. By virtue of (A.7), it is easily seen that

$$\nabla^2 f_{\text{reg-free}}(\boldsymbol{z}) = \nabla^2 f_{\mathsf{clean}}(\boldsymbol{z}) + \begin{bmatrix} \boldsymbol{M} & \boldsymbol{0} \\ 0 & \overline{\boldsymbol{M}} \end{bmatrix}, \tag{A.25}$$

where

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

By setting

$$\boldsymbol{u} = \begin{bmatrix} \boldsymbol{h}_1 - \boldsymbol{h}_2 \\ \boldsymbol{x}_1 - \boldsymbol{x}_2 \\ \overline{\boldsymbol{h}_1 - \boldsymbol{h}_2} \\ \overline{\boldsymbol{x}_1 - \boldsymbol{x}_2} \end{bmatrix} =: \begin{bmatrix} \boldsymbol{u}_{\boldsymbol{h}} \\ \boldsymbol{u}_{\boldsymbol{x}} \\ \overline{\boldsymbol{u}_{\boldsymbol{h}}} \\ \overline{\boldsymbol{u}_{\boldsymbol{x}}} \end{bmatrix}$$

and recalling the definitions of $\boldsymbol{D}$, γ_1, γ_2 in the statement of Lemma 4, we arrive at

$$
\begin{aligned}
&\boldsymbol{u}^{\mathsf{H}}\left[\boldsymbol{D}\nabla^2 f_{\text{reg-free}}\left(\boldsymbol{z}\right)+\nabla^2 f_{\text{reg-free}}\left(\boldsymbol{z}\right)\boldsymbol{D}\right]\boldsymbol{u}\\
&=\boldsymbol{u}^{\mathsf{H}}\left[\boldsymbol{D}\nabla^2 f_{\text{clean}}\left(\boldsymbol{z}\right)+\nabla^2 f_{\text{clean}}\left(\boldsymbol{z}\right)\boldsymbol{D}\right]\boldsymbol{u}-2\left(\gamma_1+\gamma_2\right)\mathrm{Re}\left(\boldsymbol{u}_{\boldsymbol{h}}^{\mathsf{H}}\sum_{j=1}^{m}\xi_j\boldsymbol{b}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{u}_{\boldsymbol{x}}\right)\\
&\quad-2\left(\gamma_1+\gamma_2\right)\mathrm{Re}\left(\overline{\boldsymbol{u}_{\boldsymbol{h}}}^{\mathsf{H}}\overline{\sum_{j=1}^{m}\xi_j\boldsymbol{b}_j\boldsymbol{a}_j^{\mathsf{H}}}\overline{\boldsymbol{u}_{\boldsymbol{x}}}\right)\\
&=\boldsymbol{u}^{\mathsf{H}}\left[\boldsymbol{D}\nabla^2 f_{\text{clean}}\left(\boldsymbol{z}\right)+\nabla^2 f_{\text{clean}}\left(\boldsymbol{z}\right)\boldsymbol{D}\right]\boldsymbol{u}-4\left(\gamma_1+\gamma_2\right)\mathrm{Re}\left(\boldsymbol{u}_{\boldsymbol{h}}^{\mathsf{H}}\sum_{j=1}^{m}\xi_j\boldsymbol{b}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{u}_{\boldsymbol{x}}\right).
\end{aligned}
$$

Consequently, with high probability one has

$$
\begin{aligned}
&\left|\boldsymbol{u}^{\mathsf{H}}\left[\boldsymbol{D}\nabla^2 f_{\text{reg-free}}\left(\boldsymbol{z}\right)+\nabla^2 f_{\text{reg-free}}\left(\boldsymbol{z}\right)\boldsymbol{D}\right]\boldsymbol{u}-\boldsymbol{u}^{\mathsf{H}}\left[\boldsymbol{D}\nabla^2 f_{\text{clean}}\left(\boldsymbol{z}\right)+\nabla^2 f_{\text{clean}}\left(\boldsymbol{z}\right)\boldsymbol{D}\right]\boldsymbol{u}\right|\\
&\leq 4\left(\gamma_1+\gamma_2\right)\left|\mathrm{Re}\left(\boldsymbol{u}_{\boldsymbol{h}}^{\mathsf{H}}\sum_{j=1}^{m}\xi_j\boldsymbol{b}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{u}_{\boldsymbol{x}}\right)\right|\leq 4\left(\gamma_1+\gamma_2\right)\left\|\sum_{j=1}^{m}\xi_j\boldsymbol{b}_j\boldsymbol{a}_j^{\mathsf{H}}\right\|\left\|\boldsymbol{u}\right\|_2^2\\
&\lesssim\sigma\sqrt{K\log m}\left\|\boldsymbol{u}\right\|_2^2=:\mathcal{E}_{\text{res}}
\end{aligned}
\tag{A.26}
$$

for any vector $\boldsymbol{u}$, where the last inequality follows from Lemma 38 as well as the assumptions $\gamma_1,\gamma_2\asymp 1$.

The above bound allows us to turn attention to $\nabla^2 f_{\text{clean}}$, which has been studied in Ma et al. [2018]. In particular, it has been shown in Ma et al. [2018] that

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

under the assumptions stated in the lemma. These bounds together with (A.26) yield

$$
\boldsymbol{u}^{\mathsf{H}}\left[\boldsymbol{D}\nabla^2 f_{\text{reg-free}}\left(\boldsymbol{z}\right)+\nabla^2 f_{\text{reg-free}}\left(\boldsymbol{z}\right)\boldsymbol{D}\right]\boldsymbol{u}\geq(1/4)\cdot\left\|\boldsymbol{u}\right\|_2^2-\mathcal{E}_{\text{res}}\geq(1/8)\cdot\left\|\boldsymbol{u}\right\|_2^2,\tag{A.27a}
$$

$$
\text{and}\quad\left\|\nabla^2 f_{\text{reg-free}}\left(\boldsymbol{z}\right)\right\|\leq\left\|\nabla^2 f_{\text{clean}}\left(\boldsymbol{z}\right)\right\|+\sup_{\boldsymbol{u}\neq\boldsymbol{0}}\frac{\mathcal{E}_{\text{res}}}{\left\|\boldsymbol{u}\right\|_2^2}\leq 7/2,\tag{A.27b}
$$

provided that $\sigma\sqrt{K\log m}\leq 0.5$. To finish up, we recall that

$$
\nabla^2 f\left(\boldsymbol{z}\right)=\nabla^2 f_{\text{reg-free}}\left(\boldsymbol{z}\right)+\lambda\boldsymbol{I},
$$

which combined with (A.27) and the assumption $\lambda\leq C_\lambda\sigma\sqrt{K\log m}\leq C_\lambda c_1/\log^2 m\ll 1$ yields

$$
\begin{aligned}
\boldsymbol{u}^{\mathsf{H}}\left[\boldsymbol{D}\nabla^2 f\left(\boldsymbol{z}\right)+\nabla^2 f\left(\boldsymbol{z}\right)\boldsymbol{D}\right]\boldsymbol{u}&=\boldsymbol{u}^{\mathsf{H}}\left[\boldsymbol{D}\nabla^2 f_{\text{reg-free}}\left(\boldsymbol{z}\right)+\nabla^2 f_{\text{reg-free}}\left(\boldsymbol{z}\right)\boldsymbol{D}\right]\boldsymbol{u}+2\lambda\boldsymbol{u}^{\mathsf{H}}\boldsymbol{D}\boldsymbol{u}\\
&\geq\boldsymbol{u}^{\mathsf{H}}\left[\boldsymbol{D}\nabla^2 f_{\text{reg-free}}\left(\boldsymbol{z}\right)+\nabla^2 f_{\text{reg-free}}\left(\boldsymbol{z}\right)\boldsymbol{D}\right]\boldsymbol{u}\\
&\geq\left\|\boldsymbol{u}\right\|_2^2/8
\end{aligned}
$$

and

$$
\left\|\nabla^2 f\left(\boldsymbol{z}\right)\right\|\leq\left\|\nabla^2 f_{\text{reg-free}}\left(\boldsymbol{z}\right)\right\|+\lambda\leq 4.
$$

A.7 Proof of Lemma 5

Recognizing that

$$
f_{\text{reg-free}}\left(\boldsymbol{h},\boldsymbol{x}\right)=f_{\text{reg-free}}\left(\frac{1}{\alpha}\boldsymbol{h},\alpha\boldsymbol{x}\right)\quad\text{and}\quad\nabla f_{\text{reg-free}}\left(\boldsymbol{h},\boldsymbol{x}\right)=\begin{bmatrix}\frac{1}{\alpha}\nabla_{\boldsymbol{h}}f_{\text{reg-free}}\left(\frac{1}{\alpha}\boldsymbol{h},\alpha\boldsymbol{x}\right)\\ \overline{\alpha}\nabla_{\boldsymbol{x}}f_{\text{reg-free}}\left(\frac{1}{\alpha}\boldsymbol{h},\alpha\boldsymbol{x}\right)\end{bmatrix}
$$

and recalling the definitions of $\left(\widetilde{\boldsymbol{h}}^{t}, \widetilde{\boldsymbol{x}}^{t}\right) := \left(\frac{1}{\overline{\alpha^{t}}}\boldsymbol{h}^{t}, \alpha^{t}\boldsymbol{x}^{t}\right)$, we can deduce that

$$\mathsf{dist}\left(\boldsymbol{z}^{t+1}, \boldsymbol{z}^{\star}\right) = \mathsf{dist}\left(\boldsymbol{z}^{t+1/2}, \boldsymbol{z}^{\star}\right) \leq \left\| \begin{bmatrix} \frac{1}{\overline{\alpha^{t}}}\boldsymbol{h}^{t+1/2} - \boldsymbol{h}^{\star} \\ \alpha^{t}\boldsymbol{x}^{t+1/2} - \boldsymbol{x}^{\star} \end{bmatrix} \right\|_{2} \tag{A.28}$$

$$= \left\| \begin{bmatrix} \widetilde{\boldsymbol{h}}^{t} - \frac{\eta}{|\alpha^{t}|^{2}}\nabla_{\boldsymbol{h}} f_{\textsf{reg-free}}\left(\widetilde{\boldsymbol{z}}^{t}\right) - \eta\lambda\widetilde{\boldsymbol{h}}^{t} - \left(\boldsymbol{h}^{\star} - \frac{\eta}{|\alpha^{t}|^{2}}\nabla_{\boldsymbol{h}} f_{\textsf{reg-free}}\left(\boldsymbol{z}^{\star}\right)\right) - \frac{\eta}{|\alpha^{t}|^{2}}\nabla_{\boldsymbol{h}} f_{\textsf{reg-free}}\left(\boldsymbol{z}^{\star}\right) \\ \widetilde{\boldsymbol{x}}^{t} - \eta\left|\alpha^{t}\right|^{2}\nabla_{\boldsymbol{x}} f_{\textsf{reg-free}}\left(\widetilde{\boldsymbol{z}}^{t}\right) - \eta\lambda\widetilde{\boldsymbol{x}}^{t} - \left(\boldsymbol{x}^{\star} - \eta\left|\alpha^{t}\right|^{2}\nabla_{\boldsymbol{x}} f_{\textsf{reg-free}}\left(\boldsymbol{z}^{\star}\right)\right) - \eta\left|\alpha^{t}\right|^{2}\nabla_{\boldsymbol{x}} f_{\textsf{reg-free}}\left(\boldsymbol{z}^{\star}\right) \end{bmatrix} \right\|_{2}$$

$$\leq \underbrace{\left\| \begin{bmatrix} \widetilde{\boldsymbol{h}}^{t} - \frac{\eta}{|\alpha^{t}|^{2}}\nabla_{\boldsymbol{h}} f_{\textsf{reg-free}}\left(\widetilde{\boldsymbol{z}}^{t}\right) - \left(\boldsymbol{h}^{\star} - \frac{\eta}{|\alpha^{t}|^{2}}\nabla_{\boldsymbol{h}} f_{\textsf{reg-free}}\left(\boldsymbol{z}^{\star}\right)\right) \\ \widetilde{\boldsymbol{x}}^{t} - \eta\left|\alpha^{t}\right|^{2}\nabla_{\boldsymbol{x}} f_{\textsf{reg-free}}\left(\widetilde{\boldsymbol{z}}^{t}\right) - \left(\boldsymbol{x}^{\star} - \eta\left|\alpha^{t}\right|^{2}\nabla_{\boldsymbol{x}} f_{\textsf{reg-free}}\left(\boldsymbol{z}^{\star}\right)\right) \end{bmatrix} \right\|_{2}}_{=:\beta_{1}}$$

$$+ \underbrace{\left\| \begin{bmatrix} \frac{\eta}{|\alpha^{t}|^{2}}\nabla_{\boldsymbol{h}} f_{\textsf{reg-free}}\left(\boldsymbol{z}^{\star}\right) \\ \eta\left|\alpha^{t}\right|^{2}\nabla_{\boldsymbol{x}} f_{\textsf{reg-free}}\left(\boldsymbol{z}^{\star}\right) \end{bmatrix} \right\|_{2}}_{=:\beta_{2}} + \underbrace{\eta\lambda\left\| \begin{bmatrix} \widetilde{\boldsymbol{h}}^{t} \\ \widetilde{\boldsymbol{x}}^{t} \end{bmatrix} \right\|_{2}}_{=:\beta_{3}}. \tag{A.29}$$

Using an argument similar to the proof idea of Ma et al. [2018, Equation (210)], we can obtain

$$\begin{aligned} \beta_{1}^{2} &= \left\| \widetilde{\boldsymbol{h}}^{t} - \frac{\eta}{|\alpha^{t}|^{2}}\nabla_{\boldsymbol{h}} f_{\textsf{reg-free}}\left(\widetilde{\boldsymbol{z}}^{t}\right) - \left(\boldsymbol{h}^{\star} - \frac{\eta}{|\alpha^{t}|^{2}}\nabla_{\boldsymbol{h}} f_{\textsf{reg-free}}\left(\boldsymbol{z}^{\star}\right)\right) \right\|_{2}^{2} \\ &\quad + \left\| \widetilde{\boldsymbol{x}}^{t} - \eta\left|\alpha^{t}\right|^{2}\nabla_{\boldsymbol{x}} f_{\textsf{reg-free}}\left(\widetilde{\boldsymbol{z}}^{t}\right) - \left(\boldsymbol{x}^{\star} - \eta\left|\alpha^{t}\right|^{2}\nabla_{\boldsymbol{x}} f_{\textsf{reg-free}}\left(\boldsymbol{z}^{\star}\right)\right) \right\|_{2}^{2} \\ &\leq \left(1 - \frac{\eta}{8}\right)\left\|\widetilde{\boldsymbol{z}}^{t} - \boldsymbol{z}^{\star}\right\|_{2}^{2}. \end{aligned} \tag{A.30}$$

Regarding β_{2}, we first invoke Lemma 14 and the fact $\nabla f_{\textsf{clean}}\left(\boldsymbol{z}^{\star}\right) = \boldsymbol{0}$ to derive

$$\begin{aligned} \left\|\nabla f_{\textsf{reg-free}}\left(\boldsymbol{z}^{\star}\right)\right\|_{2} &\leq \left\|\nabla f_{\textsf{clean}}\left(\boldsymbol{z}^{\star}\right)\right\|_{2} + \left\|\mathcal{A}^{*}\left(\boldsymbol{\xi}\right)\right\|\left\|\boldsymbol{h}^{\star}\right\|_{2} + \left\|\mathcal{A}^{*}\left(\boldsymbol{\xi}\right)\right\|\left\|\boldsymbol{x}^{\star}\right\|_{2} \\ &\lesssim \sigma\sqrt{K\log m}. \end{aligned} \tag{A.31}$$

A little algebra then yields

$$\begin{aligned} \beta_{2}^{2} &= \left\| \frac{\eta}{|\alpha^{t}|^{2}}\nabla_{\boldsymbol{h}} f_{\textsf{reg-free}}\left(\boldsymbol{z}^{\star}\right) \right\|_{2}^{2} + \left\| \eta\left|\alpha^{t}\right|^{2}\nabla_{\boldsymbol{x}} f_{\textsf{reg-free}}\left(\boldsymbol{z}^{\star}\right) \right\|_{2}^{2} \\ &\leq \left(\frac{\eta^{2}}{|\alpha^{t}|^{4}} + \eta^{2}\left|\alpha^{t}\right|^{4}\right)\left\|\nabla f_{\textsf{reg-free}}\left(\boldsymbol{z}^{\star}\right)\right\|_{2}^{2} \\ &\lesssim \eta^{2}\left(\sigma\sqrt{K\log m}\right)^{2}, \end{aligned}$$

which relies on the observation that $|\alpha^{t}| \asymp 1$ (see Corollary 1). Finally, when it comes to β_{3}, we have

$$\beta_{3}^{2} = \eta^{2}\lambda^{2}\left\|\widetilde{\boldsymbol{h}}^{t}\right\|_{2}^{2} + \eta^{2}\lambda^{2}\left\|\widetilde{\boldsymbol{x}}^{t}\right\|_{2}^{2} \leq 8\eta^{2}\lambda^{2},$$

using the fact that $\left\|\widetilde{\boldsymbol{x}}^{t}\right\|_{2} \asymp \left\|\widetilde{\boldsymbol{h}}^{t}\right\|_{2} \asymp 1$ (see Lemma 3).

As a result, as long as $\eta > 0$ is taken to be some constant small enough, combining (A.29) and the above bounds on β_{1}, β_{2} gives

$$\mathsf{dist}\left(\boldsymbol{z}^{t+1}, \boldsymbol{z}^{\star}\right) \leq \left\|\widehat{\boldsymbol{z}}^{t+1/2} - \boldsymbol{z}^{\star}\right\|_{2}^{2} \leq \sqrt{(1-\eta/8)}\left\|\widetilde{\boldsymbol{z}}^{t} - \boldsymbol{z}^{\star}\right\|_{2} + C_{1}\eta\left(\lambda + \sigma\sqrt{K\log m}\right),$$

which together with the elementary fact $\sqrt{1-x} \leq 1 - x/2$ leads to

$$\begin{aligned} \mathsf{dist}\left(\boldsymbol{z}^{t+1}, \boldsymbol{z}^{\star}\right) &\leq \left\|\widehat{\boldsymbol{z}}^{t+1/2} - \boldsymbol{z}^{\star}\right\|_{2} \leq (1-\eta/16)\left\|\widetilde{\boldsymbol{z}}^{t} - \boldsymbol{z}^{\star}\right\|_{2} + C_{1}\eta\left(\lambda + \sigma\sqrt{K\log m}\right) \\ &= (1-\eta/16)\,\mathsf{dist}\left(\boldsymbol{z}^{t}, \boldsymbol{z}^{\star}\right) + C_{1}\eta\left(\lambda + \sigma\sqrt{K\log m}\right). \end{aligned}$$

The advertised claim then follows, provided that C_{1} is large enough.

A.8 Proof of Lemma 6

The lemma can be established in a similar manner as Ma et al. [2018, Lemma 17]. We have

$$\begin{aligned}\mathsf{dist}\big(\boldsymbol{z}^{t+1,(l)}, \widetilde{\boldsymbol{z}}^{t+1}\big) &= \mathsf{dist}\big(\boldsymbol{z}^{t+1/2,(l)}, \widetilde{\boldsymbol{z}}^{t+1/2}\big) \\ &\leq \max\left\{\left|\frac{\alpha^{t+1/2}}{\alpha^t}\right|, \left|\frac{\alpha^t}{\alpha^{t+1/2}}\right|\right\}\left\|\begin{bmatrix}\frac{1}{\alpha_{\text{mutual}}^{t,(l)}}\boldsymbol{h}^{t+1/2,(l)} - \frac{1}{\alpha^t}\boldsymbol{h}^{t+1/2} \\ \alpha_{\text{mutual}}^{t,(l)}\boldsymbol{x}^{t+1,(l)} - \alpha^t \boldsymbol{x}^{t+1/2}\end{bmatrix}\right\|_2,\end{aligned} \tag{A.32}$$

where the second line comes from the same calculation as Ma et al. [2018, Eqn. (212)]. Repeating the analysis in Ma et al. [2018, Appendix C.3] and using the gradient update rule, we obtain

$$\begin{aligned}&\begin{bmatrix}\frac{1}{\alpha_{\text{mutual}}^{t,(l)}}\boldsymbol{h}^{t+1/2,(l)} - \frac{1}{\alpha^t}\boldsymbol{h}^{t+1/2} \\ \alpha_{\text{mutual}}^{t,(l)}\boldsymbol{x}^{t+1,(l)} - \alpha^t \boldsymbol{x}^{t+1/2}\end{bmatrix} \\ &= \underbrace{\begin{bmatrix}\widehat{\boldsymbol{h}}^{t,(l)} - \frac{\eta}{|\alpha_{\text{mutual}}^{t,(l)}|^2}\nabla_{\boldsymbol{h}} f_{\text{reg-free}}\big(\widehat{\boldsymbol{z}}^{t,(l)}\big) - \left(\widetilde{\boldsymbol{h}}^t - \frac{\eta}{|\alpha_{\text{mutual}}^{t,(l)}|^2}\nabla_{\boldsymbol{h}} f_{\text{reg-free}}\big(\widetilde{\boldsymbol{z}}^t\big)\right) \\ \widehat{\boldsymbol{x}}^{t,(l)} - \eta|\alpha_{\text{mutual}}^{t,(l)}|^2\nabla_{\boldsymbol{x}} f_{\text{reg-free}}\big(\widehat{\boldsymbol{z}}^{t,(l)}\big) - \left(\widetilde{\boldsymbol{x}}^t - \eta|\alpha_{\text{mutual}}^{t,(l)}|^2\nabla_{\boldsymbol{x}} f_{\text{reg-free}}\big(\widetilde{\boldsymbol{z}}^t\big)\right)\end{bmatrix}}_{=:\boldsymbol{\nu}_1} \\ &\quad + \eta\underbrace{\begin{bmatrix}\left(\frac{1}{|\alpha^t|^2} - \frac{1}{|\alpha_{\text{mutual}}^{t,(l)}|^2}\right)\nabla_{\boldsymbol{h}} f_{\text{reg-free}}\big(\widetilde{\boldsymbol{z}}^t\big) \\ \left(|\alpha^t|^2 - |\alpha_{\text{mutual}}^{t,(l)}|^2\right)\nabla_{\boldsymbol{x}} f_{\text{reg-free}}\big(\widetilde{\boldsymbol{z}}^t\big)\end{bmatrix}}_{=:\boldsymbol{\nu}_2} - \eta\underbrace{\begin{bmatrix}\frac{1}{|\alpha_{\text{mutual}}^{t,(l)}|^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)} \\ |\alpha_{\text{mutual}}^{t,(l)}|^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{\nu}_3} \\ &\quad + \eta\lambda\underbrace{\begin{bmatrix}\widehat{\boldsymbol{h}}^{t,(l)} - \widetilde{\boldsymbol{h}}^t \\ \widehat{\boldsymbol{x}}^{t,(l)} - \widetilde{\boldsymbol{x}}^t\end{bmatrix}}_{=:\boldsymbol{\nu}_4}.\end{aligned} \tag{A.33}$$

In what follows, we shall look at $\boldsymbol{\nu}_1$, $\boldsymbol{\nu}_2$, $\boldsymbol{\nu}_3$ and $\boldsymbol{\nu}_4$ separately.

- It has been shown in Ma et al. [2018, Lemma 17] that

$$\|\boldsymbol{\nu}_1\|_2 \leq (1-\eta/16)\left\|\widehat{\boldsymbol{z}}^{t,(l)} - \widetilde{\boldsymbol{z}}^t\right\|_2; \qquad \|\boldsymbol{\nu}_2\|_2 \lesssim C_1\frac{1}{\log^2 m}\left\|\widehat{\boldsymbol{z}}^{t,(l)} - \widetilde{\boldsymbol{z}}^t\right\|_2. \tag{A.34}$$

- Regarding $\boldsymbol{\nu}_3$, we have

$$\begin{aligned}\|\boldsymbol{\nu}_3\|_2 &= \sqrt{\frac{1}{|\alpha_{\text{mutual}}^{t,(l)}|^4}\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\|_2^2 + |\alpha_{\text{mutual}}^{t,(l)}|^4\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\|_2^2} \\ &\leq \frac{1}{|\alpha_{\text{mutual}}^{t,(l)}|^2}\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 + |\alpha_{\text{mutual}}^{t,(l)}|^2\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\|_2 \\ &\leq \frac{1}{|\alpha_{\text{mutual}}^{t,(l)}|^2}\underbrace{\left\|\boldsymbol{b}_l^{\mathsf{H}}\left(\widehat{\boldsymbol{h}}^{t,(l)}\widehat{\boldsymbol{x}}^{t,(l)\mathsf{H}} - \boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\right)\boldsymbol{a}_l\boldsymbol{b}_l\boldsymbol{a}_l^{\mathsf{H}}\widehat{\boldsymbol{x}}^{t,(l)}\right\|_2}_{=:\nu_{31}} \\ &\quad + |\alpha_{\text{mutual}}^{t,(l)}|^2\underbrace{\left\|\overline{\boldsymbol{b}_l^{\mathsf{H}}\left(\widehat{\boldsymbol{h}}^{t,(l)}\widehat{\boldsymbol{x}}^{t,(l)\mathsf{H}} - \boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\right)\boldsymbol{a}_l}\boldsymbol{a}_l\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\right\|_2}_{=:\nu_{32}} \\ &\quad + \frac{1}{|\alpha_{\text{mutual}}^{t,(l)}|^2}\underbrace{\left\|\xi_l\boldsymbol{b}_l\boldsymbol{a}_l^{\mathsf{H}}\widehat{\boldsymbol{x}}^{t,(l)}\right\|_2}_{=:\nu_{33}} + |\alpha_{\text{mutual}}^{t,(l)}|^2\underbrace{\left\|\overline{\xi_l}\boldsymbol{a}_l\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\right\|_2}_{=:\nu_{34}},\end{aligned} \tag{A.35a}$$

where the first inequality comes from the elementary inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$, and the second inequality follows from the triangle inequality. The bounds of ν_{31} and ν_{32} follow from the same derivation as Ma et al. [2018, Equation (217)] and are thus omitted here for simplicity. The quantity ν_{31} can be upper bounded by

$$\begin{aligned}\nu_{31} &\le \left|\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\widehat{\boldsymbol{x}}^{t,(l)\mathsf{H}}\boldsymbol{a}_l - \boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_l\right| \left\|\boldsymbol{b}_l\right\|_2 \left|\boldsymbol{a}_l^{\mathsf{H}}\widehat{\boldsymbol{x}}^{t,(l)}\right| \\ &\le \left|\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\widehat{\boldsymbol{x}}^{t,(l)\mathsf{H}}\boldsymbol{a}_l - \boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_l\right| \cdot \sqrt{\frac{K}{m}} \cdot 20\sqrt{\log m} \cdot \left\|\widehat{\boldsymbol{x}}^{t,(l)}\right\|_2 \\ &\le 40\sqrt{\frac{K\log m}{m}}\left|\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\widehat{\boldsymbol{x}}^{t,(l)\mathsf{H}}\boldsymbol{a}_l - \boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_l\right|, \end{aligned} \tag{A.35b}$$

where the penultimate inequality follows from the fact that $\left\|\boldsymbol{b}_l\right\|_2 = \sqrt{K/m}$ and (F.1), and the last line makes use of (A.15f). Regarding ν_{32}, one has

$$\begin{aligned}\nu_{32} &\le \left|\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\widehat{\boldsymbol{x}}^{t,(l)\mathsf{H}}\boldsymbol{a}_l - \boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_l\right| \left\|\boldsymbol{a}_l\right\|_2 \left|\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\right| \\ &\le \left|\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\widehat{\boldsymbol{x}}^{t,(l)\mathsf{H}}\boldsymbol{a}_l - \boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_l\right| \cdot 10\sqrt{K} \cdot \left(\sqrt{\frac{K}{m}}\left\|\widehat{\boldsymbol{h}}^{t,(l)} - \widetilde{\boldsymbol{h}}^{t}\right\|_2 + \left|\boldsymbol{b}_l^{\mathsf{H}}\widetilde{\boldsymbol{h}}^{t}\right|\right) \\ &\le \left|\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\widehat{\boldsymbol{x}}^{t,(l)\mathsf{H}}\boldsymbol{a}_l - \boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_l\right| \cdot 10\sqrt{K} \cdot \sqrt{\frac{K}{m}}C_2\left(\frac{\mu}{\sqrt{m}}\sqrt{\frac{\mu^2 K\log^9 m}{m}} + \frac{\sigma}{\log^2 m}\right) \\ &\quad + \left|\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\widehat{\boldsymbol{x}}^{t,(l)\mathsf{H}}\boldsymbol{a}_l - \boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_l\right| \cdot 10\sqrt{K} \cdot C_4\left(\frac{\mu}{\sqrt{m}}\log^2 m + \sigma\right) \\ &\le 20C_4\left(\frac{\mu\sqrt{K}}{\sqrt{m}}\log^2 m + \sigma\sqrt{K}\right)\left|\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\widehat{\boldsymbol{x}}^{t,(l)\mathsf{H}}\boldsymbol{a}_l - \boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_l\right|, \end{aligned} \tag{A.35c}$$

where the second line follows from (F.2), triangle inequality and the fact that $\left\|\boldsymbol{b}_l\right\|_2 = \sqrt{K/m}$; the penultimate inequality follows from (A.14a) and (A.6c); the last line holds as long as $m \gg \mu^2 K\log^3 m$. Further we have

$$\begin{aligned}\left|\boldsymbol{b}_l^{\mathsf{H}}\left(\widehat{\boldsymbol{h}}^{t,(l)} - \boldsymbol{h}^{\star}\right)\right| &\le \left|\boldsymbol{b}_l^{\mathsf{H}}\left(\widehat{\boldsymbol{h}}^{t,(l)} - \widetilde{\boldsymbol{h}}^{t}\right)\right| + \left|\boldsymbol{b}_l^{\mathsf{H}}\widetilde{\boldsymbol{h}}^{t}\right| + \left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^{\star}\right| \\ &\le \sqrt{\frac{K}{m}}\left\|\widehat{\boldsymbol{h}}^{t,(l)} - \widetilde{\boldsymbol{h}}^{t}\right\|_2 + \left|\boldsymbol{b}_l^{\mathsf{H}}\widetilde{\boldsymbol{h}}^{t}\right| + \left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^{\star}\right| \\ &\le \sqrt{\frac{K}{m}}C_2\left(\frac{\mu}{\sqrt{m}}\sqrt{\frac{\mu^2 K\log^9 m}{m}} + \frac{\sigma}{\log^2 m}\right) + C_4\left(\frac{\mu}{\sqrt{m}}\log^2 m + \sigma\right) + \frac{\mu}{\sqrt{m}} \\ &\le 2C_4\left(\frac{\mu}{\sqrt{m}}\log^2 m + \sigma\right), \end{aligned} \tag{A.35d}$$

where the second line follows from the fact that $\left\|\boldsymbol{b}_l\right\|_2 = \sqrt{K/m}$; the penultimate inequality follows from (A.14a), (A.6c) and (7); the last line holds as long as $m \gg \mu^2 K\log^3 m$. Therefore,

$$\begin{aligned}&\left|\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\widehat{\boldsymbol{x}}^{t,(l)\mathsf{H}}\boldsymbol{a}_l - \boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_l\right| \\ &\le \left|\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\left(\widehat{\boldsymbol{x}}^{t,(l)} - \boldsymbol{x}^{\star}\right)^{\mathsf{H}}\boldsymbol{a}_l\right| + \left|\boldsymbol{b}_l^{\mathsf{H}}\left(\widehat{\boldsymbol{h}}^{t,(l)} - \boldsymbol{h}^{\star}\right)\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_l\right| \\ &\le \left(\left|\boldsymbol{b}_l^{\mathsf{H}}\left(\widehat{\boldsymbol{h}}^{t,(l)} - \boldsymbol{h}^{\star}\right)\right| + \left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^{\star}\right|\right) \cdot 20\sqrt{\log m}\left(\left\|\widehat{\boldsymbol{x}}^{t,(l)} - \widetilde{\boldsymbol{x}}^{t}\right\|_2 + \left\|\widetilde{\boldsymbol{x}}^{t} - \boldsymbol{x}^{\star}\right\|_2\right) + \left|\boldsymbol{b}_l^{\mathsf{H}}\left(\widehat{\boldsymbol{h}}^{t,(l)} - \boldsymbol{h}^{\star}\right)\right| \cdot \left|\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_l\right| \\ &\le 2C_4\left(\frac{\mu}{\sqrt{m}}\log^2 m + \sigma\right) \cdot 20\sqrt{\log m} \cdot C_2\left(\frac{\mu}{\sqrt{m}}\sqrt{\frac{\mu^2 K\log^9 m}{m}} + \frac{\sigma}{\log^2 m}\right) \\ &\quad + 2C_4\left(\frac{\mu}{\sqrt{m}}\log^2 m + \sigma\right) \cdot C_1\left(\sqrt{\frac{\mu^2 K\log m}{m}} + \lambda + \sigma\sqrt{K\log m}\right) \cdot 20\sqrt{\log m}\end{aligned}$$

$$
\begin{aligned}
&+ 2C_4\left(\frac{\mu}{\sqrt{m}}\log^2 m+\sigma\right)\cdot 20\sqrt{\log m}\\
&\lesssim C_4\left(\frac{\mu}{\sqrt{m}}\log^{2.5} m+\sigma\sqrt{\log m}\right), \qquad \text{(A.35e)}
\end{aligned}
$$

where the second inequality follows from triangle inequality and (F.1); the penultimate inequality follows from (A.35d), (A.14a), (A.15a) and (F.1); the last line holds as long as $m \gg \mu^2 K \log m$. Substituting (A.35e) into (A.35b) and (A.35c), we reach

$$
\begin{aligned}
\nu_{31}+\nu_{32} &\lesssim \left(40\sqrt{\frac{K\log m}{m}}+20C_4\left(\frac{\mu\sqrt{K}}{\sqrt{m}}\log^2 m+\sigma\sqrt{K}\right)\right)C_4\left(\frac{\mu}{\sqrt{m}}\log^{2.5} m+\sigma\sqrt{\log m}\right)\\
&\le (C_4)^2\,\frac{\mu}{\sqrt{m}}\sqrt{\frac{\mu^2 K\log^9 m}{m}}+C_4\frac{\sigma}{\log^2 m}, \qquad \text{(A.35f)}
\end{aligned}
$$

as long as $m \gg \mu^2 K \log^9 m$. Regarding ν_{33} and ν_{34}, it is seen that

$$
\left\|\xi_l \boldsymbol{b}_l \boldsymbol{a}_l^{\mathsf{H}} \widehat{\boldsymbol{x}}^{t,(l)}\right\|_2 \le |\xi_l|\,\|\boldsymbol{b}_l\|_2\left|\boldsymbol{a}_l^{\mathsf{H}}\widehat{\boldsymbol{x}}^{t,(l)}\right| \overset{\text{(i)}}{\lesssim} \sigma\sqrt{\frac{K}{m}}\left\|\widehat{\boldsymbol{x}}^{t,(l)}\right\|_2\log m \overset{\text{(ii)}}{\le} 2\sigma\sqrt{\frac{K}{m}}\log m, \qquad \text{(A.35g)}
$$

$$
\begin{aligned}
\left\|\overline{\xi_l}\boldsymbol{a}_l \boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\right\|_2 &\le |\xi_l|\,\|\boldsymbol{a}_l\|_2\left|\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\right| \overset{\text{(iii)}}{\lesssim} \sigma\sqrt{K}\left(\left|\boldsymbol{b}_l^{\mathsf{H}}\left(\widehat{\boldsymbol{h}}^{t,(l)}-\boldsymbol{h}^{\star}\right)\right|+\left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^{\star}\right|\right)\\
&\overset{\text{(iv)}}{\lesssim} \sigma\sqrt{K}\left(2C_4\left(\frac{\mu}{\sqrt{m}}\log^2 m+\sigma\right)+\frac{\mu}{\sqrt{m}}\right)\\
&\lesssim C_4\frac{\sigma}{\log^{2.5} m}+C_4\sigma\sqrt{\frac{\mu^2 K\log^4 m}{m}}, \qquad \text{(A.35h)}
\end{aligned}
$$

where (i) holds by the property of sub-Gaussian variables (cf. Vershynin [2018, Proposition 2.5.2]) and the independence between $\xi_l, \boldsymbol{a}_l$ and $\widehat{\boldsymbol{x}}^{t,(l)}$, (ii) holds by (A.15f), (iii) is due to Lemma (38), the triangle inequality and (7), and (iv) follows from (A.35d) and (7). Consequently, by (A.35f)-(A.35h) we have

$$
\|\boldsymbol{\nu}_3\|_2 \lesssim (C_4)^2\,\frac{\mu}{\sqrt{m}}\sqrt{\frac{\mu^2 K\log^9 m}{m}}+C_4\frac{\sigma}{\log^2 m}. \qquad \text{(A.36)}
$$

- Finally, in terms of $\boldsymbol{\nu}_4$ one has

$$
\|\boldsymbol{\nu}_4\|_2 = \left\|\begin{bmatrix}\widehat{\boldsymbol{h}}^{t,(l)}-\widetilde{\boldsymbol{h}}^t\\ \widehat{\boldsymbol{x}}^{t,(l)}-\widetilde{\boldsymbol{x}}^t\end{bmatrix}\right\|_2 = \left\|\widehat{\boldsymbol{z}}^{t,(l)}-\widetilde{\boldsymbol{z}}^t\right\|_2. \qquad \text{(A.37)}
$$

With the above bounds in place, we can demonstrate that

$$
\begin{aligned}
\mathsf{dist}\left(\boldsymbol{z}^{t+1,(l)},\widetilde{\boldsymbol{z}}^{t+1}\right) &\le \max\left\{\left|\frac{\alpha^{t+1/2}}{\alpha^t}\right|,\left|\frac{\alpha^t}{\alpha^{t+1/2}}\right|\right\}\left\|\begin{bmatrix}\frac{1}{\overline{\alpha_{\text{mutual}}^{t,(l)}}}\boldsymbol{h}^{t+1/2,(l)}-\frac{1}{\overline{\alpha^t}}\boldsymbol{h}^{t+1/2}\\ \alpha_{\text{mutual}}^{t,(l)}\boldsymbol{x}^{t+1/2,(l)}-\alpha^t\boldsymbol{x}^{t+1/2}\end{bmatrix}\right\|_2\\
&\overset{\text{(i)}}{\le} \frac{1-\eta/32}{1-\eta/16}\left(\|\boldsymbol{\nu}_1\|_2+\|\boldsymbol{\nu}_2\|_2+\|\boldsymbol{\nu}_3\|_2+\|\boldsymbol{\nu}_4\|_2\right)\\
&\overset{\text{(ii)}}{\le} (1-\eta/32)\left\|\widehat{\boldsymbol{z}}^{t,(l)}-\widetilde{\boldsymbol{z}}^t\right\|_2+\frac{1-\eta/32}{1-\eta/16}C\eta\times C_1\frac{1}{\log^2 m}\left\|\widehat{\boldsymbol{z}}^{t,(l)}-\widetilde{\boldsymbol{z}}^t\right\|_2\\
&\quad+\frac{1-\eta/32}{1-\eta/16}C\eta\left((C_4)^2\,\frac{\mu}{\sqrt{m}}\sqrt{\frac{\mu^2 K\log^9 m}{m}}+C_4\frac{\sigma}{\log^2 m}\right)+\frac{1-\eta/32}{1-\eta/16}\eta\lambda\left\|\widehat{\boldsymbol{z}}^{t,(l)}-\widetilde{\boldsymbol{z}}^t\right\|_2\\
&\le \left(1-\eta/32+\frac{1-\eta/32}{1-\eta/16}\eta\lambda+\frac{1-\eta/32}{1-\eta/16}CC_1\frac{\eta}{\log^2 m}\right)\left\|\widehat{\boldsymbol{z}}^{t,(l)}-\widetilde{\boldsymbol{z}}^t\right\|_2
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{1-\eta/32}{1-\eta/16} C\eta \left((C_4)^2 \frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^9 m}{m}} + C_4 \frac{\sigma}{\log^2 m} \right) \\
&\leq \left(1 - \frac{\eta}{64}\right) \mathsf{dist}\big(\boldsymbol{z}^{t,(l)}, \widetilde{\boldsymbol{z}}^{t}\big) + \eta C (C_4)^2 \frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^9 m}{m}} + \eta C C_4 \frac{\sigma}{\log^2 m} \\
&\leq C_2 \left(\frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^9 m}{m}} + \frac{\sigma}{\log^2 m} \right), \qquad \text{(A.38)}
\end{aligned}
$$

provided that $\eta > 0$ is some sufficiently small constant and $C_2 \gg C_4^2$. To see why (i) holds, we observe that

$$
\left| \left| \frac{\alpha^{t+1/2}}{\alpha^t} \right| - 1 \right| \leq \left| \frac{\alpha^{t+1/2}}{\alpha^t} - 1 \right| \leq C \left(\sqrt{\frac{\mu^2 K \log m}{m}} + \lambda + \sigma \sqrt{K \log m} \right)
$$

as shown in Corollary 1, which implies that

$$
\left| \frac{\alpha^{t+1/2}}{\alpha^t} \right| \leq 1 + \frac{\eta/32}{1-\eta/16} = \frac{1-\eta/32}{1-\eta/16}
$$

as long as $m \gg \mu^2 K \log m$ and $\sigma\sqrt{K \log m} \ll 1$; a similar argument also reveals that

$$
\left| \frac{\alpha^{t}}{\alpha^{t+1/2}} \right| \leq \frac{1-\eta/32}{1-\eta/16}.
$$

In addition, (ii) follows from (A.34), (A.36) and (A.37), whereas the last inequality of (A.38) relies on the hypothesis (A.14a).

Next, we turn to the second inequality claimed in the lemma. In view of (A.15a) in Lemma 3, we have

$$
\left\| \widetilde{\boldsymbol{z}}^{t+1} - \boldsymbol{z}^\star \right\|_2 \leq C_1 \left(\sqrt{\frac{\mu^2 K \log m}{m}} + \lambda + \sigma \sqrt{K \log m} \right),
$$

which together with the triangle inequality and (A.38) yields

$$
\begin{aligned}
\left\| \widehat{\boldsymbol{z}}^{t+1,(l)} - \boldsymbol{z}^\star \right\|_2 &\leq \left\| \widehat{\boldsymbol{z}}^{t+1,(l)} - \widetilde{\boldsymbol{z}}^{t+1} \right\|_2 + \left\| \widetilde{\boldsymbol{z}}^{t+1} - \boldsymbol{z}^\star \right\|_2 \\
&\leq C_2 \left(\frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^9 m}{m}} + \frac{\sigma}{\log^2 m} \right) + C_1 \left(\sqrt{\frac{\mu^2 K \log m}{m}} + \lambda + \sigma \sqrt{K \log m} \right) \\
&\lesssim \sqrt{\frac{\mu^2 K \log m}{m}} + \sigma \sqrt{K \log m} + \lambda. \qquad \text{(A.39)}
\end{aligned}
$$

In other words, both $\widetilde{\boldsymbol{z}}^{t+1}$ and $\widehat{\boldsymbol{z}}^{t+1,(l)}$ are sufficiently close to the truth $\boldsymbol{z}^\star$. Consequently, we are ready to invoke Ma et al. [2018, Lemma 55]. Taking $\boldsymbol{h}_1 = \widetilde{\boldsymbol{h}}^{t+1}$, $\boldsymbol{x}_1 = \widetilde{\boldsymbol{x}}^{t+1}$, $\boldsymbol{h}_2 = \widehat{\boldsymbol{h}}^{t+1,(l)}$ and $\boldsymbol{x}_2 = \widehat{\boldsymbol{x}}^{t+1,(l)}$ in Ma et al. [2018, Lemma 55] yields

$$
\left\| \widetilde{\boldsymbol{z}}^{t+1,(l)} - \widetilde{\boldsymbol{z}}^{t+1} \right\|_2 \lesssim \left\| \widehat{\boldsymbol{z}}^{t+1,(l)} - \widetilde{\boldsymbol{z}}^{t+1} \right\|_2 \leq C_2 \left(\frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^9 m}{m}} + \frac{\sigma}{\log^2 m} \right), \qquad \text{(A.40)}
$$

where the last inequality follows from (A.39).

A.9 Proof of Lemma 7

Recall from Corollary 1 that there exist some constant $C > 0$ such that

$$
\left| \frac{\alpha^{t+1/2}}{\alpha^t} - 1 \right| \leq C\eta \left(\sqrt{\frac{\mu^2 K \log m}{m}} + \lambda + \sigma \sqrt{K \log m} \right) =: \delta, \qquad \text{(A.41)}
$$

with $\delta \ll 1$, thus indicating that

$$\max_{1\le l\le m}\left|\boldsymbol{b}_l^{\mathsf{H}}\frac{1}{\overline{\alpha^{t+1}}}\boldsymbol{h}^{t+1}\right| = \max_{1\le l\le m}\left|\boldsymbol{b}_l^{\mathsf{H}}\frac{1}{\overline{\alpha^{t+1/2}}}\boldsymbol{h}^{t+1/2}\right| \le \left|\frac{\alpha^t}{\alpha^{t+1/2}}\right|\max_{1\le l\le m}\left|\boldsymbol{b}_l^{\mathsf{H}}\frac{1}{\overline{\alpha^t}}\boldsymbol{h}^{t+1/2}\right|$$
$$\le (1+\delta)\max_{1\le l\le m}\left|\boldsymbol{b}_l^{\mathsf{H}}\frac{1}{\overline{\alpha^t}}\boldsymbol{h}^{t+1/2}\right|.$$

The gradient update rule regarding $\boldsymbol{h}^{t+1}$ then leads to

$$\frac{1}{\overline{\alpha^t}}\boldsymbol{h}^{t+1/2} = \widetilde{\boldsymbol{h}}^t - \frac{\eta}{|\alpha^t|^2}\sum_{j=1}^m\left(\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\widetilde{\boldsymbol{x}}^{t\mathsf{H}}\boldsymbol{a}_j - y_j\right)\boldsymbol{b}_j\boldsymbol{a}_j^{\mathsf{H}}\widetilde{\boldsymbol{x}}^t - \eta\lambda\widetilde{\boldsymbol{h}}^t,$$

where we recall that $\widetilde{\boldsymbol{h}}^t = \boldsymbol{h}^t/\overline{\alpha^t}$ and $\widetilde{\boldsymbol{x}}^t = \alpha^t\boldsymbol{x}^t$. Expanding terms further and using the assumption $\sum_{j=1}^m \boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}} = \boldsymbol{I}$ give

$$\begin{aligned}\frac{1}{\overline{\alpha^t}}\boldsymbol{h}^{t+1/2} &= \widetilde{\boldsymbol{h}}^t - \frac{\eta}{|\alpha^t|^2}\sum_{j=1}^m\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^t\widetilde{\boldsymbol{x}}^{t\mathsf{H}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right)\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\widetilde{\boldsymbol{x}}^t + \frac{\eta}{|\alpha^t|^2}\sum_{j=1}^m\xi_j\boldsymbol{b}_j\boldsymbol{a}_j^{\mathsf{H}}\widetilde{\boldsymbol{x}}^t - \eta\lambda\widetilde{\boldsymbol{h}}^t\\
&= \underbrace{\left(1-\eta\lambda-\frac{\eta}{|\alpha^t|^2}\|\boldsymbol{x}^\star\|_2^2\right)\widetilde{\boldsymbol{h}}^t}_{=:\boldsymbol{\nu}_0} - \frac{\eta}{|\alpha^t|^2}\underbrace{\sum_{j=1}^m\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\left(\left|\boldsymbol{a}_j^{\mathsf{H}}\widetilde{\boldsymbol{x}}^t\right|^2 - \left|\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}^\star\right|^2\right)}_{=:\boldsymbol{\nu}_1}\\
&\quad - \frac{\eta}{|\alpha^t|^2}\underbrace{\sum_{j=1}^m\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\left(\left|\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}^\star\right|^2 - \|\boldsymbol{x}^\star\|_2^2\right)}_{=:\boldsymbol{\nu}_2} + \frac{\eta}{|\alpha^t|^2}\underbrace{\sum_{j=1}^m\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\widetilde{\boldsymbol{x}}^t}_{=:\boldsymbol{\nu}_3} + \frac{\eta}{|\alpha^t|^2}\underbrace{\sum_{j=1}^m\xi_j\boldsymbol{b}_j\boldsymbol{a}_j^{\mathsf{H}}\widetilde{\boldsymbol{x}}^t}_{=:\boldsymbol{\nu}_4}.\end{aligned} \tag{A.42}$$

The first three terms can be controlled via the same arguments as Ma et al. [2018, Appendix C.4], which are built upon the induction hypotheses (A.6a)-(A.6c) at the tth iteration as well as the following claim (which is the counterpart of Ma et al. [2018, Claim 224]).

Claim 1. Suppose that $m \gg \tau K\log^4 m$. For some sufficiently small constant $c > 0$, it holds that

$$\max_{1\le j\le\tau}\left|(\boldsymbol{b}_j - \boldsymbol{b}_1)^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right| \le cC_4\left(\frac{\mu}{\sqrt{m}}\log m + \frac{\sigma}{\log m}\right).$$

The corresponding bounds obtained from Ma et al. [2018, Appendix C.4] are listed below:

$$\left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{\nu}_1\right| \le 0.1\max_{1\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right|, \tag{A.43a}$$

$$\left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{\nu}_2\right| \le 0.2\max_{1\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right| + \max_{1\le j\le\tau}\left|(\boldsymbol{b}_j - \boldsymbol{b}_1)^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right|\log m, \tag{A.43b}$$

$$\left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{\nu}_3\right| \lesssim \frac{\mu}{\sqrt{m}} + \frac{\mu}{\sqrt{m}}\log^{3/2} m\max_{1\le j\le m}\left|\boldsymbol{a}_j^{\mathsf{H}}(\widetilde{\boldsymbol{x}}^t - \boldsymbol{x}^\star)\right|. \tag{A.43c}$$

When it comes to the last term of (A.42) concerning $\boldsymbol{\nu}_4$, it is seen that

$$\left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{\nu}_4\right| \le \underbrace{\left|\sum_{j=1}^m\xi_j\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{a}_j^{\mathsf{H}}(\widetilde{\boldsymbol{x}}^t - \boldsymbol{x}^\star)\right|}_{=:\varsigma_1} + \underbrace{\left|\sum_{j=1}^m\xi_j\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}^\star\right|}_{=:\varsigma_2},$$

leaving us with two terms to control.

- With regards to ς_1, we have

$$\varsigma_1 \le \sum_{j=1}^m\left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{b}_j\right|\cdot\max_{1\le j\le m}|\xi_j|\cdot\max_{1\le j\le m}\left|\boldsymbol{a}_j^{\mathsf{H}}(\widetilde{\boldsymbol{x}}^t - \boldsymbol{x}^\star)\right|$$

$$\lesssim (4\log m)\cdot\sigma\sqrt{\log m}\cdot\max_{1\le j\le m}\left|\boldsymbol{a}_j^{\mathsf{H}}(\widetilde{\boldsymbol{x}}^t-\boldsymbol{x}^\star)\right|$$
$$\asymp \sigma\log^{1.5}m\max_{1\le j\le m}\left|\boldsymbol{a}_j^{\mathsf{H}}(\widetilde{\boldsymbol{x}}^t-\boldsymbol{x}^\star)\right|,$$

where the second inequality follows from Ma et al. [2018, Lemma 48] and standard sub-Gaussian concentration inequalities.

- Regarding ς_2, since $\left\{\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}^\star\right\}_{j=1}^m$ are i.i.d. Gaussian variables with variance $\|\boldsymbol{x}^\star\|_2=1$, we see that

$$\left\|\xi_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}^\star\right\|_{\psi_1}\le\|\xi_j\|_{\psi_2}\left\|\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}^\star\right\|_{\psi_2}\le\sigma,$$

where $\|\cdot\|_{\psi_1}$ and $\|\cdot\|_{\psi_2}$ denote the sub-exponential norm and the sub-Gaussian norm, respectively. In view of the Bernstein inequality Vershynin [2018, Theorem 2.8.2], we have

$$\mathbb{P}\left\{\left|\sum_{j=1}^m\xi_j\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}^\star\right|\ge t\right\}\le 2\exp\left(-c\min\left(\frac{\tau^2}{\sigma^2\sum_{j=1}^m\left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{b}_j\right|^2},\frac{\tau}{\sigma\max_{1\le j\le m}\left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{b}_j\right|}\right)\right)\tag{A.44}$$

for any $\tau>0$. Recognizing that

$$\sum_{j=1}^m\left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{b}_j\right|^2=\boldsymbol{b}_l^{\mathsf{H}}\Big(\sum_{j=1}^m\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\Big)\boldsymbol{b}_l=\frac{K}{m}\qquad\text{and}\qquad\max_{1\le j\le m}\left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{b}_j\right|\le\max_{1\le j\le m}\|\boldsymbol{b}_l\|_2\|\boldsymbol{b}_j\|_2=\frac{K}{m}$$

and setting $\tau=C\sigma\sqrt{\frac{K}{m}\log m}$ for some large enough constant $C>0$, one obtains

$$\mathbb{P}\left\{\varsigma_2\ge C\sigma\sqrt{\frac{K}{m}\log m}\right\}\le 2\exp\left(-c\min\left(C^2\log m, C\sqrt{\frac{m\log m}{K}}\right)\right)\lesssim m^{-100},$$

provided that $m\gg K\log m$.

- Combining the above two pieces implies that, with probability exceeding $1-O\left(m^{-100}\right)$,

$$\left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{\nu}_4\right|\lesssim\sigma\log^{1.5}m\max_{1\le j\le m}\left|\boldsymbol{a}_j^{\mathsf{H}}(\widetilde{\boldsymbol{x}}^t-\boldsymbol{x}^\star)\right|+\sigma\sqrt{\frac{K}{m}\log m},\tag{A.45}$$
$$\le\sigma\log^{1.5}m\cdot C_3\left(\sqrt{\frac{\mu^2K\log^2m}{m}}+\sqrt{\log m}\left(\lambda+\sigma\sqrt{K\log m}\right)\right)+\sigma\sqrt{\frac{K}{m}\log m}$$
$$\lesssim C_3\sigma.\tag{A.46}$$

where the penultimate inequality follows from the hypothesis (A.6b), and the last line holds as long as $m\gg\mu^2K\log^5m$, $\sigma\sqrt{K\log^5m}\ll 1$.

Combining the bounds (A.43) with (A.42) and (A.45), we arrive at

$$\left|\boldsymbol{b}_l^{\mathsf{H}}\widetilde{\boldsymbol{h}}^{t+1}\right|\le(1+\delta)\left(1-\eta\lambda-\frac{\eta}{|\alpha^t|^2}\|\boldsymbol{x}^\star\|_2^2\right)\left|\boldsymbol{b}_l^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right|+(1+\delta)\,0.3\frac{\eta}{|\alpha^t|^2}\max_{1\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right|$$
$$+(1+\delta)\frac{\eta}{|\alpha^t|^2}\times C\left(\frac{\mu}{\sqrt{m}}+\frac{\mu}{\sqrt{m}}\log^{3/2}m\max_{1\le j\le m}\left|\boldsymbol{a}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^t-\boldsymbol{x}^\star\right)\right|\right)$$
$$+(1+\delta)\frac{\eta}{|\alpha^t|^2}\max_{1\le j\le\tau}\left|\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right|\log m+(1+\delta)\frac{\eta}{|\alpha^t|^2}\left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{\nu}_4\right|$$
$$\le C_4\left(\frac{\mu}{\sqrt{m}}\log^2m+\sigma\right),$$

as long as $m\gg\mu^2K\log^9m$ for some large enough constant $C_4\gg C_3$. Here, the last inequality invokes the induction hypotheses (A.6) at the tth iteration, Claim 1, as well as the fact $|\alpha^t|\asymp 1$ (cf. Corollary 1).

A.9.1 Proof of Claim 1

To begin with, we make the observation that

$$\left|(\boldsymbol{b}_j-\boldsymbol{b}_1)^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right| = \left|(\boldsymbol{b}_j-\boldsymbol{b}_1)^{\mathsf{H}}\widetilde{\boldsymbol{h}}^{t-1/2}\right| = \left|\frac{\alpha^{t-1}}{\alpha^{t-1/2}}\right|\left|(\boldsymbol{b}_j-\boldsymbol{b}_1)^{\mathsf{H}}\frac{\boldsymbol{h}^{t-1/2}}{\overline{\alpha^{t-1}}}\right|$$
$$\leq (1+\delta)\left|(\boldsymbol{b}_j-\boldsymbol{b}_1)^{\mathsf{H}}\frac{\boldsymbol{h}^{t-1/2}}{\overline{\alpha^{t-1}}}\right|,$$

with $\delta \ll 1$ defined in (A.41). This inequality allows us to turn attention to $\frac{1}{\overline{\alpha^{t-1}}}(\boldsymbol{b}_j-\boldsymbol{b}_1)^{\mathsf{H}}\boldsymbol{h}^{t-1/2}$ instead.

Use the gradient update rule with respect to $\boldsymbol{h}^t$, we obtain

$$\frac{1}{\overline{\alpha^{t-1}}}\boldsymbol{h}^{t-1/2} = \frac{1}{\overline{\alpha^{t-1}}}\left(\boldsymbol{h}^{t-1}-\eta\left(\sum_{l=1}^{m}\boldsymbol{b}_l\boldsymbol{b}_l^{\mathsf{H}}\left(\boldsymbol{h}^{t-1}\boldsymbol{x}^{t-1\mathsf{H}}-\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\right)\boldsymbol{a}_l\boldsymbol{a}_l^{\mathsf{H}}\boldsymbol{x}^{t-1}-\sum_{l=1}^{m}\xi_l\boldsymbol{b}_l\boldsymbol{a}_l^{\mathsf{H}}\boldsymbol{x}^{t-1}+\lambda\boldsymbol{h}^{t-1}\right)\right).$$

Therefore, one can decompose

$$\begin{aligned}(\boldsymbol{b}_j-\boldsymbol{b}_1)^{\mathsf{H}}\frac{1}{\overline{\alpha^{t-1}}}\boldsymbol{h}^t &= \left(1-\eta\lambda-\frac{\eta}{|\alpha^t|^2}\left\|\widetilde{\boldsymbol{x}}^{t-1}\right\|_2^2\right)(\boldsymbol{b}_j-\boldsymbol{b}_1)^{\mathsf{H}}\widetilde{\boldsymbol{h}}^{t-1}+\frac{\eta}{|\alpha^t|^2}\underbrace{(\boldsymbol{b}_j-\boldsymbol{b}_1)\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\widetilde{\boldsymbol{x}}^{t-1}}_{=:\beta_1}\\ &\quad-\frac{\eta}{|\alpha^t|^2}\underbrace{(\boldsymbol{b}_j-\boldsymbol{b}_1)^{\mathsf{H}}\sum_{l=1}^{m}\boldsymbol{b}_l\boldsymbol{b}_l^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^{t-1}\widetilde{\boldsymbol{x}}^{t-1\mathsf{H}}-\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\right)\left(\boldsymbol{a}_l\boldsymbol{a}_l^{\mathsf{H}}-\boldsymbol{I}_k\right)\widetilde{\boldsymbol{x}}^{t-1}}_{=:\beta_2}\\ &\quad+\frac{\eta}{|\alpha^t|^2}\underbrace{(\boldsymbol{b}_j-\boldsymbol{b}_1)^{\mathsf{H}}\sum_{l=1}^{m}\xi_l\boldsymbol{b}_l\boldsymbol{a}_l^{\mathsf{H}}\widetilde{\boldsymbol{x}}^{t-1}}_{=:\beta_3}.\end{aligned}\tag{A.47}$$

Except β_3, the bounds of the other terms can be obtained by the same arguments as in Ma et al. [2018, Appendix C.4.3]; we thus omit the detailed proof but only list the results below:

$$|\beta_1| \leq 4\frac{\mu}{\sqrt{m}}$$
$$|\beta_2| \leq \frac{c}{\log m}\left(\max_{1\leq l\leq m}\left|\boldsymbol{b}_l^{\mathsf{H}}\widetilde{\boldsymbol{h}}^{t-1}\right|+\frac{\mu}{\sqrt{m}}\right)$$

with c some small constant $c>0$, as long as $m \gg K\log^8 m$. When it comes to the remaining term β_3, the triangle inequality yields

$$|\beta_3| \leq \underbrace{\left|\sum_{l=1}^{m}\xi_l(\boldsymbol{b}_j-\boldsymbol{b}_1)^{\mathsf{H}}\boldsymbol{b}_l\boldsymbol{a}_l^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^{t-1}-\boldsymbol{x}^{\star}\right)\right|}_{=:\omega_1}+\underbrace{\left|\sum_{l=1}^{m}\xi_l(\boldsymbol{b}_j-\boldsymbol{b}_1)^{\mathsf{H}}\boldsymbol{b}_l\boldsymbol{a}_l^{\mathsf{H}}\boldsymbol{x}^{\star}\right|}_{=:\omega_2}.$$

- Regarding ω_1, we have

$$\begin{aligned}\omega_1 &\leq \sum_{j=1}^{m}\left|(\boldsymbol{b}_j-\boldsymbol{b}_1)^{\mathsf{H}}\boldsymbol{b}_l\right|\cdot\max_{1\leq j\leq m}|\xi_j|\cdot\max_{1\leq j\leq m}\left|\boldsymbol{a}_j^{\mathsf{H}}(\widetilde{\boldsymbol{x}}^t-\boldsymbol{x}^{\star})\right|\\ &\lesssim \frac{1}{\log^2 m}\cdot\sigma\sqrt{\log m}\cdot\max_{1\leq j\leq m}\left|\boldsymbol{a}_j^{\mathsf{H}}(\widetilde{\boldsymbol{x}}^t-\boldsymbol{x}^{\star})\right|\\ &\lesssim \frac{\sigma}{\log^{1.5} m}\max_{1\leq j\leq m}\left|\boldsymbol{a}_j^{\mathsf{H}}(\widetilde{\boldsymbol{x}}^t-\boldsymbol{x}^{\star})\right|,\end{aligned}$$

  where the second inequality follows from Ma et al. [2018, Lemma 50] and standard sub-Gaussian concentration inequalities.

- For ω_2, similar to (A.44), we can invoke the Bernstein inequality Vershynin [2018, Theorem 2.8.2] to reach

$$\begin{aligned}
&\mathbb{P}\left\{\left|\sum_{l=1}^{m}\xi_l\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\boldsymbol{b}_l\boldsymbol{a}_l^{\mathsf{H}}\boldsymbol{x}^{\star}\right|\geq\tau\right\} \\
&\qquad\leq 2\exp\left(-c\min\left(\frac{\tau^2}{\sigma^2\sum_{l=1}^{m}\left|\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\boldsymbol{b}_l\right|^2},\frac{\tau}{\sigma\max_{1\leq j\leq m}\left|\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\boldsymbol{b}_l\right|}\right)\right)
\end{aligned} \tag{A.48}$$

for any $\tau\geq 0$. In addition, observe that

$$\begin{aligned}
\sum_{j=1}^{m}\left|\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\boldsymbol{b}_l\right|^2 &\leq\left\{\max_{1\leq j\leq m}\left|\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\boldsymbol{b}_l\right|\right\}\cdot\sum_{j=1}^{m}\left|\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\boldsymbol{b}_l\right| \\
&\leq 2\frac{K}{m}\cdot\frac{c}{\log^2 m},
\end{aligned}$$

where the last inequality follows from Ma et al. [2018, Lemma 48, 49]. Taking $\tau=C\sigma\sqrt{K\log^2 m/m}$ in (A.48) for some large enough constant $C>0$, one arrives at

$$\mathbb{P}\left\{\omega_2\geq C\sigma\sqrt{\frac{K\log m}{m}}\right\}\leq 2\exp\left(-c\min\left(C^2\log^3 m, C\frac{m}{K}\sqrt{\log m}\right)\right)\lesssim m^{-100}.$$

- The above bounds taken collectively imply that: with probability exceeding $1-O\left(m^{-100}\right)$,

$$\begin{aligned}
|\beta_3| &\lesssim\frac{\sigma}{\log^{1.5}m}\max_{1\leq j\leq m}\left|\boldsymbol{a}_j^{\mathsf{H}}(\widetilde{\boldsymbol{x}}^t-\boldsymbol{x}^{\star})\right|+\sigma\sqrt{\frac{K\log m}{m}} \\
&\lesssim C_3\frac{\sigma}{\log^{1.5}m}\left(\sqrt{\frac{\mu^2K\log^2 m}{m}}+\sqrt{\log m}\left(\lambda+\sigma\sqrt{K\log m}\right)\right)+\sigma\sqrt{\frac{K\log m}{m}} \\
&\lesssim\frac{\sigma}{\log^3 m}.
\end{aligned} \tag{A.49}$$

Putting together the above results, we demonstrate that

$$\begin{aligned}
\left|\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right| &\leq(1+\delta)\left(1-\eta\lambda-\frac{\eta}{|\alpha^t|^2}\left\|\widetilde{\boldsymbol{x}}^{t-1}\right\|_2^2\right)\left|\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\widetilde{\boldsymbol{h}}^{t-1}\right|+4(1+\delta)\frac{\eta}{|\alpha^t|^2}\frac{\mu}{\sqrt{m}} \\
&\quad+c(1+\delta)\frac{\eta}{|\alpha^t|^2}\frac{1}{\log m}\left[\max_{1\leq l\leq m}\left|\boldsymbol{b}_l^{\mathsf{H}}\widetilde{\boldsymbol{h}}^{t-1}\right|+\frac{\mu}{\sqrt{m}}\right]+(1+\delta)\frac{\eta}{|\alpha^t|^2}\frac{\sigma}{\log^3 m} \\
&\leq cC_4\left(\frac{\mu}{\sqrt{m}}\log m+\frac{\sigma}{\log m}\right)
\end{aligned}$$

if $\eta>0$ is sufficiently small, where the last inequality utilizes $\|\widetilde{\boldsymbol{x}}^{t-1}\|_2\asymp 1$ and $|\alpha^t|\asymp 1$ in Lemma 3.

B Analysis under Fourier design: connections between convex and nonconvex solutions

B.1 Proof outline for Theorem 1

As the empirical evidence (cf. Figure 1) suggests, an approximate nonconvex optimizer produced by a simple gradient-type algorithm is exceedingly close to the convex minimizer of (3). In what follows, we shall start by introducing an auxiliary nonconvex gradient method, and formalize its connection to the convex program. Without loss of generality, we assume that $\|\boldsymbol{h}^{\star}\|_2=\|\boldsymbol{x}^{\star}\|_2=1$ throughout the proof.

An auxiliary nonconvex algorithm. Let us consider the iterates obtained by running a variant of (Wirtinger) gradient descent, as summarized in Algorithm 2. A crucial difference from Algorithm 1 lies in the initialization stage — namely, Algorithm 2 initializes the algorithm from the ground truth $(\boldsymbol{h}^\star, \boldsymbol{x}^\star)$ rather than a spectral estimate as adopted in Algorithm 1. While initialization at the truth is not practically implementable, it is introduced here solely for analytical purpose, namely, it creates a sequence of ancillary random variables that approximate our estimators and are close to the ground truth. This is how we establish the convergence rate of our estimators.

Algorithm 2 Auxiliary gradient descent for blind deconvolution (for analysis purpose only)

Input: $\{\boldsymbol{a}_j\}_{1\le j\le m}$, $\{\boldsymbol{b}_j\}_{1\le j\le m}$, $\{y_j\}_{1\le j\le m}$, $\boldsymbol{h}^\star$ and $\boldsymbol{x}^\star$.
Initialization: $\boldsymbol{h}^0 = \boldsymbol{h}^\star$ and $\boldsymbol{x}^0 = \boldsymbol{x}^\star$.
Gradient updates: **for** $t = 0, 1, \ldots, t_0 - 1$ **do**

$$\begin{bmatrix} \boldsymbol{h}^{t+1/2} \\ \boldsymbol{x}^{t+1/2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{h}^{t} \\ \boldsymbol{x}^{t} \end{bmatrix} - \eta \begin{bmatrix} \nabla_{\boldsymbol{h}} f\left(\boldsymbol{h}^t, \boldsymbol{x}^t\right) \\ \nabla_{\boldsymbol{x}} f\left(\boldsymbol{h}^t, \boldsymbol{x}^t\right) \end{bmatrix}, \tag{B.1a}$$

$$\begin{bmatrix} \boldsymbol{h}^{t+1} \\ \boldsymbol{x}^{t+1} \end{bmatrix} = \begin{bmatrix} \sqrt{\frac{\left\|\boldsymbol{x}^{t+1/2}\right\|_2}{\left\|\boldsymbol{h}^{t+1/2}\right\|_2}}\boldsymbol{h}^{t+1/2} \\ \sqrt{\frac{\left\|\boldsymbol{h}^{t+1/2}\right\|_2}{\left\|\boldsymbol{x}^{t+1/2}\right\|_2}}\boldsymbol{x}^{t+1/2} \end{bmatrix}, \tag{B.1b}$$

where $\nabla_{\boldsymbol{h}} f(\cdot)$ and $\nabla_{\boldsymbol{x}} f(\cdot)$ represent the Wirtinger gradient (see [Li et al., 2019, Section 3.3] and Appendix A.2.1) of $f(\cdot)$ w.r.t. $\boldsymbol{h}$ and $\boldsymbol{x}$, respectively.

Properties of the auxiliary nonconvex algorithm. The trajectory of this auxiliary nonconvex algorithm enjoys several important properties. In the following lemma, the results are stated for the properly rescaled iterate

$$\widetilde{\boldsymbol{z}}^t = \left(\widetilde{\boldsymbol{h}}^t, \widetilde{\boldsymbol{x}}^t\right) := \left(\frac{1}{\overline{\alpha^t}}\boldsymbol{h}^t, \alpha^t \boldsymbol{x}^t\right),$$

with alignment parameter defined by

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

Lemma 11. *Take $\lambda = C_\lambda \sigma\sqrt{K\log m}$ for some large enough constant $C_\lambda > 0$. Assume the number of measurements obeys $m \ge C\mu^2 K \log^9 m$ for some sufficiently large constant $C > 0$, and the noise satisfies $\sigma\sqrt{K\log m} \le c/\log^2 m$ for some sufficiently small constant $c > 0$. Then, with probability at least $1 - O\left(m^{-100} + me^{-cK}\right)$ for some constant $c > 0$, the iterates $\{\boldsymbol{h}^t, \boldsymbol{x}^t\}_{0<t\le t_0}$ of Algorithm (2) satisfy*

$$\mathsf{dist}\left(\boldsymbol{z}^t, \boldsymbol{z}^\star\right) \le \rho \mathsf{dist}\left(\boldsymbol{z}^{t-1}, \boldsymbol{z}^\star\right) + C_5\eta\left(\lambda + \sigma\sqrt{K\log m}\right) \tag{B.2a}$$

$$\max_{1\le j\le m}\left|\boldsymbol{a}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^t - \boldsymbol{x}^\star\right)\right| \le C_7\sqrt{\log m}\left(\lambda + \sigma\sqrt{K\log m}\right) \tag{B.2b}$$

$$\max_{1\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right| \le C_8\left(\frac{\mu}{\sqrt{m}}\log m + \sigma\right) \tag{B.2c}$$

$$\max_{1\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^t - \boldsymbol{h}^\star\right)\right| \le C_9\sigma \tag{B.2d}$$

for any $0 < t \le t_0$, where $\rho = 1 - c_\rho\eta \in (0,1)$ for some small constant $c_\rho > 0$, and we take $t_0 = m^{20}$. Here, C_5, …, C_9 are constants obeying $C_7 \gg C_5$. In addition, we have

$$\min_{0\le t\le t_0}\left\|\nabla f\left(\boldsymbol{h}^t, \boldsymbol{x}^t\right)\right\|_2 \le \frac{\lambda}{m^{10}}. \tag{B.2e}$$

Most of the inequalities of this lemma (as well as their proofs) resemble the ones derived for Algorithm 1 in Appendix A. It is worth emphasizing, however, that the establishment of the inequality (B.2d) relies heavily on the idealized initialization $(\boldsymbol{h}^0, \boldsymbol{x}^0) = (\boldsymbol{h}^\star, \boldsymbol{x}^\star)$, and the current proof does not work if the algorithm is spectrally initialized. The proof of this lemma is deferred to Appendix B.3.

Connection between the approximate nonconvex minimizer and the convex solution. As it turns out, the above type of features of the nonconvex iterates together with the first-order optimality of the convex program allows us to control the proximity of the convex minimizer and the approximate nonconvex optimizer. Before proceeding to develop this idea formally, we first introduce the following operators for notational convenience. For any $\boldsymbol{z} = [z_j]_{1\leq j\leq m}$ and any $\boldsymbol{Z} \in \mathbb{C}^{K\times K}$, we define

$$\mathcal{A}(\boldsymbol{Z}) \coloneqq \left\{\boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{Z} \boldsymbol{a}_j\right\}_{j=1}^m, \qquad \mathcal{A}^*(\boldsymbol{z}) = \sum_{j=1}^m z_j \boldsymbol{b}_j \boldsymbol{a}_j^{\mathsf{H}},$$
$$\mathcal{T}(\boldsymbol{Z}) \coloneqq \mathcal{A}^*\mathcal{A}(\boldsymbol{Z}) = \sum_{j=1}^m \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{Z} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}}. \tag{B.3}$$

Below are several key conditions on these operators concerned with the interplay between the noise size, the estimation accuracy of the nonconvex estimate $(\boldsymbol{h}, \boldsymbol{x})$ and the regularization parameters λ.

Condition 1. The regularization parameter λ satisfies

1. $\|\mathcal{T}\left(\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right) - \left(\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right)\| < \lambda/8$.
2. $\|\mathcal{A}^*(\boldsymbol{\xi})\| = \|\sum_{j=1}^m \xi_j \boldsymbol{b}_j \boldsymbol{a}_j^{\mathsf{H}}\| \leq c\lambda$, for some small constant $c > 0$.

Condition 1 requires that the regularization parameter λ dominate the norm of the deviation of $\mathcal{T}(\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}})$ from its mean $\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}$, and also the norm of the noise operated on by $\mathcal{A}^*$. As can be seen shortly, these two conditions can be met with high probability when $(\boldsymbol{h}, \boldsymbol{x})$ is sufficiently close to $(\boldsymbol{h}^\star, \boldsymbol{x}^\star)$.

Another critical condition is the following injectivity condition on $\mathcal{A}$.

Condition 2. Let T be the tangent space of $\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}$. Then for all $\boldsymbol{Z} \in T$, one has

$$\|\mathcal{A}(\boldsymbol{Z})\|_2^2 \geq \frac{1}{16}\|\boldsymbol{Z}\|_{\mathrm{F}}^2.$$

When these two conditions hold, the aforementioned intimate connection between approximate nonconvex minimizer and the convex solution can be formalized in the following crucial lemma.

Lemma 12. *Suppose that* $(\boldsymbol{h}, \boldsymbol{x})$ *obeys*

$$\|\nabla f(\boldsymbol{h}, \boldsymbol{x})\|_2 \leq C\frac{\lambda}{m^{10}}, \tag{B.4a}$$

for some constants $C > 0$. *Then under Conditions 1 and 2, any minimizer* $\boldsymbol{Z}_{\mathsf{cvx}}$ *of the convex problem (3) satisfies*

$$\left\|\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} - \boldsymbol{Z}_{\mathsf{cvx}}\right\|_{\mathrm{F}} \lesssim \|\nabla f(\boldsymbol{h}, \boldsymbol{x})\|_2.$$

Proof. See Appendix B.4. □

In words, if we can find a point $(\boldsymbol{h}, \boldsymbol{x})$ that has vanishingly small gradient (cf. (B.4a)) and that satisfies the additional Conditions 1 and 2, then the matrix $\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}$ is guaranteed to be exceedingly close to the solution of the convex program. Encouragingly, Lemma 11 hints at the existence of a point along the trajectory of Algorithm (2) satisfying these conditions (B.5); if this were true, then one could transfer the properties of the approximate nonconvex optimizer to the convex solution, as a means to certify the statistical efficiency of convex programming. As we will see soon, this is indeed the case that with Assumption 1, we can prove that under some mild sample size and noise level conditions, Conditions 1 and 2 would hold with high probability. To begin with, the following lemma corresponds to the first point in Condition 1.

Lemma 13. *Suppose that the sample complexity satisfies $m \geq C\mu^2 K \log^4 m$ for some sufficiently large constant $C > 0$. Take $\lambda = C_\lambda \sigma\sqrt{K\log m}$ for some large enough constant $C_\lambda > 0$. Then with probability at least $1 - O\left(m^{-10} + me^{-CK}\right)$, we have*

$$\left\|\mathcal{T}\left(\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right) - \left(\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right)\right\| < \lambda/8,$$

simultaneously for any $(\boldsymbol{h}, \boldsymbol{x})$ obeying

$$\|\boldsymbol{h}\|_2 = \|\boldsymbol{x}\|_2, \quad \|\boldsymbol{h} - \boldsymbol{h}^\star\|_2 \leq \frac{C_5}{1-\rho}\eta\left(\lambda + \sigma\sqrt{K\log m}\right), \quad \|\boldsymbol{x} - \boldsymbol{x}^\star\|_2 \leq \frac{C_5}{1-\rho}\eta\left(\lambda + \sigma\sqrt{K\log m}\right), \tag{B.5a}$$

$$\max_{1\leq j\leq m}\left|\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{h} - \boldsymbol{h}^\star\right)\right| \leq C_9\sigma \quad \text{and} \quad \max_{1\leq j\leq m}\left|\boldsymbol{a}_j^{\mathsf{H}}\left(\boldsymbol{x} - \boldsymbol{x}^\star\right)\right| \leq C_7\sqrt{\log m}\left(\lambda + \sigma\sqrt{K\log m}\right), \tag{B.5b}$$

for some constants $C_5, C_7, C_9 > 0$.

Proof. See Appendix B.5. ☐

Recall the definition of operator $\mathcal{T}$ in (B.3). The lemma above states that for all $(\boldsymbol{h}, \boldsymbol{x})$ sufficiently close to $(\boldsymbol{h}^\star, \boldsymbol{x}^\star)$, the matrix $\mathcal{T}\left(\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right)$ is close to the expectation $\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}$.

Next we turn to the second point in Condition 1.

Lemma 14. *Suppose that Asumption 1 holds and $m \gtrsim K\log^3 m$. With probability at least $1 - O\left(m^{-100}\right)$, one has*

$$\|\mathcal{A}^*\left(\boldsymbol{\xi}\right)\| = \left\|\sum_{j=1}^m \xi_j \boldsymbol{b}_j \boldsymbol{a}_j^{\mathsf{H}}\right\| \lesssim \sigma\sqrt{K\log m}.$$

Proof. See Appendix B.6. ☐

Regarding Condition 2, we have the following lemma.

Lemma 15. *Suppose that the sample complexity satisfies $m \geq C\mu^2 K\log m$ for some sufficiently large constant $C > 0$. Then with probability at least $1 - O\left(m^{-10}\right)$,*

$$\|\mathcal{A}\left(\boldsymbol{Z}\right)\|_2^2 \geq \frac{1}{16}\|\boldsymbol{Z}\|_{\mathrm{F}}^2, \quad \forall \boldsymbol{Z} \in T$$

holds simultaneously for all T for which the associated point $(\boldsymbol{h}, \boldsymbol{x})$ obeys (B.5a) and (B.5b). Here, T denotes the tangent space of $\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}$.

Proof. See Appendix B.7. ☐

Basically, this lemma reveals that when $(\boldsymbol{h}, \boldsymbol{x})$ is sufficiently close to $(\boldsymbol{h}^\star, \boldsymbol{x}^\star)$, the operator $\mathcal{A}(\cdot)$ — restricted to the tangent space T of $\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}$ — is injective.

Now we are ready to present the proof of Theorem 1.

Proof of Theorem 1. Armed with this result and the properties about the nonconvex trajectory, we are ready to establish Theorem 1 as follows. Let $\bar{t} := \arg\min_{0\leq t\leq t_0}\|\nabla f(\boldsymbol{h}^t, \boldsymbol{x}^t)\|_{\mathrm{F}}$, and take $(\boldsymbol{h}_{\mathsf{ncvx}}, \boldsymbol{x}_{\mathsf{ncvx}}) = \left(\frac{1}{\overline{\alpha^{\bar{t}}}}\boldsymbol{h}^{\bar{t}}, \alpha^{\bar{t}}\boldsymbol{x}^{\bar{t}}\right)$. By virtue of Lemma 11, we see that $(\boldsymbol{h}_{\mathsf{ncvx}}, \boldsymbol{x}_{\mathsf{ncvx}})$ satisfies — with high probability — the small gradient property (B.2e) as well as all conditions required to invoke Lemma 12. As a consequence, invoke Lemma 12 to obtain

$$\left\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{h}_{\mathsf{ncvx}}\boldsymbol{x}_{\mathsf{ncvx}}^{\mathsf{H}}\right\|_{\mathrm{F}} \lesssim \frac{1}{c_{\mathrm{inj}}}\left\|\nabla f\left(\boldsymbol{h}_{\mathsf{ncvx}}, \boldsymbol{x}_{\mathsf{ncvx}}\right)\right\|_{\mathrm{F}} \lesssim \frac{\lambda}{m^{10}}. \tag{B.6}$$

Further, it is seen that

$$\left\|\boldsymbol{h}_{\mathsf{ncvx}}\left(\boldsymbol{x}_{\mathsf{ncvx}}\right)^{\mathsf{H}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}} \leq \left\|\boldsymbol{h}_{\mathsf{ncvx}}\left(\boldsymbol{x}_{\mathsf{ncvx}}\right)^{\mathsf{H}} - \boldsymbol{h}^\star\left(\boldsymbol{x}_{\mathsf{ncvx}}\right)^{\mathsf{H}}\right\|_{\mathrm{F}} + \left\|\boldsymbol{h}^\star\left(\boldsymbol{x}_{\mathsf{ncvx}}\right)^{\mathsf{H}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}}$$

$$
\begin{aligned}
&\le \|\boldsymbol{h}_{\mathsf{ncvx}} - \boldsymbol{h}^\star\|_2 \|\boldsymbol{x}_{\mathsf{ncvx}}\|_2 + \|\boldsymbol{h}^\star\|_2 \|\boldsymbol{x}_{\mathsf{ncvx}} - \boldsymbol{x}^\star\|_2 \\
&\le 2\|\boldsymbol{z}^\star\|_2 \cdot \frac{C_5 \eta}{(1-\rho)\|\boldsymbol{z}^\star\|_2}\left(\lambda + \sigma\sqrt{K\log m}\right) \\
&= \frac{2C_5}{c_\rho}\left(\lambda + \sigma\sqrt{K\log m}\right), \qquad \text{(B.7)}
\end{aligned}
$$

where the penultimate line follows from (B.8a) and the inequality

$$
\|\boldsymbol{x}_{\mathsf{ncvx}}\|_2 \le \|\boldsymbol{x}^\star\|_2 + \|\boldsymbol{x}_{\mathsf{ncvx}} - \boldsymbol{x}^\star\|_2 \le \|\boldsymbol{z}^\star\|_2 + \frac{C_5\eta}{(1-\rho)\|\boldsymbol{z}^\star\|_2}\left(\lambda + \sigma\sqrt{K\log m}\right) \le 2\|\boldsymbol{z}^\star\|_2 .
$$

Taking (B.6) and (B.7) collectively yields

$$
\begin{aligned}
\left\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}} &\le \left\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{h}_{\mathsf{ncvx}}\boldsymbol{x}_{\mathsf{ncvx}}^{\mathsf{H}}\right\|_{\mathrm{F}} + \left\|\boldsymbol{h}_{\mathsf{ncvx}}\boldsymbol{x}_{\mathsf{ncvx}}^{\mathsf{H}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}} \\
&\lesssim \frac{\lambda}{m^{10}} + \lambda + \sigma\sqrt{K\log m} \\
&\lesssim \lambda + \sigma\sqrt{K\log m}.
\end{aligned}
$$

This together with the elementary bound $\left\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right\| \le \left\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}}$ concludes the proof, as long as the above key lemmas can be justified.

To prove the results also holds for $\boldsymbol{Z}_{\mathsf{cvx},1}$, we recall that $\boldsymbol{Z}_{\mathsf{cvx},1}$ is the best rank-1 approximation of $\boldsymbol{Z}_{\mathsf{cvx}}$ and this implies that,

$$
\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{Z}_{\mathsf{cvx},1}\|_{\mathrm{F}} \le \left\|\boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{h}_{\mathsf{ncvx}}\boldsymbol{x}_{\mathsf{ncvx}}^{\mathsf{H}}\right\|_{\mathrm{F}} \lesssim \frac{\lambda}{m^{10}}.
$$

Hence, repeating the above calculations for $\boldsymbol{Z}_{\mathsf{cvx},1}$ reveals that (14) continues to holds if $\boldsymbol{Z}_{\mathsf{cvx}}$ is replaced by $\boldsymbol{Z}_{\mathsf{cvx},1}$.

In what follows, we establish the key lemmas stated above.

B.2 Preliminary facts

Before proceeding, there are a couple of immediate consequences of Lemma 11 that will prove useful, which we summarize as follows.

Lemma 16. *Instate the notation and assumptions in Theorem 2. For $t \ge 0$, suppose that the hypotheses (B.9) hold in the first t iterations. Then there exist some constants $C_5 > 0$ such that for any $1 \le l \le m$,*

$$
\mathsf{dist}\left(\boldsymbol{z}^t, \boldsymbol{z}^\star\right) \le \frac{C_5}{c_\rho}\left(\lambda + \sigma\sqrt{K\log m}\right), \qquad \text{(B.8a)}
$$

$$
\left\|\widetilde{\boldsymbol{z}}^{t,(l)} - \boldsymbol{z}^\star\right\|_2 \le 2\frac{C_5}{c_\rho}\left(\lambda + \sigma\sqrt{K\log m}\right), \qquad \text{(B.8b)}
$$

$$
\frac{1}{2} \le \left\|\widetilde{\boldsymbol{x}}^t\right\|_2 \le \frac{3}{2}, \qquad \frac{1}{2} \le \left\|\widetilde{\boldsymbol{h}}^t\right\|_2 \le \frac{3}{2}, \qquad \text{(B.8c)}
$$

$$
\frac{1}{2} \le \left\|\widetilde{\boldsymbol{x}}^{t,(l)}\right\|_2 \le \frac{3}{2}, \qquad \frac{1}{2} \le \left\|\widetilde{\boldsymbol{h}}^{t,(l)}\right\|_2 \le \frac{3}{2}, \qquad \text{(B.8d)}
$$

$$
\frac{1}{2} \le \left\|\widehat{\boldsymbol{x}}^{t,(l)}\right\|_2 \le \frac{3}{2}, \qquad \frac{1}{2} \le \left\|\widehat{\boldsymbol{h}}^{t,(l)}\right\|_2 \le \frac{3}{2}, \qquad \text{(B.8e)}
$$

$$
\left\|\boldsymbol{h}^t\right\|_2^2 = \left\|\boldsymbol{x}^t\right\|_2^2 = \left\|\boldsymbol{h}^t\right\|_2\left\|\boldsymbol{x}^t\right\|_2 = \left\|\widetilde{\boldsymbol{h}}^{t-1/2}\right\|_2\left\|\widetilde{\boldsymbol{x}}^{t-1/2}\right\|_2 = \left\|\widetilde{\boldsymbol{h}}^t\right\|_2\left\|\widetilde{\boldsymbol{x}}^t\right\|_2. \qquad \text{(B.8f)}
$$

In addition, for an integer $t > 0$, suppose that the hypotheses (B.9) hold in the first $t-1$ iterations. Then there exists some constant $C > 0$ such that with probability at least $1 - O\left(m^{-100} + e^{-CK}\log m\right)$, there holds

$$
\left\|\widehat{\boldsymbol{z}}^t - \boldsymbol{z}^\star\right\|_2 \le \frac{C_5}{c_\rho}\left(\lambda + \sigma\sqrt{K\log m}\right), \qquad \text{(B.8g)}
$$

$$
\left|\left|\alpha^t\right| - 1\right| \lesssim \frac{C_5}{c_\rho}\left(\lambda + \sigma\sqrt{K\log m}\right), \qquad \text{(B.8h)}
$$

$$\left|\frac{\alpha^{t-1/2}}{\alpha^{t-1}}-1\right| \lesssim \eta\frac{C_5}{c_\rho}\left(\lambda+\sigma\sqrt{K\log m}\right), \tag{B.8i}$$

$$\left|\alpha^{t-1/2}-\alpha^{t-1}\right| \lesssim \eta\frac{C_5}{c_\rho}\left(\lambda+\sigma\sqrt{K\log m}\right), \tag{B.8j}$$

$$\frac{1}{2} \le \left|\frac{\alpha^{t-1}}{\alpha^{t-1/2}}\right| \le \frac{3}{2}, \tag{B.8k}$$

$$\frac{1}{2} \le \left|\alpha^{t}\right| \le \frac{3}{2}. \tag{B.8l}$$

Proof. The proof follows from the same argument as in the proof of Lemma 3 and Corollary 1, and is thus omitted here for brevity. □

B.3 Proof of Lemma 11

After the introduction of the proof idea in Appendix A, we state a more complete version of Lemma 11 here.

Lemma 17. *Take $\lambda = C_\lambda\sigma\sqrt{K\log m}$ for some large enough constant $C_\lambda > 0$. Assume the number of measurements obeys $m \ge C\mu^2 K\log^9 m$ for some sufficiently large constant $C > 0$, and the noise satisfies $\sigma\sqrt{K\log m} \le c/\log^2 m$ for some sufficiently small constant $c > 0$. Then, with probability at least $1 - O\left(m^{-100} + me^{-cK}\right)$ for some constant $c > 0$, the iterates $\{\boldsymbol{h}^t, \boldsymbol{x}^t\}_{0<t\le t_0}$ of Algorithm (2) satisfy*

$$\mathsf{dist}\left(\boldsymbol{z}^t, \boldsymbol{z}^\star\right) \le \rho\,\mathsf{dist}\left(\boldsymbol{z}^{t-1}, \boldsymbol{z}^\star\right) + C_5\eta\left(\lambda+\sigma\sqrt{K\log m}\right) \tag{B.9a}$$

$$\max_{1\le l\le m} \mathsf{dist}\big(\boldsymbol{z}^{t,(l)}, \widetilde{\boldsymbol{z}}^t\big) \le C_6\frac{\sigma}{\log^2 m} \tag{B.9b}$$

$$\max_{1\le l\le m} \left\|\widetilde{\boldsymbol{z}}^{t,(l)} - \widetilde{\boldsymbol{z}}^t\right\|_2 \lesssim C_6\frac{\sigma}{\log^2 m} \tag{B.9c}$$

$$\max_{1\le j\le m} \left|\boldsymbol{a}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^t - \boldsymbol{x}^\star\right)\right| \le C_7\sqrt{\log m}\left(\lambda+\sigma\sqrt{K\log m}\right) \tag{B.9d}$$

$$\max_{1\le j\le m} \left|\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right| \le C_8\left(\frac{\mu}{\sqrt{m}}\log m + \sigma\right) \tag{B.9e}$$

$$\max_{1\le j\le m} \left|\boldsymbol{b}_j^{\mathsf{H}}\big(\widetilde{\boldsymbol{h}}^t - \boldsymbol{h}^\star\big)\right| \le C_9\sigma \tag{B.9f}$$

for any $0 < t \le t_0$, where $\rho = 1 - c_\rho\eta \in (0,1)$ for some small constant $c_\rho > 0$, and we take $t_0 = m^{20}$. Here, C_5, …, C_9 are constants obeying $C_7 \gg C_5$. In addition, we have

$$\min_{0\le t\le t_0} \left\|\nabla f\left(\boldsymbol{h}^t, \boldsymbol{x}^t\right)\right\|_2 \le \frac{\lambda}{m^{10}}. \tag{B.9g}$$

The claims (B.9a)-(B.9e) are direct consequences of Lemma 5, Lemma 6, the relation (A.21), and Lemma 7. As a result, the remaining steps lie in proving (B.2d) and (B.2e).

B.3.1 Proof of the claim (B.2d)

Recall the definition $\widetilde{\boldsymbol{h}}^t := \boldsymbol{h}^t/\overline{\alpha^t}$. We aim to prove inductively that

$$\max_{1\le j\le m} \left|\boldsymbol{b}_j^{\mathsf{H}}\big(\widetilde{\boldsymbol{h}}^t - \boldsymbol{h}^\star\big)\right| \le C_9\sigma \tag{B.10}$$

holds for some constant $C_9 > 0$, provided that the algorithm is initialized at the truth.

It is self-evident that (B.10) holds for the base case (i.e. $t = 0$) when $\boldsymbol{h}^0 = \boldsymbol{h}^\star$. Assume for the moment that (B.10) holds true at the tth iteration. In view of the simple relation between α^{t+1} and $\alpha^{t+1/2}$ in (A.3) and the balancing step (B.1), one has

$$\alpha^{t+1} = \sqrt{\frac{\left\|\boldsymbol{x}^{t+1/2}\right\|_2}{\left\|\boldsymbol{h}^{t+1/2}\right\|_2}}\,\alpha^{t+1/2}, \qquad \text{and} \qquad \boldsymbol{h}^{t+1} = \sqrt{\frac{\left\|\boldsymbol{x}^{t+1/2}\right\|_2}{\left\|\boldsymbol{h}^{t+1/2}\right\|_2}}\,\boldsymbol{h}^{t+1/2}.$$

It then follows that $\boldsymbol{h}^{t+1}/\overline{\alpha^{t+1}} = \boldsymbol{h}^{t+1/2}/\overline{\alpha^{t+1/2}}$ and, therefore,

$$
\begin{aligned}
&\frac{\overline{\alpha^{t+1/2}}}{\overline{\alpha^{t}}}\left(\frac{\boldsymbol{h}^{t+1}}{\overline{\alpha^{t+1}}}-\boldsymbol{h}^{\star}\right)=\frac{\overline{\alpha^{t+1/2}}}{\overline{\alpha^{t}}}\left(\frac{\boldsymbol{h}^{t+1/2}}{\overline{\alpha^{t+1/2}}}-\boldsymbol{h}^{\star}\right)\\
&\overset{\text{(i)}}{=}\frac{\overline{\alpha^{t+1/2}}}{\overline{\alpha^{t}}}\left(\frac{1}{\overline{\alpha^{t+1/2}}}\left(\boldsymbol{h}^{t}-\eta\nabla_{\boldsymbol{h}}f\left(\boldsymbol{h}^{t},\boldsymbol{x}^{t}\right)\right)-\boldsymbol{h}^{\star}\right)\\
&=\widetilde{\boldsymbol{h}}^{t}-\frac{\eta}{\left|\alpha^{t}\right|^{2}}\nabla_{\boldsymbol{h}}f\big(\widetilde{\boldsymbol{h}}^{t},\widetilde{\boldsymbol{x}}^{t}\big)-\frac{\overline{\alpha^{t+1/2}}}{\overline{\alpha^{t}}}\boldsymbol{h}^{\star}\\
&\overset{\text{(ii)}}{=}\left(1-\eta\lambda-\frac{\overline{\alpha^{t+1/2}}}{\overline{\alpha^{t}}}\right)\boldsymbol{h}^{\star}+(1-\eta\lambda)\left(\widetilde{\boldsymbol{h}}^{t}-\boldsymbol{h}^{\star}\right)-\frac{\eta}{\left|\alpha^{t}\right|^{2}}\sum_{j=1}^{m}\boldsymbol{b}_{j}\boldsymbol{b}_{j}^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^{t}\widetilde{\boldsymbol{x}}^{t\mathsf{H}}-\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\right)\boldsymbol{a}_{j}\boldsymbol{a}_{j}^{\mathsf{H}}\widetilde{\boldsymbol{x}}^{t}+\frac{\eta}{\left|\alpha^{t}\right|^{2}}\sum_{j=1}^{m}\xi_{j}\boldsymbol{b}_{j}\boldsymbol{a}_{j}^{\mathsf{H}}\widetilde{\boldsymbol{x}}^{t}\\
&=\left(1-\eta\lambda-\frac{\overline{\alpha^{t+1/2}}}{\overline{\alpha^{t}}}\right)\boldsymbol{h}^{\star}+(1-\eta\lambda)\left(\widetilde{\boldsymbol{h}}^{t}-\boldsymbol{h}^{\star}\right)-\frac{\eta}{\left|\alpha^{t}\right|^{2}}\sum_{j=1}^{m}\boldsymbol{b}_{j}\boldsymbol{b}_{j}^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^{t}-\boldsymbol{h}^{\star}\right)\widetilde{\boldsymbol{x}}^{t\mathsf{H}}\boldsymbol{a}_{j}\boldsymbol{a}_{j}^{\mathsf{H}}\widetilde{\boldsymbol{x}}^{t}\\
&\qquad-\frac{\eta}{\left|\alpha^{t}\right|^{2}}\sum_{j=1}^{m}\boldsymbol{b}_{j}\boldsymbol{b}_{j}^{\mathsf{H}}\boldsymbol{h}^{\star}\left(\widetilde{\boldsymbol{x}}^{t}-\boldsymbol{x}^{\star}\right)^{\mathsf{H}}\boldsymbol{a}_{j}\boldsymbol{a}_{j}^{\mathsf{H}}\widetilde{\boldsymbol{x}}^{t}+\frac{\eta}{\left|\alpha^{t}\right|^{2}}\sum_{j=1}^{m}\xi_{j}\boldsymbol{b}_{j}\boldsymbol{a}_{j}^{\mathsf{H}}\widetilde{\boldsymbol{x}}^{t}\\
&=\left(1-\eta\lambda-\frac{\overline{\alpha^{t+1/2}}}{\overline{\alpha^{t}}}\right)\boldsymbol{h}^{\star}+\left(1-\eta\lambda-\frac{\eta}{\left|\alpha^{t}\right|^{2}}\right)\left(\widetilde{\boldsymbol{h}}^{t}-\boldsymbol{h}^{\star}\right)-\frac{\eta}{\left|\alpha^{t}\right|^{2}}\underbrace{\sum_{j=1}^{m}\boldsymbol{b}_{j}\boldsymbol{b}_{j}^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^{t}-\boldsymbol{h}^{\star}\right)\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{\nu}_{1}}\\
&\quad-\frac{\eta}{\left|\alpha^{t}\right|^{2}}\underbrace{\sum_{j=1}^{m}\boldsymbol{b}_{j}\boldsymbol{b}_{j}^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^{t}-\boldsymbol{h}^{\star}\right)\left(\left|\boldsymbol{a}_{j}^{\mathsf{H}}\boldsymbol{x}^{\star}\right|^{2}-\left\|\boldsymbol{x}^{\star}\right\|_{2}^{2}\right)}_{=:\boldsymbol{\nu}_{2}}-\frac{\eta}{\left|\alpha^{t}\right|^{2}}\underbrace{\sum_{j=1}^{m}\boldsymbol{b}_{j}\boldsymbol{b}_{j}^{\mathsf{H}}\boldsymbol{h}^{\star}\left(\widetilde{\boldsymbol{x}}^{t}-\boldsymbol{x}^{\star}\right)^{\mathsf{H}}\boldsymbol{a}_{j}\boldsymbol{a}_{j}^{\mathsf{H}}\widetilde{\boldsymbol{x}}^{t}}_{=:\boldsymbol{\nu}_{3}}+\frac{\eta}{\left|\alpha^{t}\right|^{2}}\underbrace{\sum_{j=1}^{m}\xi_{j}\boldsymbol{b}_{j}\boldsymbol{a}_{j}^{\mathsf{H}}\widetilde{\boldsymbol{x}}^{t}}_{=:\boldsymbol{\nu}_{4}},
\end{aligned}
\tag{B.11}
$$

where (i) comes from the gradient update rule (B.1) and (ii) is due to the expression (A.7).

- Applying a similar argument as for Ma et al. [2018, Equation (219)] yields

$$\left|\boldsymbol{b}_{l}^{\mathsf{H}}\boldsymbol{\nu}_{1}\right|\leq 0.1\max_{1\leq j\leq m}\left|\boldsymbol{b}_{j}^{\mathsf{H}}\big(\widetilde{\boldsymbol{h}}^{t}-\boldsymbol{h}^{\star}\big)\right|.$$

- The $\boldsymbol{\nu}_2$ can be controlled as follows

$$
\begin{aligned}
\left|\boldsymbol{b}_{l}^{\mathsf{H}}\boldsymbol{\nu}_{2}\right| &\leq 0.2\max_{1\leq j\leq m}\left|\boldsymbol{b}_{j}^{\mathsf{H}}\big(\widetilde{\boldsymbol{h}}^{t}-\boldsymbol{h}^{\star}\big)\right|+C\log m\max_{0\leq l\leq m-\tau,1\leq j\leq\tau}\left|\left(\boldsymbol{b}_{l+j}-\boldsymbol{b}_{l+1}\right)^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^{t}-\boldsymbol{h}^{\star}\right)\right|\\
&\leq 0.2\max_{1\leq j\leq m}\left|\boldsymbol{b}_{j}^{\mathsf{H}}\big(\widetilde{\boldsymbol{h}}^{t}-\boldsymbol{h}^{\star}\big)\right|+\left(C\log m\right)C_{11}\frac{\sigma}{\log^{3}m}.
\end{aligned}
$$

  The first inequality can be derived via a similar argument as in Ma et al. [2018, Equation (221)] (the detailed proof is omitted here for the sake of simplicity), whereas the second inequality results from the following claim.

Claim 2. For some constant $C_{11}\gg C_7$, we have

$$\max_{0\leq l\leq m-\tau,1\leq j\leq\tau}\left|\left(\boldsymbol{b}_{l+j}-\boldsymbol{b}_{l+1}\right)^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^{t}-\boldsymbol{h}^{\star}\right)\right|\leq C_{11}\frac{\sigma}{\log^{3}m}.$$

Proof. See Appendix B.3.3. □

- When it comes to the term $\boldsymbol{\nu}_3$, we observe that

$$\left|\boldsymbol{b}_{l}^{\mathsf{H}}\boldsymbol{\nu}_{3}\right|\leq\left|\sum_{j=1}^{m}\boldsymbol{b}_{l}^{\mathsf{H}}\boldsymbol{b}_{j}\boldsymbol{b}_{j}^{\mathsf{H}}\boldsymbol{h}^{\star}\left(\widetilde{\boldsymbol{x}}^{t}-\boldsymbol{x}^{\star}\right)^{\mathsf{H}}\boldsymbol{a}_{j}\boldsymbol{a}_{j}^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^{t}-\boldsymbol{x}^{\star}\right)\right|+\left|\sum_{j=1}^{m}\boldsymbol{b}_{l}^{\mathsf{H}}\boldsymbol{b}_{j}\boldsymbol{b}_{j}^{\mathsf{H}}\boldsymbol{h}^{\star}\left(\widetilde{\boldsymbol{x}}^{t}-\boldsymbol{x}^{\star}\right)^{\mathsf{H}}\boldsymbol{a}_{j}\boldsymbol{a}_{j}^{\mathsf{H}}\boldsymbol{x}^{\star}\right|$$

$$
\begin{aligned}
&\le \sum_{j=1}^{m} \left| \boldsymbol{b}_l^{\mathsf{H}} \boldsymbol{b}_j \right| \max_{1\le j\le m} \left| \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \right| \max_{1\le j\le m} \left| \boldsymbol{a}_j^{\mathsf{H}} \left(\widetilde{\boldsymbol{x}}^t - \boldsymbol{x}^\star \right) \right|^2 + \sum_{j=1}^{m} \left| \boldsymbol{b}_l^{\mathsf{H}} \boldsymbol{b}_j \right| \left| \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \right| \max_{1\le j\le m} \left| \boldsymbol{a}_j^{\mathsf{H}} \left(\widetilde{\boldsymbol{x}}^t - \boldsymbol{x}^\star \right) \right| \max_{1\le j\le m} \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right| \\
&\le (4\log m) \frac{\mu}{\sqrt{m}} \left(\max_{1\le j\le m} \left| \boldsymbol{a}_j^{\mathsf{H}} \left(\widetilde{\boldsymbol{x}}^t - \boldsymbol{x}^\star \right) \right| \right)^2 + (4\log m) \frac{\mu}{\sqrt{m}} \max_{1\le j\le m} \left| \boldsymbol{a}_j^{\mathsf{H}} \left(\widetilde{\boldsymbol{x}}^t - \boldsymbol{x}^\star \right) \right| \max_{1\le j\le m} \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right| \\
&\lesssim C_7 \frac{\mu}{\sqrt{m}} \log^2 m \left(\lambda + \sigma \sqrt{K \log m} \right).
\end{aligned}
$$

Here, the penultimate inequality follows from the incoherence condition (B.9d) and Lemma 2, whereas the last inequality follows from the induction hypothesis (B.9d).

- Finally, we turn to the term $\boldsymbol{\nu}_4$. Clearly, it is of the same form as $\boldsymbol{\nu}_4$ in (A.42); therefore, via the same line of analysis, one can deduce the following bound (similar to (A.45))

$$
\begin{aligned}
\left| \boldsymbol{b}_l^{\mathsf{H}} \boldsymbol{\nu}_4 \right| &\lesssim \left(\sigma \log^{1.5} m \right) \max_{1\le j\le m} \left| \boldsymbol{a}_j^{\mathsf{H}} (\widetilde{\boldsymbol{x}}^t - \boldsymbol{x}^\star) \right| + \sigma \sqrt{\frac{K}{m} \log m} \\
&\lesssim \sigma \log^{1.5} m \left(C_7 \sqrt{\log m} \left(\lambda + \sigma \sqrt{K \log m} \right) \right) + \sigma \sqrt{\frac{K}{m} \log m},
\end{aligned}
$$

where the last inequality invokes (B.9d).

With all the preceding results in place, we can combine them to demonstrate that

$$
\begin{aligned}
&\left| \frac{\overline{\alpha^{t+1/2}}}{\overline{\alpha^t}} \right| \max_{1\le j\le m} \left| \boldsymbol{b}_j^{\mathsf{H}} \left(\widetilde{\boldsymbol{h}}^{t+1} - \boldsymbol{h}^\star \right) \right| \\
&\le \left(1 - \eta\lambda - \frac{\overline{\alpha^{t+1/2}}}{\overline{\alpha^t}} \right) \max_{1\le j\le m} \left| \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \right| + \left(1 - \eta\lambda - \frac{\eta}{|\alpha^t|^2} \right) \max_{1\le j\le m} \left| \boldsymbol{b}_j^{\mathsf{H}} \left(\widetilde{\boldsymbol{h}}^t - \boldsymbol{h}^\star \right) \right| \\
&\quad + \frac{\eta}{|\alpha^t|^2} \left(0.3 \max_{1\le j\le m} \left| \boldsymbol{b}_j^{\mathsf{H}} \left(\widetilde{\boldsymbol{h}}^t - \boldsymbol{h}^\star \right) \right| + \log m \times C_{11} \frac{\sigma}{\log^3 m} \right) \\
&\quad + \frac{\eta}{|\alpha^t|^2} C C_7 \frac{\mu}{\sqrt{m}} \log^2 m \left(\lambda + \sigma \sqrt{K \log m} \right) + \frac{\eta C}{|\alpha^t|^2} \left(\sigma \log^{1.5} m \left(C_7 \sqrt{\log m} \left(\lambda + \sigma \sqrt{K \log m} \right) \right) + \sigma \sqrt{\frac{K}{m} \log m} \right) \\
&\overset{\text{(i)}}{\le} \left(1 - \frac{7\eta}{40} \right) \max_{1\le j\le m} \left| \boldsymbol{b}_j^{\mathsf{H}} \left(\widetilde{\boldsymbol{h}}^t - \boldsymbol{h}^\star \right) \right| + \left(\eta\lambda + \left| 1 - \frac{\alpha^{t+1/2}}{\alpha^t} \right| \right) \frac{\mu}{\sqrt{m}} + \frac{4 C_{11} \eta \sigma}{\log^2 m} + C C_7 \frac{\mu}{\sqrt{m}} \log^2 m \left(\lambda + \sigma \sqrt{K \log m} \right) \\
&\quad + 4\eta C \left[\sigma \log^{1.5} m \left(C_7 \sqrt{\log m} \left(\lambda + \sigma \sqrt{K \log m} \right) \right) + \sigma \sqrt{\frac{K}{m} \log m} \right] \\
&\le \left(1 - \frac{7\eta}{40} \right) C_9 \sigma + c\eta\sigma,
\end{aligned}
$$

for some constant $C > 0$ and sufficiently small constant $c > 0$. Here (i) uses triangle inequality and (B.8l) and the proviso that $m \gg \mu^2 K \log^5 m$ and $\sigma \sqrt{K \log^4 m} \ll 1$.

Finally, making use of (B.8i) we obtain

$$
\begin{aligned}
\max_{1\le j\le m} \left| \boldsymbol{b}_j^{\mathsf{H}} \left(\widetilde{\boldsymbol{h}}^{t+1} - \boldsymbol{h}^\star \right) \right| &\le \frac{\left(1 - \frac{7\eta}{40} \right) C_9 \sigma + c\eta\sigma}{\left| \frac{\alpha^{t+1/2}}{\alpha^t} \right|} \le \frac{\left(1 - \frac{7\eta}{40} \right) C_9 \sigma + c\eta\sigma}{1 - \left| \frac{\alpha^{t+1/2}}{\alpha^t} - 1 \right|} \\
&\le \frac{\left(1 - \frac{7\eta}{40} \right) C_9 \sigma + c\eta\sigma}{1 - \eta \frac{C C_5}{c_\rho} \left(\lambda + \sigma \sqrt{K \log m} \right)} \\
&\le C_9 \sigma,
\end{aligned}
$$

where $C > 0$ is some constant and the last inequality holds since c is sufficiently small.

B.3.2 Proof of the claim (B.2e)

To prove (B.2e), we need to show that the objective value decreases as the algorithm progresses.

Claim 3. If the iterates satisfy the induction hypotheses (B.9a)-(B.9e) in the tth iteration, then with probability exceeding $1 - O\left(m^{-100} + e^{-CK}\log m\right)$,

$$f\left(\boldsymbol{h}^{t+1}, \boldsymbol{x}^{t+1}\right) \leq f\left(\boldsymbol{h}^{t}, \boldsymbol{x}^{t}\right) - \frac{\eta}{2}\left\|\nabla f\left(\boldsymbol{h}^{t}, \boldsymbol{x}^{t}\right)\right\|_2^2. \tag{B.12}$$

Proof. See Appendix B.3.4. □

When summed over t, the inequality in Lemma 3 leads to the following telescopic sum

$$f\left(\boldsymbol{z}^{t_0}\right) \leq f\left(\boldsymbol{z}^{0}\right) - \frac{\eta}{2}\sum_{t=0}^{t_0-1}\left\|\nabla f\left(\boldsymbol{z}^{t}\right)\right\|_2^2.$$

This further gives

$$\min_{0\leq t<t_0}\left\|\nabla f\left(\boldsymbol{z}^{t}\right)\right\|_2 \leq \left\{\frac{1}{t_0}\sum_{t=0}^{t_0-1}\left\|\nabla f\left(\boldsymbol{z}^{t}\right)\right\|_2^2\right\}^{1/2} \leq \left\{\frac{2}{\eta t_0}\left[f\left(\boldsymbol{z}^{\star}\right) - f\left(\boldsymbol{z}^{t_0}\right)\right]\right\}^{1/2}, \tag{B.13}$$

where we have assumed that $\boldsymbol{z}^0 = \boldsymbol{z}^{\star}$.

We then proceed to control $f\left(\boldsymbol{z}^{\star}\right) - f\left(\boldsymbol{z}^{t_0}\right)$. From the mean value theorem (cf. Ma et al. [2018, Appendix D.3.1]), we can write

$$\begin{aligned}
f\left(\boldsymbol{z}^{t_0}\right) &= f\left(\frac{\boldsymbol{h}^{t_0}}{\overline{\alpha^{t_0}}/\left|\alpha^{t_0}\right|}, \frac{\alpha^{t_0}}{\left|\alpha^{t_0}\right|}\boldsymbol{x}^{t_0}\right)\\
&= f\left(\boldsymbol{z}^{\star}\right) + \begin{bmatrix}\nabla f\left(\boldsymbol{z}^{\star}\right)\\ \overline{\nabla f\left(\boldsymbol{z}^{\star}\right)}\end{bmatrix}^{\mathsf{H}}\begin{bmatrix}\overline{\boldsymbol{z}}^{t_0} - \boldsymbol{z}^{\star}\\ \overline{\overline{\boldsymbol{z}}^{t_0} - \boldsymbol{z}^{\star}}\end{bmatrix} + \frac{1}{2}\begin{bmatrix}\overline{\boldsymbol{z}}^{t_0} - \boldsymbol{z}^{\star}\\ \overline{\overline{\boldsymbol{z}}^{t_0} - \boldsymbol{z}^{\star}}\end{bmatrix}^{\mathsf{H}}\nabla^2 f\left(\widehat{\boldsymbol{z}}\right)\begin{bmatrix}\overline{\boldsymbol{z}}^{t_0} - \boldsymbol{z}^{\star}\\ \overline{\overline{\boldsymbol{z}}^{t_0} - \boldsymbol{z}^{\star}}\end{bmatrix}
\end{aligned}$$

for some $\widehat{\boldsymbol{z}}$ lying between $\left(\frac{\boldsymbol{h}^{t_0}}{\overline{\alpha^{t_0}}/\left|\alpha^{t_0}\right|}, \frac{\alpha^{t_0}}{\left|\alpha^{t_0}\right|}\boldsymbol{x}^{t_0}\right)$ and $\boldsymbol{z}^{\star}$. Then one has

$$f\left(\boldsymbol{z}^{\star}\right) - f\left(\boldsymbol{z}^{t_0}\right) \leq 2\left\|\nabla f\left(\boldsymbol{z}^{\star}\right)\right\|_2\left\|\overline{\boldsymbol{z}}^{t_0} - \boldsymbol{z}^{\star}\right\|_2 + 4\left\|\overline{\boldsymbol{z}}^{t_0} - \boldsymbol{z}^{\star}\right\|_2^2.$$

The last inequality in the above formula invokes Lemma 4, whose assumptions are verified in the proof of Claim 3 (see Appendix (B.3.4)). Further, the relations (B.24) and (B.18) in the proof of Claim 3 lead to

$$f\left(\boldsymbol{z}^{\star}\right) - f\left(\boldsymbol{z}^{t_0}\right) \lesssim \left(\lambda + \sigma\sqrt{K\log m}\right)^2. \tag{B.14}$$

It then follows from (B.13) and (B.14) that

$$\min_{0\leq t<t_0}\left\|\nabla f\left(\boldsymbol{z}^{t}\right)\right\|_2 \lesssim \sqrt{\frac{2}{\eta t_0}}\left(\lambda + \sigma\sqrt{K\log m}\right) \leq \frac{\lambda}{m^{10}}.$$

B.3.3 Proof of Claim 2

We aim to prove by induction that there exists some constant $C_{11} > 0$ such that

$$\max_{0\leq l\leq m-\tau,\, 1\leq j\leq\tau}\left|\left(\boldsymbol{b}_{l+j} - \boldsymbol{b}_{l+1}\right)^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^{t} - \boldsymbol{h}^{\star}\right)\right| \leq C_{11}\frac{\sigma}{\log^3 m}. \tag{B.15}$$

Apparently, (B.15) holds when $t = 0$ given that $\boldsymbol{h}^0 = \boldsymbol{h}^{\star}$. In what follows, we shall assume that (B.15) holds true at the tth iteration, and examine this condition for the $(t+1)$th iteration.

Similar to the derivation of (B.11), we have the following decomposition

$$\frac{\overline{\alpha^{t+1/2}}}{\overline{\alpha^{t}}}\left(\frac{1}{\overline{\alpha^{t+1}}}\boldsymbol{h}^{t+1} - \boldsymbol{h}^{\star}\right) = \frac{\overline{\alpha^{t+1/2}}}{\overline{\alpha^{t}}}\left(\frac{1}{\overline{\alpha^{t+1/2}}}\boldsymbol{h}^{t+1/2} - \boldsymbol{h}^{\star}\right)$$

$$
\begin{aligned}
&= \left(1-\eta\lambda-\frac{\overline{\alpha^{t+1/2}}}{\overline{\alpha^{t}}}\right)\boldsymbol{h}^\star + \left(1-\eta\lambda-\eta\left\|\boldsymbol{x}^t\right\|_2^2\right)\left(\widetilde{\boldsymbol{h}}^t-\boldsymbol{h}^\star\right) \\
&\quad - \frac{\eta}{\left|\alpha^t\right|^2}\underbrace{\sum_{j=1}^m \boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}^t}-\boldsymbol{h}^\star\right)\widetilde{\boldsymbol{x}}^{t\mathsf{H}}\left(\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}-\boldsymbol{I}_k\right)\widetilde{\boldsymbol{x}}^t}_{=:\boldsymbol{\nu}_1} \\
&\quad - \frac{\eta}{\left|\alpha^t\right|^2}\underbrace{\sum_{j=1}^m \boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\left(\widetilde{\boldsymbol{x}}^t-\boldsymbol{x}^\star\right)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\widetilde{\boldsymbol{x}}^t}_{=:\boldsymbol{\nu}_2} + \frac{\eta}{\left|\alpha^t\right|^2}\underbrace{\sum_{j=1}^m \xi_j\boldsymbol{b}_j\boldsymbol{a}_j^{\mathsf{H}}\widetilde{\boldsymbol{x}}^t}_{=:\boldsymbol{\nu}_3},
\end{aligned}
$$

leaving us with several terms to control.

- For $\boldsymbol{\nu}_1$, we have that

$$
\begin{aligned}
\left|\left(\boldsymbol{b}_l-\boldsymbol{b}_1\right)^{\mathsf{H}}\boldsymbol{\nu}_1\right| &\le \sum_{j=1}^m\left|\left(\boldsymbol{b}_l-\boldsymbol{b}_1\right)^{\mathsf{H}}\boldsymbol{b}_j\right|\max_{1\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}^t}-\boldsymbol{h}^\star\right)\widetilde{\boldsymbol{x}}^{t\mathsf{H}}\left(\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}-\boldsymbol{I}_k\right)\widetilde{\boldsymbol{x}}^t\right| \\
&\le \frac{c}{\log^2 m}\max_{1\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^t-\boldsymbol{h}^\star\right)\right|\max_{1\le j\le m}\left|\widetilde{\boldsymbol{x}}^{t\mathsf{H}}\left(\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}-\boldsymbol{I}_k\right)\widetilde{\boldsymbol{x}}^t\right| \\
&\le \frac{c}{\log^2 m}\max_{1\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^t-\boldsymbol{h}^\star\right)\right|\max_{1\le j\le m}\left(\left\|\boldsymbol{a}_j^{\mathsf{H}}\widetilde{\boldsymbol{x}}^t\right\|_2^2+\left\|\widetilde{\boldsymbol{x}}^t\right\|_2^2\right) \\
&\lesssim \frac{c}{\log m}\max_{1\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^t-\boldsymbol{h}^\star\right)\right|,
\end{aligned}
$$

  where the second inequality follows from Ma et al. [2018, Lemma 50] and the last inequality utilizes the following consequence of (B.9d) and Lemma 38:

$$
\max_{1\le j\le m}\left(\left\|\boldsymbol{a}_j^{\mathsf{H}}\widetilde{\boldsymbol{x}}^t\right\|_2^2+\left\|\widetilde{\boldsymbol{x}}^t\right\|_2^2\right)\lesssim\max_{1\le j\le m}\left(2\left\|\boldsymbol{a}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^t-\boldsymbol{x}^\star\right)\right\|_2^2+2\left\|\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}^\star\right\|_2^2+\left\|\widetilde{\boldsymbol{x}}^t\right\|_2^2\right)\lesssim\log m.
$$

- With regards to $\boldsymbol{\nu}_2$, we invoke the induction hypothesis (B.9d) at the tth iteration to obtain

$$
\begin{aligned}
\left|\left(\boldsymbol{b}_l-\boldsymbol{b}_1\right)^{\mathsf{H}}\boldsymbol{\nu}_2\right| &\le \sum_{j=1}^m\left|\left(\boldsymbol{b}_l-\boldsymbol{b}_1\right)^{\mathsf{H}}\boldsymbol{b}_j\right|\max_{1\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\right|\max_{1\le j\le m}\left|\left(\widetilde{\boldsymbol{x}}^t-\boldsymbol{x}^\star\right)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\widetilde{\boldsymbol{x}}^t\right| \\
&\le \frac{c}{\log^2 m}\frac{\mu}{\sqrt{m}}\left(\max_{1\le j\le m}\left|\left(\widetilde{\boldsymbol{x}}^t-\boldsymbol{x}^\star\right)^{\mathsf{H}}\boldsymbol{a}_j\right|^2+\max_{1\le j\le m}\left|\left(\widetilde{\boldsymbol{x}}^t-\boldsymbol{x}^\star\right)^{\mathsf{H}}\boldsymbol{a}_j\right|\max_{1\le j\le m}\left|\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}^\star\right|\right) \\
&\lesssim C_8\frac{\mu}{\log m\sqrt{m}}\left(\lambda+\sigma\sqrt{\log m}\right),
\end{aligned}
$$

  where the second inequality applies Ma et al. [2018, Lemma 50] and (7), and the last inequality results from (B.9d) and (F.1).

- Finally, since $\left(\boldsymbol{b}_l-\boldsymbol{b}_1\right)^{\mathsf{H}}\boldsymbol{\nu}_3$ is of the same form as the quantity β_3 in (A.47), we can apply the analysis leading to (A.49) to derive

$$
\begin{aligned}
\left|\left(\boldsymbol{b}_l-\boldsymbol{b}_1\right)^{\mathsf{H}}\boldsymbol{\nu}_3\right| &\lesssim \frac{\sigma}{\log^{1.5} m}\max_{1\le j\le m}\left|\boldsymbol{a}_j^{\mathsf{H}}(\widetilde{\boldsymbol{x}}^t-\boldsymbol{x}^\star)\right|+\sigma\sqrt{\frac{K\log^2 m}{m}} \\
&\lesssim \frac{\sigma}{\log^{1.5} m}\left(C_7\sqrt{\log m}\left(\lambda+\sigma\sqrt{K\log m}\right)\right)+\sigma\sqrt{\frac{K\log^2 m}{m}}
\end{aligned}
$$

With the preceding results in hand, we have

$$
\left|\frac{\overline{\alpha^{t+1/2}}}{\overline{\alpha^t}}\right|\max_{0\le l\le m-\tau,1\le j\le\tau}\left|\left(\boldsymbol{b}_{l+j}-\boldsymbol{b}_{l+1}\right)^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^{t+1}-\boldsymbol{h}^\star\right)\right|
$$

$$
\begin{aligned}
&\le \left|1-\eta\lambda-\frac{\overline{\alpha^{t+1/2}}}{\overline{\alpha^{t}}}\right| \max_{0\le l\le m-\tau, 1\le j\le\tau}\left|\left(\boldsymbol{b}_{l+j}-\boldsymbol{b}_{l+1}\right)^{\mathsf{H}}\boldsymbol{h}^{\star}\right| \\
&\quad+\left(1-\eta\lambda-\eta\left\|\boldsymbol{x}^{t}\right\|_2^2\right)\max_{0\le l\le m-\tau, 1\le j\le\tau}\left|\left(\boldsymbol{b}_{l+j}-\boldsymbol{b}_{l+1}\right)^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^{t}-\boldsymbol{h}^{\star}\right)\right| \\
&\quad+\frac{\eta C C_9}{|\alpha^t|^2}\frac{\mu\log m}{\sqrt{m}}\left(\lambda+\sigma\sqrt{K\log m}\right)+\frac{\eta C C_8}{|\alpha^t|^2}\left(\frac{\mu}{\log m\sqrt{m}}\left(\lambda+\sigma\sqrt{K\log m}\right)\right) \\
&\quad+\frac{\eta C}{|\alpha^t|^2}\left[\frac{\sigma}{\log^{1.5} m}\left(C_7\sqrt{\log m}\left(\lambda+\sigma\sqrt{K\log m}\right)\right)+\sigma\sqrt{\frac{K\log^2 m}{m}}\right] \\
&\overset{\text{(i)}}{\le}\left(\eta\lambda+\left|1-\frac{\alpha^{t+1/2}}{\alpha^{t}}\right|\right)\frac{2\mu}{\sqrt{m}}+\left(1-\frac{\eta}{16}\right)C_{11}\frac{\sigma}{\log^3 m} \\
&\quad+4\eta C C_9\frac{\mu\log m}{\sqrt{m}}\left(\lambda+\sigma\sqrt{K\log m}\right)+4\eta C C_8\left(\frac{\mu}{\log m\sqrt{m}}\left(\lambda+\sigma\sqrt{K\log m}\right)\right) \\
&\quad+4\eta C\left[\frac{\sigma}{\log^{1.5} m}\left(C_7\sqrt{\log m}\left(\lambda+\sigma\sqrt{K\log m}\right)\right)+\sigma\sqrt{\frac{K\log^2 m}{m}}\right] \\
&\overset{\text{(ii)}}{\le}\left(1-\frac{\eta}{16}\right)\frac{C_{11}\sigma}{\log^3 m}+c\frac{\eta\sigma}{\log^3 m}
\end{aligned}
$$

for some constant $C > 0$ and some sufficiently small constant $c > 0$. Here, the relation (i) comes from the triangle inequality, (B.8l), as well as the consequence of (B.8c) and (B.8l)

$$
\left\|\boldsymbol{x}^{t}\right\|_2=\frac{\|\widetilde{\boldsymbol{x}}^{t}\|_2}{|\alpha^t|}\ge\frac{1/2}{2}=\frac{1}{4};
$$

the inequality (ii) invokes (B.8i) and holds with the proviso that $m \gg \mu^2 K\log^8 m$ and $\sigma\sqrt{K\log^5 m}\ll 1$.

Finally, by (B.8i) we obtain

$$
\begin{aligned}
\max_{0\le l\le m-\tau, 1\le j\le\tau}\left|\left(\boldsymbol{b}_{l+j}-\boldsymbol{b}_{l+1}\right)^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^{t+1}-\boldsymbol{h}^{\star}\right)\right| &\le \frac{\left(1-\frac{\eta}{16}\right)C_{11}\frac{\sigma}{\log^3 m}+c\frac{\eta\sigma}{\log^3 m}}{\left|\frac{\overline{\alpha^{t+1/2}}}{\overline{\alpha^{t}}}\right|} \\
&\le \frac{\left(1-\frac{\eta}{16}\right)C_{11}\frac{\sigma}{\log^3 m}+c\frac{\eta\sigma}{\log^3 m}}{1-\left|\frac{\alpha^{t+1/2}}{\alpha^{t}}-1\right|} \\
&\le \frac{\left(1-\frac{\eta}{16}\right)C_{11}\frac{\sigma}{\log^3 m}+c\frac{\eta\sigma}{\log^3 m}}{1-\eta\frac{CC_5}{c_\rho}\left(\lambda+\sigma\sqrt{K\log m}\right)} \\
&\le C_{11}\frac{\sigma}{\log^3 m},
\end{aligned}
$$

where $C > 0$ is some constant. Here, the last inequality holds as long as c is sufficiently small.

B.3.4 Proof of Claim 3

Before proceeding, we note that

$$
\nabla f(\boldsymbol{z})=\nabla f_{\textsf{reg-free}}(\boldsymbol{z})+\lambda\boldsymbol{z},
$$

and

$$
\begin{bmatrix}\nabla_{\boldsymbol{h}} f\left(\frac{\boldsymbol{h}}{\overline{\alpha}},\alpha\boldsymbol{x}\right)\\ \nabla_{\boldsymbol{x}} f\left(\frac{\boldsymbol{h}}{\overline{\alpha}},\alpha\boldsymbol{x}\right)\end{bmatrix}=\begin{bmatrix}\alpha\nabla_{\boldsymbol{h}} f_{\textsf{reg-free}}(\boldsymbol{h},\boldsymbol{x})\\ \frac{1}{\overline{\alpha}}\nabla_{\boldsymbol{x}} f_{\textsf{reg-free}}(\boldsymbol{h},\boldsymbol{x})\end{bmatrix}+\lambda\begin{bmatrix}\frac{\boldsymbol{h}}{\overline{\alpha}}\\ \alpha\boldsymbol{x}\end{bmatrix}. \tag{B.16}
$$

Another fact of use is that

$$
\nabla^2 f(\boldsymbol{h},\boldsymbol{x})=\nabla^2 f_{\textsf{reg-free}}(\boldsymbol{h},\boldsymbol{x})+\lambda\boldsymbol{I}_{4K}.
$$

Letting

$$\beta^t = \frac{\alpha^t}{|\alpha^t|}, \qquad \overline{\boldsymbol{h}}^t = \frac{1}{\overline{\beta^t}}\boldsymbol{h}^t, \qquad \text{and} \qquad \overline{\boldsymbol{x}}^t = \beta^t \boldsymbol{x}^t,$$

we can write

$$\begin{aligned}
\left\|\nabla f(\overline{\boldsymbol{h}}^t, \overline{\boldsymbol{x}}^t)\right\|_2 &= \left\| \begin{bmatrix} \beta^t \nabla_{\boldsymbol{h}} f_{\text{reg-free}}(\boldsymbol{h}^t, \boldsymbol{x}^t) \\ \frac{1}{\overline{\beta^t}} \nabla_{\boldsymbol{x}} f_{\text{reg-free}}(\boldsymbol{h}^t, \boldsymbol{x}^t) \end{bmatrix} + \lambda \begin{bmatrix} \frac{\boldsymbol{h}^t}{\overline{\beta^t}} \\ \beta^t \boldsymbol{x}^t \end{bmatrix} \right\|_2 \\
&= \left\| \begin{bmatrix} \nabla_{\boldsymbol{h}} f_{\text{reg-free}}(\boldsymbol{h}^t, \boldsymbol{x}^t) \\ \nabla_{\boldsymbol{x}} f_{\text{reg-free}}(\boldsymbol{h}^t, \boldsymbol{x}^t) \end{bmatrix} + \lambda \begin{bmatrix} \boldsymbol{h}^t \\ \boldsymbol{x}^t \end{bmatrix} \right\|_2 \\
&= \left\|\nabla f(\boldsymbol{h}^t, \boldsymbol{x}^t)\right\|_2, \qquad \text{(B.17)}
\end{aligned}$$

where the first inequality is due to (B.16), and the second inequality comes from the simple fact that $\beta^t \overline{\beta^t} = 1$ (by definition of β^t).

To begin with, we show that $f(\boldsymbol{h}^{t+1}, \boldsymbol{x}^{t+1})$ is upper bounded by $f(\boldsymbol{h}^{t+1/2}, \boldsymbol{x}^{t+1/2})$, that is,

$$\begin{aligned}
f(\boldsymbol{h}^{t+1}, \boldsymbol{x}^{t+1}) &= \sum_{j=1}^m \left| \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^{t+1} (\boldsymbol{x}^{t+1})^{\mathsf{H}} \boldsymbol{a}_j - y_j \right|^2 + \lambda \left\|\boldsymbol{h}^{t+1}\right\|_2^2 + \lambda \left\|\boldsymbol{x}^{t+1}\right\|_2^2 \\
&\overset{\text{(i)}}{=} \sum_{j=1}^m \left| \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^{t+1/2} (\boldsymbol{x}^{t+1/2})^{\mathsf{H}} \boldsymbol{a}_j - y_j \right|^2 + 2\lambda \left\|\boldsymbol{h}^{t+1}\right\|_2 \left\|\boldsymbol{x}^{t+1}\right\|_2 \\
&\overset{\text{(ii)}}{=} \sum_{j=1}^m \left| \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^{t+1/2} (\boldsymbol{x}^{t+1/2})^{\mathsf{H}} \boldsymbol{a}_j - y_j \right|^2 + 2\lambda \left\|\boldsymbol{h}^{t+1/2}\right\|_2 \left\|\boldsymbol{x}^{t+1/2}\right\|_2 \\
&\overset{\text{(iii)}}{\leq} \sum_{j=1}^m \left| \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^{t+1/2} (\boldsymbol{x}^{t+1/2})^{\mathsf{H}} \boldsymbol{a}_j - y_j \right|^2 + \lambda \left\|\boldsymbol{h}^{t+1/2}\right\|_2^2 + \lambda \left\|\boldsymbol{x}^{t+1/2}\right\|_2^2 \\
&= f\left(\boldsymbol{h}^{t+1/2}, \boldsymbol{x}^{t+1/2}\right),
\end{aligned}$$

where (i) and (ii) come from (B.8f), and (iii) is due to the elementary inequality $2ab \leq a^2 + b^2$. In order to control $f(\boldsymbol{h}^{t+1/2}, \boldsymbol{x}^{t+1/2})$, one observes that

$$\begin{aligned}
f\left(\boldsymbol{h}^{t+1/2}, \boldsymbol{x}^{t+1/2}\right) &= f\left(\frac{\boldsymbol{h}^{t+1/2}}{\overline{\beta^t}}, \beta^t \boldsymbol{x}^{t+1/2}\right) \\
&\overset{\text{(i)}}{=} f\left(\overline{\boldsymbol{h}}^t - \frac{\eta}{\overline{\beta^t}} \left(\nabla_{\boldsymbol{h}} f_{\text{reg-free}}(\boldsymbol{z}^t) + \lambda \boldsymbol{h}^t\right), \overline{\boldsymbol{x}}^t - \eta \beta^t \left(\nabla_{\boldsymbol{x}} f_{\text{reg-free}}(\boldsymbol{z}^t) + \lambda \boldsymbol{x}^t\right)\right) \\
&\overset{\text{(ii)}}{=} f\left(\overline{\boldsymbol{h}}^t - \eta \nabla_{\boldsymbol{h}} f(\overline{\boldsymbol{z}}^t), \overline{\boldsymbol{x}}^t - \eta \nabla_{\boldsymbol{x}} f(\overline{\boldsymbol{z}}^t)\right) \\
&\overset{\text{(iii)}}{=} f\left(\overline{\boldsymbol{h}}^t, \overline{\boldsymbol{x}}^t\right) - \eta \begin{bmatrix} \nabla_{\boldsymbol{h}} f\left(\overline{\boldsymbol{h}}^t, \overline{\boldsymbol{x}}^t\right) \\ \nabla_{\boldsymbol{x}} f\left(\overline{\boldsymbol{h}}^t, \overline{\boldsymbol{x}}^t\right) \\ \hline \overline{\nabla_{\boldsymbol{h}} f\left(\overline{\boldsymbol{h}}^t, \overline{\boldsymbol{x}}^t\right)} \\ \hline \overline{\nabla_{\boldsymbol{x}} f\left(\overline{\boldsymbol{h}}^t, \overline{\boldsymbol{x}}^t\right)} \end{bmatrix}^{\mathsf{H}} \begin{bmatrix} \nabla_{\boldsymbol{h}} f\left(\overline{\boldsymbol{h}}^t, \overline{\boldsymbol{x}}^t\right) \\ \nabla_{\boldsymbol{x}} f\left(\overline{\boldsymbol{h}}^t, \overline{\boldsymbol{x}}^t\right) \\ \hline \overline{\nabla_{\boldsymbol{h}} f\left(\overline{\boldsymbol{h}}^t, \overline{\boldsymbol{x}}^t\right)} \\ \hline \overline{\nabla_{\boldsymbol{x}} f\left(\overline{\boldsymbol{h}}^t, \overline{\boldsymbol{x}}^t\right)} \end{bmatrix} \\
&\quad + \frac{\eta^2}{2} \begin{bmatrix} \nabla_{\boldsymbol{h}} f\left(\overline{\boldsymbol{h}}^t, \overline{\boldsymbol{x}}^t\right) \\ \nabla_{\boldsymbol{x}} f\left(\overline{\boldsymbol{h}}^t, \overline{\boldsymbol{x}}^t\right) \\ \hline \overline{\nabla_{\boldsymbol{h}} f\left(\overline{\boldsymbol{h}}^t, \overline{\boldsymbol{x}}^t\right)} \\ \hline \overline{\nabla_{\boldsymbol{x}} f\left(\overline{\boldsymbol{h}}^t, \overline{\boldsymbol{x}}^t\right)} \end{bmatrix}^{\mathsf{H}} \nabla^2 f(\widehat{\boldsymbol{z}}) \begin{bmatrix} \nabla_{\boldsymbol{h}} f\left(\overline{\boldsymbol{h}}^t, \overline{\boldsymbol{x}}^t\right) \\ \nabla_{\boldsymbol{x}} f\left(\overline{\boldsymbol{h}}^t, \overline{\boldsymbol{x}}^t\right) \\ \hline \overline{\nabla_{\boldsymbol{h}} f\left(\overline{\boldsymbol{h}}^t, \overline{\boldsymbol{x}}^t\right)} \\ \hline \overline{\nabla_{\boldsymbol{x}} f\left(\overline{\boldsymbol{h}}^t, \overline{\boldsymbol{x}}^t\right)} \end{bmatrix} \\
&\overset{\text{(iv)}}{\leq} f\left(\overline{\boldsymbol{h}}^t, \overline{\boldsymbol{x}}^t\right) - 2\eta \left\|\nabla_{\boldsymbol{h}} f\left(\overline{\boldsymbol{h}}^t, \overline{\boldsymbol{x}}^t\right)\right\|_2^2 - 2\eta \left\|\nabla_{\boldsymbol{x}} f\left(\overline{\boldsymbol{h}}^t, \overline{\boldsymbol{x}}^t\right)\right\|_2^2
\end{aligned}$$

$$
\begin{aligned}
&+\frac{\eta^2}{2}\cdot 4\left[2\left\|\nabla_{\boldsymbol{h}} f\left(\overline{\boldsymbol{h}}^t,\overline{\boldsymbol{x}}^t\right)\right\|_2^2+2\left\|\nabla_{\boldsymbol{x}} f\left(\overline{\boldsymbol{h}}^t,\overline{\boldsymbol{x}}^t\right)\right\|_2^2\right]\\
&\overset{\text{(v)}}{\leq} f\left(\overline{\boldsymbol{h}}^t,\overline{\boldsymbol{x}}^t\right)-\frac{\eta}{2}\left\|\nabla_{\boldsymbol{h}} f\left(\overline{\boldsymbol{h}}^t,\overline{\boldsymbol{x}}^t\right)\right\|_2^2-\frac{\eta}{2}\left\|\nabla_{\boldsymbol{x}} f\left(\overline{\boldsymbol{h}}^t,\overline{\boldsymbol{x}}^t\right)\right\|_2^2\\
&= f\left(\boldsymbol{h}^t,\boldsymbol{x}^t\right)-\frac{\eta}{2}\left\|\nabla f\left(\boldsymbol{h}^t,\boldsymbol{x}^t\right)\right\|_2^2,
\end{aligned}
$$

where $\widehat{\boldsymbol{z}}$ is a point lying between $\overline{\boldsymbol{z}}^t-\eta\nabla f\left(\overline{\boldsymbol{z}}^t\right)$ and $\overline{\boldsymbol{z}}^t$. Here, (i) resorts to the gradient update rule (B.1); (ii) utilizes the relation (B.16); (iii) comes from the mean value theorem Ma et al. [2018, Appendix D.3.1]; (iv) follows from Lemma 4 (which we shall verify shortly); (v) holds true for sufficiently small $\eta>0$; and the last equality follows from the identity (B.17). Therefore, it only remains to verify the conditions required to invoke Lemma 4 in Step (iv). In particular, we would need to justify that both $\overline{\boldsymbol{z}}^t$ and $\overline{\boldsymbol{z}}^t-\eta\nabla f\left(\overline{\boldsymbol{z}}^t\right)$ satisfy the conditions of Lemma 4.

- We first show that $\overline{\boldsymbol{z}}^t$ satisfies the conditions of Lemma 4. Towards this, it is first seen that

$$
\begin{aligned}
\left\|\overline{\boldsymbol{h}}^t-\boldsymbol{h}^\star\right\|_2^2+\left\|\overline{\boldsymbol{x}}^t-\boldsymbol{x}^\star\right\|_2^2&=\left\|\frac{\boldsymbol{h}^t}{\overline{\alpha^t/|\alpha^t|}}-\boldsymbol{h}^\star\right\|_2^2+\left\|\frac{\alpha^t}{|\alpha^t|}\boldsymbol{x}^t-\boldsymbol{x}^\star\right\|_2^2\\
&\leq\left(\left\|\frac{\boldsymbol{h}^t}{\overline{\alpha^t/|\alpha^t|}}-\frac{\boldsymbol{h}^t}{\overline{\alpha^t}}\right\|_2+\left\|\frac{\boldsymbol{h}^t}{\overline{\alpha^t}}-\boldsymbol{h}^\star\right\|_2\right)^2+\left(\left\|\frac{\alpha^t}{|\alpha^t|}\boldsymbol{x}^t-\alpha^t\boldsymbol{x}^t\right\|_2+\left\|\alpha^t\boldsymbol{x}^t-\boldsymbol{x}^\star\right\|_2\right)^2\\
&=\left(\left||\alpha^t|-1\right|\left\|\widetilde{\boldsymbol{h}}^t\right\|_2+\left\|\widetilde{\boldsymbol{h}}^t-\boldsymbol{h}^\star\right\|_2\right)^2+\left(\left|\frac{|\alpha^t|-1}{|\alpha^t|}\right|\left\|\widetilde{\boldsymbol{x}}^t\right\|_2+\left\|\widetilde{\boldsymbol{x}}^t-\boldsymbol{x}^\star\right\|_2\right)^2\\
&\lesssim\left(\frac{C_5}{c_\rho}\left(\lambda+\sigma\sqrt{K\log m}\right)\right)^2,
\end{aligned}
\tag{B.18}
$$

where the last inequality comes from (B.8a) and (B.8h). Further,

$$
\begin{aligned}
\max_{1\leq j\leq m}\left|\boldsymbol{a}_j^{\mathsf{H}}\left(\overline{\boldsymbol{x}}^t-\boldsymbol{x}^\star\right)\right|&\leq\max_{1\leq j\leq m}\left|\boldsymbol{a}_j^{\mathsf{H}}\left(\frac{\alpha^t}{|\alpha^t|}\boldsymbol{x}^t-\alpha^t\boldsymbol{x}^t\right)\right|+\max_{1\leq j\leq m}\left|\boldsymbol{a}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^t-\boldsymbol{x}^\star\right)\right|\\
&\leq\left|\frac{|\alpha^t|-1}{|\alpha^t|}\right|\max_{1\leq j\leq m}\left|\boldsymbol{a}_j^{\mathsf{H}}\widetilde{\boldsymbol{x}}^t\right|+\max_{1\leq j\leq m}\left|\boldsymbol{a}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^t-\boldsymbol{x}^\star\right)\right|\\
&\leq\left|\frac{|\alpha^t|-1}{|\alpha^t|}\right|\left(\max_{1\leq j\leq m}\left|\boldsymbol{a}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^t-\boldsymbol{x}^\star\right)\right|+\max_{1\leq j\leq m}\left|\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}^\star\right|\right)+\max_{1\leq j\leq m}\left|\boldsymbol{a}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^t-\boldsymbol{x}^\star\right)\right|\\
&\lesssim\left(\lambda+\sigma\sqrt{K\log m}\right)\sqrt{\log m},
\end{aligned}
\tag{B.19}
$$

where the last inequality follows from (B.8h), (B.9d) and Lemma 38. Similarly, one has

$$
\begin{aligned}
\max_{1\leq j\leq m}\left|\boldsymbol{b}_j^{\mathsf{H}}\overline{\boldsymbol{h}}^t\right|&\leq\max_{1\leq j\leq m}\left|\boldsymbol{b}_j^{\mathsf{H}}\left(\frac{\boldsymbol{h}^t}{\overline{\alpha^t/|\alpha^t|}}-\frac{\boldsymbol{h}^t}{\overline{\alpha^t}}\right)\right|+\max_{1\leq j\leq m}\left|\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right|\\
&\leq\left||\alpha^t|-1\right|\max_{1\leq j\leq m}\left|\boldsymbol{b}_j^{\mathsf{H}}\frac{\boldsymbol{h}^t}{\overline{\alpha^t}}\right|+\max_{1\leq j\leq m}\left|\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right|\\
&\leq 2\max_{1\leq j\leq m}\left|\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right| &\text{(B.20)}\\
&\lesssim\frac{\mu}{\sqrt{m}}\log m+\sigma, &\text{(B.21)}
\end{aligned}
$$

where the last inequality comes from (B.9e). Given that $\overline{\boldsymbol{z}}^t$ satisfies the conditions in Lemma 4, we can invoke Lemma 4 to demonstrate that

$$
\left\|\nabla_{\boldsymbol{h}} f\left(\overline{\boldsymbol{z}}^t\right)-\nabla_{\boldsymbol{h}} f\left(\boldsymbol{z}^\star\right)\right\|_2\leq 4\left\|\overline{\boldsymbol{z}}^t-\boldsymbol{z}^\star\right\|_2. \tag{B.22}
$$

- Next, we move on to show that $\overline{\boldsymbol{z}}^t - \eta \nabla f\left(\overline{\boldsymbol{z}}^t\right)$ also satisfies the conditions of Lemma 4. To begin with,

$$\left\|\overline{\boldsymbol{z}}^t - \eta \nabla f\left(\overline{\boldsymbol{z}}^t\right) - \boldsymbol{z}^\star\right\|_2 \leq \left\|\overline{\boldsymbol{z}}^t - \boldsymbol{z}^\star\right\|_2 + \eta \left\|\nabla f\left(\overline{\boldsymbol{z}}^t\right) - \nabla f\left(\boldsymbol{z}^\star\right)\right\|_2 + \eta \left\|\nabla f\left(\boldsymbol{z}^\star\right)\right\|_2. \tag{B.23}$$

We observe that

$$\begin{aligned}\left\|\nabla f\left(\boldsymbol{z}^\star\right)\right\|_2 &\leq \left\|\nabla f_{\mathsf{clean}}\left(\boldsymbol{z}^\star\right)\right\|_2 + \left\|\mathcal{A}^*\left(\boldsymbol{\xi}\right)\boldsymbol{h}^\star\right\|_2 + \left\|\mathcal{A}^*\left(\boldsymbol{\xi}\right)\boldsymbol{x}^\star\right\|_2 + \lambda\left\|\boldsymbol{h}^\star\right\|_2 + \lambda\left\|\boldsymbol{z}^\star\right\|_2 \\ &\lesssim \lambda + \sigma\sqrt{K\log m}. \end{aligned} \tag{B.24}$$

Taking (B.24), (B.22), (B.18) and (B.23) collectively, one arrives at

$$\left\|\overline{\boldsymbol{z}}^t - \eta \nabla f\left(\overline{\boldsymbol{z}}^t\right) - \boldsymbol{z}^\star\right\|_2 \lesssim \lambda + \sigma\sqrt{K\log m}.$$

With regards to the incoherence condition w.r.t. $\boldsymbol{a}_j$, we have

$$\begin{aligned}&\max_{1\leq j\leq m}\left|\boldsymbol{a}_j^{\mathsf{H}}\left(\overline{\boldsymbol{x}}^t - \eta\nabla_{\boldsymbol{x}} f\left(\overline{\boldsymbol{z}}^t\right) - \boldsymbol{x}^\star\right)\right| \\ &\leq \max_{1\leq j\leq m}\left|\boldsymbol{a}_j^{\mathsf{H}}\left(\overline{\boldsymbol{x}}^t - \boldsymbol{x}^\star\right)\right| + \eta\max_{1\leq j\leq m}\left|\boldsymbol{a}_j^{\mathsf{H}}\nabla_{\boldsymbol{x}} f\left(\overline{\boldsymbol{z}}^t\right)\right| \\ &\leq \max_{1\leq j\leq m}\left|\boldsymbol{a}_j^{\mathsf{H}}\left(\overline{\boldsymbol{x}}^t - \boldsymbol{x}^\star\right)\right| + \eta\left(\max_{1\leq j\leq m}\left|\boldsymbol{a}_j^{\mathsf{H}}\nabla_{\boldsymbol{x}} f\left(\overline{\boldsymbol{z}}^t - \widetilde{\boldsymbol{z}}^{t,(l)}\right)\right| + \max_{1\leq j\leq m}\left|\boldsymbol{a}_j^{\mathsf{H}}\nabla_{\boldsymbol{x}} f\left(\widetilde{\boldsymbol{z}}^{t,(l)}\right)\right|\right) \\ &\leq C\sqrt{\log m}\left(\lambda + \sigma\sqrt{K\log m}\right) + 4\eta\left(10\sqrt{K}\times 4\max_{1\leq j\leq m}\left\|\widetilde{\boldsymbol{z}}^t - \widetilde{\boldsymbol{z}}^{t,(l)}\right\|_2 + 20\sqrt{\log m}\max_{1\leq j\leq m}\left\|\nabla_{\boldsymbol{x}} f\left(\widetilde{\boldsymbol{z}}^{t,(l)}\right)\right\|_2\right), \end{aligned} \tag{B.25}$$

where the last inequality follows from (B.19) for some constant $C > 0$, (B.22) and Lemma 38. Further, it is self-evident that $\widetilde{\boldsymbol{z}}^{t,(l)}$ satisfies the conditions of Lemma 4, so that we have

$$\begin{aligned}\left\|\nabla_{\boldsymbol{x}} f\left(\widetilde{\boldsymbol{z}}^{t,(l)}\right)\right\|_2 &\leq \left\|\nabla_{\boldsymbol{x}} f\left(\widetilde{\boldsymbol{z}}^{t,(l)}\right) - \nabla_{\boldsymbol{x}} f\left(\boldsymbol{z}^\star\right)\right\|_2 + \left\|\nabla_{\boldsymbol{x}} f\left(\boldsymbol{z}^\star\right)\right\|_2 \\ &\leq 4\left\|\widetilde{\boldsymbol{z}}^{t,(l)} - \boldsymbol{z}^\star\right\|_2 + C\left(\lambda + \sigma\sqrt{K\log m}\right) \\ &\leq 4\left(\left\|\widetilde{\boldsymbol{z}}^{t,(l)} - \widetilde{\boldsymbol{z}}^t\right\|_2 + \left\|\widetilde{\boldsymbol{z}}^t - \boldsymbol{z}^\star\right\|_2\right) + C\left(\lambda + \sigma\sqrt{K\log m}\right), \end{aligned}$$

where the second inequality invokes Lemma 4 and (B.24). This together with (B.25) and (B.9) gives

$$\max_{1\leq j\leq m}\left|\boldsymbol{a}_j^{\mathsf{H}}\left(\overline{\boldsymbol{x}}^t - \eta\nabla_{\boldsymbol{x}} f\left(\overline{\boldsymbol{z}}^t\right) - \boldsymbol{x}^\star\right)\right| \lesssim \sqrt{\log m}\left(\lambda + \sigma\sqrt{K\log m}\right).$$

For the other incoherence condition w.r.t. $\boldsymbol{b}_j$, we can invoke similar argument to show that

$$\begin{aligned}&\max_{1\leq j\leq m}\left|\boldsymbol{b}_j^{\mathsf{H}}\left(\overline{\boldsymbol{h}}^t - \eta\nabla_{\boldsymbol{h}} f\left(\overline{\boldsymbol{z}}^t\right) - \boldsymbol{h}^\star\right)\right| \\ &\quad\leq \max_{1\leq j\leq m}\left|\boldsymbol{b}_j^{\mathsf{H}}\left(\overline{\boldsymbol{h}}^t - \boldsymbol{h}^\star\right)\right| + \eta\max_{1\leq j\leq m}\left|\boldsymbol{b}_j^{\mathsf{H}}\nabla_{\boldsymbol{h}} f\left(\overline{\boldsymbol{z}}^t\right)\right| \\ &\quad\leq \max_{1\leq j\leq m}\left|\boldsymbol{b}_j^{\mathsf{H}}\left(\overline{\boldsymbol{h}}^t - \boldsymbol{h}^\star\right)\right| + \eta\max_{1\leq j\leq m}\left|\boldsymbol{b}_j^{\mathsf{H}}\left(\sum_{l=1}^m\left(\boldsymbol{b}_l^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\widetilde{\boldsymbol{x}}^{t,\mathsf{H}}\boldsymbol{a}_l - y_l\right)\boldsymbol{b}_l\boldsymbol{a}_l^{\mathsf{H}}\overline{\boldsymbol{x}}^t + \lambda\overline{\boldsymbol{h}}^t\right)\right| \\ &\quad\leq \max_{1\leq j\leq m}\left|\boldsymbol{b}_j^{\mathsf{H}}\left(\frac{\boldsymbol{h}^t}{\overline{\alpha^t/\left|\alpha^t\right|}} - \frac{\boldsymbol{h}^t}{\overline{\alpha^t}}\right)\right| + \max_{1\leq j\leq m}\left|\boldsymbol{b}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^t - \boldsymbol{h}^\star\right)\right| \\ &\qquad + \eta\left(\lambda\left|\alpha^t\right|\max_{1\leq j\leq m}\left|\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right| + \left|\alpha^t\right|^{-1}\underbrace{\max_{1\leq j\leq m}\left|\boldsymbol{b}_j^{\mathsf{H}}\left(\sum_{l=1}^m\left(\boldsymbol{b}_l^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\widetilde{\boldsymbol{x}}^{t,\mathsf{H}}\boldsymbol{a}_l - y_l\right)\boldsymbol{b}_l\boldsymbol{a}_l^{\mathsf{H}}\widetilde{\boldsymbol{x}}^t\right)\right|}_{=:\tau}\right) \\ &\quad\leq \left|\left|\alpha^t\right| - 1\right|\max_{1\leq j\leq m}\left|\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right| + \max_{1\leq j\leq m}\left|\boldsymbol{b}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^t - \boldsymbol{h}^\star\right)\right| + \eta\left(2\lambda\max_{1\leq j\leq m}\left|\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right| + 2\tau\right). \end{aligned} \tag{B.26}$$

Here, the last inequality utilizes the fact $\|\boldsymbol{b}_j\|_2 = \sqrt{K/m}$ and (B.8h). The quantity τ can be controlled by using the same analysis as Appendix A.9. Specifically,

$$\begin{aligned}
\tau &= \max_{1\le j\le m} \left|\boldsymbol{b}_j^{\mathsf{H}} \nabla_{\boldsymbol{h}} f_{\text{reg-free}}\left(\widetilde{\boldsymbol{z}}^t\right)\right| \\
&\le \max_{1\le j\le m} \left(\left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\nu}_1\right| + \left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\nu}_2\right| + \left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\nu}_3\right| + \left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\nu}_4\right| + \|\boldsymbol{x}^\star\|_2^2 \left|\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right|\right) \\
&\lesssim \frac{\mu}{\sqrt{m}} \log m + \sigma,
\end{aligned}$$

where $\{\boldsymbol{\nu}_i\}_{i=1}^4$ are defined in (A.42), and the last inequality is a direct consequence of Appendix A.9. Finally, continue the bound (B.26) to demonstrate that

$$\begin{aligned}
&\max_{1\le j\le m} \left|\boldsymbol{b}_j^{\mathsf{H}}\left(\overline{\boldsymbol{h}}^t - \eta\nabla_{\boldsymbol{h}} f\left(\overline{\boldsymbol{z}}^t\right) - \boldsymbol{h}^\star\right)\right| \\
&\lesssim \frac{C_5}{c_\rho}\left(\lambda + \sigma\sqrt{K\log m}\right) C_8 \left(\frac{\mu}{\sqrt{m}}\log m + \sigma\right) + C_9\sigma + \eta\left(2C_8\lambda\left(\frac{\mu}{\sqrt{m}}\log m + \sigma\right) + 2\left(\frac{\mu}{\sqrt{m}}\log m + \sigma\right)\right) \\
&\lesssim \frac{\mu}{\sqrt{m}}\log m + \sigma,
\end{aligned}$$

where the penultimate inequality is due to (B.8h), (B.9e) and (B.2d).

B.4 Proof of Lemma 12

Before proceeding, let us introduce some additional convenient notation. Define

$$\boldsymbol{Z} := \boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}, \tag{B.27}$$

and denote by T the tangent space of $\boldsymbol{Z}$, namely,

$$T := \left\{\boldsymbol{X} : \boldsymbol{X} = \boldsymbol{h}\boldsymbol{v}^{\mathsf{H}} + \boldsymbol{u}\boldsymbol{x}^{\mathsf{H}}, \boldsymbol{v}\in\mathbb{C}^K, \boldsymbol{u}\in\mathbb{C}^K\right\}. \tag{B.28}$$

Further, define two associated projection operators as follows

$$\mathcal{P}_T(\boldsymbol{X}) := \frac{1}{\|\boldsymbol{h}\|_2^2}\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\boldsymbol{X} + \frac{1}{\|\boldsymbol{x}\|_2^2}\boldsymbol{X}\boldsymbol{x}\boldsymbol{x}^{\mathsf{H}} - \frac{1}{\|\boldsymbol{h}\|_2^2\|\boldsymbol{x}\|_2^2}\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\boldsymbol{X}\boldsymbol{x}\boldsymbol{x}^{\mathsf{H}}, \tag{B.29a}$$

$$\mathcal{P}_{T^\perp}(\boldsymbol{X}) := \left(\boldsymbol{I} - \frac{1}{\|\boldsymbol{h}\|_2^2}\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\right)\boldsymbol{X}\left(\boldsymbol{I} - \frac{1}{\|\boldsymbol{x}\|_2^2}\boldsymbol{x}\boldsymbol{x}^{\mathsf{H}}\right). \tag{B.29b}$$

We further introduce a key lemma below. It proves useful in connecting the first order optimality conditions of convex and nonconvex formulation.

Lemma 18. *Under the assumptions of Lemma 12, one has*

$$\mathcal{T}\left(\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right) - \mathcal{A}^*(\boldsymbol{\xi}) = -\frac{\lambda}{\|\boldsymbol{h}\|_2\|\boldsymbol{x}\|_2}\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} + \boldsymbol{R},$$

where $\boldsymbol{R}\in\mathbb{C}^{K\times K}$ is some residual matrix satisfying

$$\|\mathcal{P}_T(\boldsymbol{R})\|_{\mathrm{F}} \le 2\|\nabla f(\boldsymbol{h},\boldsymbol{x})\|_2 \qquad \text{and} \qquad \|\mathcal{P}_{T^\perp}(\boldsymbol{R})\| \le \lambda/2.$$

Proof. See Appendix B.4.1. □

With these supporting lemmas in hand, we are ready to prove Lemma 12. Suppose $\boldsymbol{Z}_{\mathsf{cvx}}$ is the minimizer of (3).

1. Let $\boldsymbol{\Delta} := \boldsymbol{Z}_{\mathsf{cvx}} - \boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}$. The optimality of $\boldsymbol{Z}_{\mathsf{cvx}}$ yields that

$$\left\|\mathcal{A}\left(\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} + \boldsymbol{\Delta} - \boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\right) - \boldsymbol{\xi}\right\|_2^2 + 2\lambda\left\|\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} + \boldsymbol{\Delta}\right\|_* \leq \left\|\mathcal{A}\left(\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} - \boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\right) - \boldsymbol{\xi}\right\|_2^2 + 2\lambda\left\|\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}\right\|_*.$$

By simple calculation, it leads to

$$\left\|\mathcal{A}\left(\boldsymbol{\Delta}\right)\right\|_2^2 \leq -\left\langle \mathcal{T}\left(\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} - \boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\right) - \mathcal{A}^*\left(\boldsymbol{\xi}\right), \boldsymbol{\Delta}\right\rangle + 2\lambda\left\|\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}\right\|_* - 2\lambda\left\|\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} + \boldsymbol{\Delta}\right\|_*.$$

The convexity of the nuclear norm gives that for any $\boldsymbol{W} \in \boldsymbol{T}^{\perp}$ with $\|\boldsymbol{W}\| \leq 1$, there holds

$$\left\|\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} + \boldsymbol{\Delta}\right\|_* \geq \left\|\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}\right\|_* + \left\langle \boldsymbol{p}\boldsymbol{q}^{\mathsf{H}} + \boldsymbol{W}, \boldsymbol{\Delta}\right\rangle,$$

where we denote by $\boldsymbol{p} := \boldsymbol{h}/\left\|\boldsymbol{h}\right\|_2$ and $\boldsymbol{q} := \boldsymbol{x}/\left\|\boldsymbol{x}\right\|_2$. We choose $\boldsymbol{W}$ such that $\langle \boldsymbol{W}, \boldsymbol{\Delta}\rangle = \left\|\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{\Delta}\right)\right\|_*$. Then, combining the above two equations gives rise to

$$\begin{aligned}
0 \leq \left\|\mathcal{A}\left(\boldsymbol{\Delta}\right)\right\|_2^2 &\leq -\left\langle \mathcal{T}\left(\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} - \boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\right) - \mathcal{A}^*\left(\boldsymbol{\xi}\right), \boldsymbol{\Delta}\right\rangle - 2\lambda\left\langle \boldsymbol{p}\boldsymbol{q}^{\mathsf{H}} + \boldsymbol{W}, \boldsymbol{\Delta}\right\rangle \\
&= -\left\langle \mathcal{T}\left(\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} - \boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\right) - \mathcal{A}^*\left(\boldsymbol{\xi}\right), \boldsymbol{\Delta}\right\rangle - 2\lambda\left\langle \boldsymbol{p}\boldsymbol{q}^{\mathsf{H}}, \boldsymbol{\Delta}\right\rangle - 2\lambda\left\|\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{\Delta}\right)\right\|_* \\
&\overset{\text{(i)}}{=} -\left\langle \boldsymbol{R}, \boldsymbol{\Delta}\right\rangle - 2\lambda\left\|\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{\Delta}\right)\right\|_* \\
&= -\left\langle \mathcal{P}_{\boldsymbol{T}}\left(\boldsymbol{R}\right), \boldsymbol{\Delta}\right\rangle - \left\langle \mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{R}\right), \boldsymbol{\Delta}\right\rangle - 2\lambda\left\|\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{\Delta}\right)\right\|_*, \qquad \text{(B.30)}
\end{aligned}$$

where $\boldsymbol{R}$ in (i) is defined in Lemma 18. Hence,

$$\begin{aligned}
&-\left\|\mathcal{P}_{\boldsymbol{T}}\left(\boldsymbol{R}\right)\right\|_{\mathrm{F}}\left\|\mathcal{P}_{\boldsymbol{T}}\left(\boldsymbol{\Delta}\right)\right\|_{\mathrm{F}} - \left\|\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{R}\right)\right\|\left\|\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{\Delta}\right)\right\|_* + 2\lambda\left\|\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{\Delta}\right)\right\|_* \\
&\leq \left\langle \mathcal{P}_{\boldsymbol{T}}\left(\boldsymbol{R}\right), \boldsymbol{\Delta}\right\rangle + \left\langle \mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{R}\right), \boldsymbol{\Delta}\right\rangle + 2\lambda\left\|\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{\Delta}\right)\right\|_* \leq 0.
\end{aligned}$$

Lemma 18 gives $\left\|\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{R}\right)\right\| \leq \lambda/2$, then we have

$$\left\|\mathcal{P}_{\boldsymbol{T}}\left(\boldsymbol{R}\right)\right\|_{\mathrm{F}}\left\|\mathcal{P}_{\boldsymbol{T}}\left(\boldsymbol{\Delta}\right)\right\|_{\mathrm{F}} \geq -\left\|\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{R}\right)\right\|\left\|\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{\Delta}\right)\right\|_* + 2\lambda\left\|\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{\Delta}\right)\right\|_* \geq \frac{3\lambda}{2}\left\|\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{\Delta}\right)\right\|_*,$$

and it immediately reveals that

$$\begin{aligned}
\left\|\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{\Delta}\right)\right\|_* &\leq \frac{2}{3\lambda}\left\|\mathcal{P}_{\boldsymbol{T}}\left(\boldsymbol{R}\right)\right\|_{\mathrm{F}}\left\|\mathcal{P}_{\boldsymbol{T}}\left(\boldsymbol{\Delta}\right)\right\|_{\mathrm{F}} \\
&\leq \frac{4}{3\lambda}\left\|\nabla f\left(\boldsymbol{h}, \boldsymbol{x}\right)\right\|_2\left\|\mathcal{P}_{\boldsymbol{T}}\left(\boldsymbol{\Delta}\right)\right\|_{\mathrm{F}} \\
&\leq C\frac{4}{3m^{10}}\left\|\mathcal{P}_{\boldsymbol{T}}\left(\boldsymbol{\Delta}\right)\right\|_{\mathrm{F}},
\end{aligned}$$

where the second inequality invokes Lemma 18. We then arrive at

$$\left\|\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{\Delta}\right)\right\|_{\mathrm{F}} \leq \left\|\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{\Delta}\right)\right\|_* \leq C\frac{4}{3m^{10}}\left\|\mathcal{P}_{\boldsymbol{T}}\left(\boldsymbol{\Delta}\right)\right\|_{\mathrm{F}} \leq \left\|\mathcal{P}_{\boldsymbol{T}}\left(\boldsymbol{\Delta}\right)\right\|_{\mathrm{F}}. \qquad \text{(B.31)}$$

2. Next, we return to (B.30) to deduce that

$$\begin{aligned}
\left\|\mathcal{A}\left(\boldsymbol{\Delta}\right)\right\|_2^2 &\leq -\left\langle \mathcal{P}_{\boldsymbol{T}}\left(\boldsymbol{R}\right), \boldsymbol{\Delta}\right\rangle - \left\langle \mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{R}\right), \boldsymbol{\Delta}\right\rangle - 2\lambda\left\|\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{\Delta}\right)\right\|_* \\
&\leq \left\|\mathcal{P}_{\boldsymbol{T}}\left(\boldsymbol{R}\right)\right\|_{\mathrm{F}}\left\|\mathcal{P}_{\boldsymbol{T}}\left(\boldsymbol{\Delta}\right)\right\|_{\mathrm{F}} + \left\|\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{R}\right)\right\|\left\|\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{\Delta}\right)\right\|_* - 2\lambda\left\|\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{\Delta}\right)\right\|_* \qquad \text{(B.32)} \\
&\overset{\text{(i)}}{\leq} \left\|\mathcal{P}_{\boldsymbol{T}}\left(\boldsymbol{R}\right)\right\|_{\mathrm{F}}\left\|\mathcal{P}_{\boldsymbol{T}}\left(\boldsymbol{\Delta}\right)\right\|_{\mathrm{F}} - \frac{3\lambda}{2}\left\|\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{\Delta}\right)\right\|_* \\
&\leq \left\|\mathcal{P}_{\boldsymbol{T}}\left(\boldsymbol{R}\right)\right\|_{\mathrm{F}}\left\|\mathcal{P}_{\boldsymbol{T}}\left(\boldsymbol{\Delta}\right)\right\|_{\mathrm{F}} \qquad \text{(B.33)} \\
&\overset{\text{(ii)}}{\leq} 2\left\|\nabla f\left(\boldsymbol{h}, \boldsymbol{x}\right)\right\|_2\left\|\boldsymbol{\Delta}\right\|_{\mathrm{F}}, \qquad \text{(B.34)}
\end{aligned}$$

where (i) and (ii) come from Lemma 18.

3. For the final step, we turn to lower bound $\|\mathcal{A}(\boldsymbol{\Delta})\|_{\mathrm{F}}$. One has

$$\begin{aligned}\|\mathcal{A}(\boldsymbol{\Delta})\|_2 &= \|\mathcal{A}(\mathcal{P}_{\boldsymbol{T}}(\boldsymbol{\Delta})) + \mathcal{A}(\mathcal{P}_{\boldsymbol{T}^\perp}(\boldsymbol{\Delta}))\|_2 \\ &\geq \|\mathcal{A}(\mathcal{P}_{\boldsymbol{T}}(\boldsymbol{\Delta}))\|_2 - \|\mathcal{A}(\mathcal{P}_{\boldsymbol{T}^\perp}(\boldsymbol{\Delta}))\|_2 \\ &\geq \|\mathcal{P}_{\boldsymbol{T}}(\boldsymbol{\Delta})\|_{\mathrm{F}}/4 - \sqrt{2K\log K + \gamma \log m}\,\|\mathcal{P}_{\boldsymbol{T}^\perp}(\boldsymbol{\Delta})\|_{\mathrm{F}}, \end{aligned} \tag{B.35}$$

where the last inequality comes from Lemma 15 and Lemma 1. Since (B.31) gives

$$\sqrt{2K\log K + \gamma\log m}\,\|\mathcal{P}_{\boldsymbol{T}^\perp}(\boldsymbol{\Delta})\|_{\mathrm{F}} \leq \sqrt{2K\log K + \gamma\log m} \times C\frac{4}{3m^{10}}\|\mathcal{P}_{\boldsymbol{T}}(\boldsymbol{\Delta})\|_{\mathrm{F}} \leq \frac{1}{8}\|\mathcal{P}_{\boldsymbol{T}}(\boldsymbol{\Delta})\|_{\mathrm{F}},$$

as long as $m \gg K$, (B.35) yields

$$\|\mathcal{A}(\boldsymbol{\Delta})\|_2 \geq \frac{1}{8}\|\mathcal{P}_{\boldsymbol{T}}(\boldsymbol{\Delta})\|_{\mathrm{F}}.$$

In addition, (B.31) implies

$$\|\boldsymbol{\Delta}\|_{\mathrm{F}} \leq \|\mathcal{P}_{\boldsymbol{T}}(\boldsymbol{\Delta})\|_{\mathrm{F}} + \|\mathcal{P}_{\boldsymbol{T}^\perp}(\boldsymbol{\Delta})\|_{\mathrm{F}} \leq 2\|\mathcal{P}_{\boldsymbol{T}}(\boldsymbol{\Delta})\|_{\mathrm{F}}.$$

Consequently,

$$\|\mathcal{A}(\boldsymbol{\Delta})\|_2 \geq \frac{1}{8}\|\mathcal{P}_{\boldsymbol{T}}(\boldsymbol{\Delta})\|_{\mathrm{F}} \geq \frac{1}{16}\|\boldsymbol{\Delta}\|_{\mathrm{F}}. \tag{B.36}$$

Combining (B.33) and (B.36), we have

$$\frac{1}{256}\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 \leq \|\mathcal{A}(\boldsymbol{\Delta})\|_2^2 \leq 2\|\nabla f(\boldsymbol{h},\boldsymbol{x})\|_2\|\boldsymbol{\Delta}\|_{\mathrm{F}},$$

and therefore

$$\|\boldsymbol{\Delta}\|_{\mathrm{F}} \lesssim \|\nabla f(\boldsymbol{h},\boldsymbol{x})\|_2.$$

B.4.1 Proof of Lemma 18

Recall the definition of $\mathcal{T}^{\mathsf{debias}}$ in (B.3). Letting

$$\boldsymbol{p} = \frac{1}{\|\boldsymbol{h}\|_2}\boldsymbol{h} \qquad \text{and} \qquad \boldsymbol{q} = \frac{1}{\|\boldsymbol{x}\|_2}\boldsymbol{x} \tag{B.37}$$

and rearranging terms, we can write

$$\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}} + \mathcal{T}^{\mathsf{debias}}\left(\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}} - \boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}\right) + \mathcal{A}^*(\boldsymbol{\xi}) = \boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} + \lambda\boldsymbol{p}\boldsymbol{q}^{\mathsf{H}} + \boldsymbol{R} \tag{B.38}$$

for some matrix $\boldsymbol{R}$. In addition, in view of the small gradient assumption (B.4a), one has

$$\left[\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}} + \mathcal{T}^{\mathsf{debias}}\left(\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}} - \boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}\right) + \mathcal{A}^*(\boldsymbol{\xi})\right]\boldsymbol{x} = \boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}\boldsymbol{x} + \lambda\boldsymbol{h} - \boldsymbol{r}_1 \tag{B.39a}$$

$$\left[\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}} + \mathcal{T}^{\mathsf{debias}}\left(\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}} - \boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}\right) + \mathcal{A}^*(\boldsymbol{\xi})\right]^{\mathsf{H}}\boldsymbol{h} = \boldsymbol{x}\boldsymbol{h}^{\mathsf{H}}\boldsymbol{h} + \lambda\boldsymbol{x} - \boldsymbol{r}_2 \tag{B.39b}$$

for some vectors $\boldsymbol{r}_1, \boldsymbol{r}_2 \in \mathbb{C}^K$ obeying

$$\|\boldsymbol{r}_1\|_2 = \left\|\lambda\boldsymbol{h} - \left(\mathcal{T}\left(\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}} - \boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}\right) + \mathcal{A}^*(\boldsymbol{\xi})\right)\boldsymbol{x}\right\|_2 \leq \|\nabla f(\boldsymbol{h},\boldsymbol{x})\|_2 \leq C\frac{\lambda}{m^{10}}, \tag{B.40a}$$

$$\|\boldsymbol{r}_2\|_2 = \left\|\lambda\boldsymbol{x} - \left(\mathcal{T}\left(\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}} - \boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}\right) + \mathcal{A}^*(\boldsymbol{\xi})\right)^{\mathsf{H}}\boldsymbol{h}\right\|_2 \leq \|\nabla f(\boldsymbol{h},\boldsymbol{x})\|_2 \leq C\frac{\lambda}{m^{10}}. \tag{B.40b}$$

In what follows, we make of these properties to control the size of $\boldsymbol{R}$.

1. We start by upper bounding $\left\|\mathcal{P}_T(\boldsymbol{R})\right\|_{\mathrm{F}}$ as follows

$$\begin{aligned}\left\|\mathcal{P}_T(\boldsymbol{R})\right\|_{\mathrm{F}} &= \left\|\boldsymbol{p}\boldsymbol{p}^{\mathsf{H}}\boldsymbol{R}\left(\boldsymbol{I}_K-\boldsymbol{q}\boldsymbol{q}^{\mathsf{H}}\right)+\boldsymbol{R}\boldsymbol{q}\boldsymbol{q}^{\mathsf{H}}\right\|_{\mathrm{F}} \\ &\le \left\|\boldsymbol{p}\right\|_2\left\|\boldsymbol{p}^{\mathsf{H}}\boldsymbol{R}\right\|_2\left\|\boldsymbol{I}_K-\boldsymbol{q}\boldsymbol{q}^{\mathsf{H}}\right\|+\left\|\boldsymbol{R}\boldsymbol{q}\right\|_2\left\|\boldsymbol{q}\right\|_2 \\ &\le \left\|\boldsymbol{p}^{\mathsf{H}}\boldsymbol{R}\right\|_2+\left\|\boldsymbol{R}\boldsymbol{q}\right\|_2,\end{aligned}$$

where $\boldsymbol{p}$ and $\boldsymbol{q}$ are unit vectors defined in (B.37). Recognizing that $\|\boldsymbol{h}\|_2=\|\boldsymbol{x}\|_2$ (cf. (B.5a)), we can use (B.38) and (B.39) to obtain

$$\boldsymbol{R}^{\mathsf{H}}\boldsymbol{p}=-\frac{\boldsymbol{r}_2}{\|\boldsymbol{h}\|_2}+\lambda\frac{\|\boldsymbol{x}\|_2}{\|\boldsymbol{h}\|_2}\boldsymbol{q}-\lambda\frac{\|\boldsymbol{h}\|_2}{\|\boldsymbol{x}\|_2}\boldsymbol{q}=-\frac{\boldsymbol{r}_2}{\|\boldsymbol{h}\|_2}\quad\text{and}\quad\boldsymbol{R}\boldsymbol{q}=-\frac{\boldsymbol{r}_1}{\|\boldsymbol{x}\|_2}.$$

These together with (B.40) yield

$$\left\|\mathcal{P}_T(\boldsymbol{R})\right\|_{\mathrm{F}}\le\left\|\boldsymbol{p}^{\mathsf{H}}\boldsymbol{R}\right\|_2+\left\|\boldsymbol{R}\boldsymbol{q}\right\|_2\le 2\left\|\nabla f(\boldsymbol{h},\boldsymbol{x})\right\|_2\le 2C\frac{\lambda}{m^{10}}. \tag{B.41}$$

2. We them move on to control $\mathcal{P}_{T^\perp}(\boldsymbol{R})$. Continue the relation (B.38) to derive

$$\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}+\mathcal{T}^{\mathsf{debias}}\left(\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}-\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}\right)+\mathcal{A}^*(\boldsymbol{\xi})-\mathcal{P}_T(\boldsymbol{R})=\boldsymbol{p}\left(\|\boldsymbol{h}\|_2\|\boldsymbol{x}\|_2+\lambda\frac{\|\boldsymbol{h}\|_2}{\|\boldsymbol{x}\|_2}\right)\boldsymbol{q}^{\mathsf{H}}+\mathcal{P}_{T^\perp}(\boldsymbol{R}), \tag{B.42}$$

where we have used the assumption $\|\boldsymbol{h}\|_2/\|\boldsymbol{x}\|_2=1$ (cf. (B.5a)). Combine this with Lemma 13, Lemma 14 and (B.41) to derive

$$\begin{aligned}\left\|\mathcal{T}^{\mathsf{debias}}\left(\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}-\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}\right)+\mathcal{A}^*(\boldsymbol{\xi})-\mathcal{P}_T(\boldsymbol{R})\right\| &\le \left\|\mathcal{T}^{\mathsf{debias}}\left(\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}-\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}\right)\right\|+\left\|\mathcal{A}^*(\boldsymbol{\xi})\right\|+\left\|\mathcal{P}_T(\boldsymbol{R})\right\|_{\mathrm{F}} \\ &\le \frac{\lambda}{8}+\frac{\lambda}{8}+2C\frac{\lambda}{m^{10}} \\ &< \frac{\lambda}{2},\end{aligned}$$

where the last inequality invokes the assumption (B.2e). Invoking (B.42) and Weyl's inequality give

$$\begin{aligned}\sigma_i\left[\boldsymbol{p}\left(\|\boldsymbol{h}\|_2\|\boldsymbol{x}\|_2+\lambda\frac{\|\boldsymbol{h}\|_2}{\|\boldsymbol{x}\|_2}\right)\boldsymbol{q}^{\mathsf{H}}+\mathcal{P}_{T^\perp}(\boldsymbol{R})\right] &\le \sigma_i\left(\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right)+\left\|\mathcal{T}^{\mathsf{debias}}\left(\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}-\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}\right)+\mathcal{A}^*(\boldsymbol{\xi})-\mathcal{P}_T(\boldsymbol{R})\right\| \\ &< \lambda/2,\end{aligned}$$

for $K\ge i\ge 2$. Additionally, when $i=1$, we have

$$\sigma_1\left[\boldsymbol{p}\left(\|\boldsymbol{h}\|_2\|\boldsymbol{x}\|_2+\lambda\frac{\|\boldsymbol{h}\|_2}{\|\boldsymbol{x}\|_2}\right)\boldsymbol{q}^{\mathsf{H}}\right]=\|\boldsymbol{h}\|_2\|\boldsymbol{x}\|_2+\lambda\frac{\|\boldsymbol{h}\|_2}{\|\boldsymbol{x}\|_2}\ge\lambda/2.$$

This indicates that at least $K-1$ singular values of $\boldsymbol{p}\left(\|\boldsymbol{h}\|_2\|\boldsymbol{x}\|_2+\lambda\|\boldsymbol{h}\|_2/\|\boldsymbol{x}\|_2\right)\boldsymbol{q}^{\mathsf{H}}+\mathcal{P}_{T^\perp}(\boldsymbol{R})$ are no larger than $\lambda/2$, and these singular values cannot correspond to the direction of $\boldsymbol{p}\boldsymbol{q}^{\mathsf{H}}$. As a consequence, we conclude that

$$\left\|\mathcal{P}_{T^\perp}(\boldsymbol{R})\right\|\le\lambda/2.$$

B.5 Proof of Lemma 13

For notational convenience, we define $\mathcal{T}^{\mathsf{debias}}$ by subtracting the expectation from $\mathcal{T}$ as follows:

$$\mathcal{T}^{\mathsf{debias}}(\boldsymbol{Z}):=\mathcal{T}(\boldsymbol{Z})-\boldsymbol{Z}=\left(\mathcal{A}^*\mathcal{A}-\mathcal{I}\right)(\boldsymbol{Z})=\sum_{j=1}^{m}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{Z}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}-\boldsymbol{Z}.$$

For any fixed vectors $\boldsymbol{h}$ and $\boldsymbol{x}$, we make note of the following decomposition

$$\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}-\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}=\left(\boldsymbol{\Delta}_{\boldsymbol{h}}+\boldsymbol{h}^\star\right)\left(\boldsymbol{\Delta}_{\boldsymbol{x}}+\boldsymbol{x}^\star\right)^{\mathsf{H}}-\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}$$

$$= \boldsymbol{h}^\star \boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}} + \boldsymbol{\Delta}_{\boldsymbol{h}} \boldsymbol{x}^{\star\mathsf{H}} + \boldsymbol{\Delta}_{\boldsymbol{h}} \boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}},$$

which together with the triangle inequality gives

$$\left\| \mathcal{T}^{\mathsf{debias}} \left(\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} - \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}} \right) \right\| \leq \underbrace{\left\| \mathcal{T}^{\mathsf{debias}} \left(\boldsymbol{h}^\star \boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}} \right) \right\|}_{=:\beta_1} + \underbrace{\left\| \mathcal{T}^{\mathsf{debias}} \left(\boldsymbol{\Delta}_{\boldsymbol{h}} \boldsymbol{x}^{\star\mathsf{H}} \right) \right\|}_{=:\beta_2} + \underbrace{\left\| \mathcal{T}^{\mathsf{debias}} \left(\boldsymbol{\Delta}_{\boldsymbol{h}} \boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}} \right) \right\|}_{=:\beta_3}.$$

In what follows, we shall upper bound β_1, β_2 and β_3 separately.

1. For any fixed $\boldsymbol{x}$, the quantity β_1 is concerned with a matrix that can be written explicitly as follows

$$\mathcal{T}^{\mathsf{debias}} \left(\boldsymbol{h}^\star \boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}} \right) = \sum_{j=1}^{m} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}} \left(\boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} - \boldsymbol{I}_K \right).$$

Consequently, for any fixed unit vectors $\boldsymbol{u}$, $\boldsymbol{v} \in \mathbb{C}^K$ one has

$$\boldsymbol{u}^{\mathsf{H}} \mathcal{T}^{\mathsf{debias}} \left(\boldsymbol{h}^\star \boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}} \right) \boldsymbol{v} = \sum_{j=1}^{m} \left(\boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{v} - \boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}} \boldsymbol{v} \right),$$

which is essentially a sum of independent variables. Letting $r := \lambda + \sigma\sqrt{K \log m}$ and $C_4 := 10 \max\{C_1, C_3, 1\}$, we can deduce that

$$\begin{aligned}
&\sum_{j=1}^{m} \Bigg(\underbrace{\boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{v} \, \mathbb{1}_{\left\{ |\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}} \boldsymbol{a}_j| \leq C_4 r \sqrt{\log m} \right\}}}_{=:z_j} - \boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}} \boldsymbol{v} \Bigg) \\
&\quad = \sum_{j=1}^{m} \left(z_j - \mathbb{E}[z_j] \right) + \sum_{j=1}^{m} \left(\mathbb{E}\left[\boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{v} \, \mathbb{1}_{\left\{ |\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}} \boldsymbol{a}_j| \leq C_4 r \sqrt{\log m} \right\}} \right] - \boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}} \boldsymbol{v} \right) \\
&\quad = \sum_{j=1}^{m} \left(z_j - \mathbb{E}[z_j] \right) + \sum_{j=1}^{m} \left(\mathbb{E}\left[\boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{v} \, \mathbb{1}_{\left\{ |\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}} \boldsymbol{a}_j| \leq C_4 r \sqrt{\log m} \right\}} \right] - \mathbb{E}\left[\boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{v} \right] \right) \\
&\quad = \underbrace{\sum_{j=1}^{m} \left(z_j - \mathbb{E}[z_j] \right)}_{=:\omega_1} - \underbrace{\sum_{j=1}^{m} \mathbb{E}\left[\boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{v} \, \mathbb{1}_{\left\{ |\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}} \boldsymbol{a}_j| > C_4 r \sqrt{\log m} \right\}} \right]}_{=:\omega_2}.
\end{aligned}$$

- The term ω_2 can be controlled by Cauchy-Schwarz as follows

$$\begin{aligned}
|\omega_2| &= \left| \sum_{j=1}^{m} \mathbb{E}\left[\boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{v} \, \mathbb{1}_{\left\{ |\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}} \boldsymbol{a}_j| > C_4 r \sqrt{\log m} \right\}} \right] \right| \\
&\overset{\text{(i)}}{\leq} \sum_{j=1}^{m} \sqrt{ \mathbb{E}\left[\left| \boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{v} \right|^2 \right] \mathbb{P}\left[|\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}} \boldsymbol{a}_j| > C_4 r \sqrt{\log m} \right] } \\
&\overset{\text{(ii)}}{\leq} \sum_{j=1}^{m} \left| \boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \right| \sqrt{ \left(2 \left| \boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}} \boldsymbol{v} \right|^2 + \|\boldsymbol{\Delta}_{\boldsymbol{x}}\|_2^2 \|\boldsymbol{v}\|_2^2 \right) 2 \exp\left(-\frac{C_4^2 r^2 \log m}{2 \|\boldsymbol{\Delta}_{\boldsymbol{x}}\|_2^2} \right) } \\
&\leq \sum_{j=1}^{m} \left| \boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \right| \sqrt{ 6 \|\boldsymbol{\Delta}_{\boldsymbol{x}}\|_2^2 \exp\left(-50 \log m \right) } \\
&\overset{\text{(iii)}}{\leq} \sum_{j=1}^{m} \left(\left| \boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \right|^2 + \left| \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \right|^2 \right) \frac{\sqrt{6} \|\boldsymbol{\Delta}_{\boldsymbol{x}}\|_2}{2 m^{25}} \qquad \text{(B.43)} \\
&\overset{\text{(iv)}}{\leq} \left(1 + \mu^2 \right) \frac{\sqrt{6} \|\boldsymbol{\Delta}_{\boldsymbol{x}}\|_2}{2 m^{25}}
\end{aligned}$$

$$\overset{\text{(v)}}{\leq} \frac{\|\boldsymbol{\Delta}_{\boldsymbol{x}}\|_2}{m^{24}}.$$

Here, (i) follows from the Cauchy-Schwarz inequality, and (ii) comes from the property of sub-Gaussian variable $\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j$ and

$$\begin{aligned} \mathbb{E}\left[\left|\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\right|^2\right] &= \left|\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\right|^2 \mathbb{E}\left[\left|\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\right|^2\right] \\ &= \left|\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\right| \left(2\left|\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{v}\right|^2 + \|\boldsymbol{\Delta}_{\boldsymbol{x}}\|_2^2\|\boldsymbol{v}\|_2^2\right), \end{aligned} \tag{B.44}$$

where the last line is due to the property of Gaussian distributions. In addition, (iii) is a consequence of the elementary inequality $|ab| \leq (|a|^2+|b|^2)/2$, (iv) comes from the incoherence condition (7) and $\sum_{j=1}^m \left|\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\right|^2 = \|\boldsymbol{u}\|_2^2$, whereas (v) holds true as long as $m \gg \mu^2$.

- Regarding ω_1, note that z_j is a sub-Gaussian random variable obeying

$$\|z_j - \mathbb{E}[z_j]\|_{\psi_2} \lesssim \left|C_4 r\sqrt{\log m}\left(\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\right)\left(\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\right)\right| \leq C_4\frac{\mu\sqrt{\log m}}{\sqrt{m}} r\left|\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\right|.$$

Therefore, by invoking Hoeffding's inequality (cf. Vershynin [2018, Theorem 2.6.2]) we reach

$$\mathbb{P}\left(\left|\sum_{j=1}^m z_j - \mathbb{E}[z_j]\right| \geq t\right) \leq 2\exp\left(-\frac{ct^2}{\frac{C_4^2\mu^2r^2\log m}{m}\sum_{j=1}^m\left|\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\right|^2}\right) = 2\exp\left(-\frac{ct^2}{\frac{C_4^2\mu^2r^2\log m}{m}}\right)$$

for any $t \geq 0$. Setting $t = \frac{C\mu r\sqrt{K}\log m}{\sqrt{m}}$ for some sufficiently large constant $C>0$ yields

$$\mathbb{P}\left(\left|\sum_{j=1}^m z_j - \mathbb{E}[z_j]\right| \geq \frac{C\mu r\sqrt{K}\log m}{\sqrt{m}}\right) \leq 2\exp\left(-10K\log m\right). \tag{B.45}$$

Next, we define $\mathcal{N}_{\boldsymbol{x}}$ to be an ε_1-net of $\mathcal{B}_{\boldsymbol{x}}\left(\frac{C_5}{1-\rho}\eta r\right) := \left\{\boldsymbol{x} : \|\boldsymbol{x}-\boldsymbol{x}^\star\| \leq \frac{C_5}{1-\rho}\eta r\right\}$, and $\mathcal{N}_0$ an ε_2-net of the unit sphere $\mathcal{S}^{K-1} = \left\{\boldsymbol{u}\in\mathbb{C}^K : \|\boldsymbol{u}\|_2 = 1\right\}$, where we take $\varepsilon_1 = r/(m\log m)$ and $\varepsilon_2 = 1/(m\log m)$. In view of Vershynin [2018, Corollary 4.2.13], one can ensure that

$$|\mathcal{N}_{\boldsymbol{x}}| \leq \left(1 + \frac{2C_5\eta r}{(1-\rho)\varepsilon_1}\right)^{2K} \quad \text{and} \quad |\mathcal{N}_0| \leq \left(1+\frac{2}{\varepsilon_2}\right)^{2K}.$$

This together with the union bound leads to

$$\left|\sum_{j=1}^m z_j - \mathbb{E}[z_j]\right| \geq \frac{C\mu r\sqrt{K}\log m}{\sqrt{m}},$$

which holds uniformly for any $\boldsymbol{x}\in\mathcal{N}_{\boldsymbol{x}}$, $\boldsymbol{u},\boldsymbol{v}\in\mathcal{N}_0$ and holds with probability at least

$$1 - \left(1 + \frac{2C_5\eta r}{(1-\rho)\varepsilon_1}\right)^{2K}\left(1+\frac{2}{\varepsilon_2}\right)^{4K}\cdot 2e^{-10K\log m} \geq 1 - O\left(m^{-100}\right).$$

As a result, with probability exceeding $1 - O\left(m^{-10} + me^{-CK}\right)$ there holds

$$\begin{aligned} &\left|\sum_{j=1}^m\left(\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{|\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j|\leq C_4r\sqrt{\log m}\right\}} - \boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{v}\right)\right| \\ &\quad \leq \left|\sum_{j=1}^m\left(z_j - \mathbb{E}[z_j]\right)\right| + \left|\sum_{j=1}^m \mathbb{E}\left[\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{|\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j|\leq C_4r\sqrt{\log m}\right\}}\right]\right| \end{aligned}$$

$$\leq \frac{C\mu r\sqrt{K}\log m}{\sqrt{m}} + \frac{\|\boldsymbol{\Delta}_{\boldsymbol{x}}\|_2}{m^{24}}$$

$$\leq \frac{\lambda}{100} \tag{B.46}$$

uniformly for any $\boldsymbol{x} \in \mathcal{N}_{\boldsymbol{x}}$, $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{N}_0$. Here, the penultimate inequality comes from (B.43) and (B.45). For any $\boldsymbol{x}$ obeying the assumption $\max_j \left| (\boldsymbol{x} - \boldsymbol{x}^\star)^{\mathsf{H}} \boldsymbol{a}_j \right| \leq C_3 r\sqrt{\log m}$ and any $\boldsymbol{u}$, $\boldsymbol{v} \in \mathcal{S}^{K-1}$, we can find $\boldsymbol{x}_0 \in \mathcal{N}_{\boldsymbol{x}}$, $\boldsymbol{u}_0 \in \mathcal{N}_0$ and $\boldsymbol{v}_0 \in \mathcal{N}_0$ satisfying $\|\boldsymbol{x} - \boldsymbol{x}_0\|_2 \leq \varepsilon_1$ and $\max\{\|\boldsymbol{u} - \boldsymbol{u}_0\|_2, \|\boldsymbol{v} - \boldsymbol{v}_0\|_2\} \leq \varepsilon_2$. Given that $\max_j \|\boldsymbol{a}_j\|_2 \leq 10\sqrt{K}$ with probability $1 - me^{-CK}$ for some constant $C > 0$, this yields that

$$\left|\boldsymbol{\Delta}_{\boldsymbol{x}_0}^{\mathsf{H}} \boldsymbol{a}_j\right| \leq \left|\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}} \boldsymbol{a}_j\right| + 10\varepsilon_1\sqrt{K} \leq 2C_3 \left(\lambda + \sigma\sqrt{K\log m}\right)\sqrt{\log m}.$$

Recalling $C_4 \geq 10C_3$, we have

$$\left|\boldsymbol{\Delta}_{\boldsymbol{x}_0}^{\mathsf{H}} \boldsymbol{a}_j\right| \leq C_4 \left(\lambda + \sigma\sqrt{K\log m}\right)\sqrt{\log m} = C_4 r\sqrt{\log m},$$

and hence $\mathbb{1}_{\left\{\left|\boldsymbol{\Delta}_{\boldsymbol{x}_0}^{\mathsf{H}} \boldsymbol{a}_j\right| \leq C_4 r\sqrt{\log m}\right\}} = 1$, $\forall j$. Therefore, if we let

$$f(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) := \sum_{j=1}^m \left(\boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star (\boldsymbol{x} - \boldsymbol{x}^\star)^{\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{v} \, \mathbb{1}_{\left\{\left|(\boldsymbol{x}-\boldsymbol{x}^\star)^{\mathsf{H}} \boldsymbol{a}_j\right| \leq C_4 r\sqrt{\log m}\right\}} - \boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star (\boldsymbol{x} - \boldsymbol{x}^\star)^{\mathsf{H}} \boldsymbol{v} \right),$$

then we can demonstrate that

$$\begin{aligned}
&|f(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) - f(\boldsymbol{x}_0, \boldsymbol{u}_0, \boldsymbol{v}_0)| \\
&\quad\leq \left| \sum_{j=1}^m \boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star (\boldsymbol{x} - \boldsymbol{x}_0)^{\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{v} \right| + \left| \sum_{j=1}^m \boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star (\boldsymbol{x} - \boldsymbol{x}_0)^{\mathsf{H}} \boldsymbol{v} \right| \\
&\qquad + \left| \sum_{j=1}^m (\boldsymbol{u} - \boldsymbol{u}_0)^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star (\boldsymbol{x}_0 - \boldsymbol{x}^\star)^{\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{v} \right| + \left| \sum_{j=1}^m (\boldsymbol{u} - \boldsymbol{u}_0)^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star (\boldsymbol{x}_0 - \boldsymbol{x}^\star)^{\mathsf{H}} \boldsymbol{v} \right| \\
&\qquad + \left| \sum_{j=1}^m \boldsymbol{u}_0^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star (\boldsymbol{x}_0 - \boldsymbol{x}^\star)^{\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} (\boldsymbol{v} - \boldsymbol{v}_0) \right| + \left| \sum_{j=1}^m \boldsymbol{u}_0^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star (\boldsymbol{x}_0 - \boldsymbol{x}^\star)^{\mathsf{H}} (\boldsymbol{v} - \boldsymbol{v}_0) \right| \\
&\quad\leq \left(\|\mathcal{A}\|^2 + 1\right) \left(\|\boldsymbol{h}^\star\|_2 \|\boldsymbol{x} - \boldsymbol{x}_0\|_2 + \|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2 \|\boldsymbol{u} - \boldsymbol{u}_0\|_2 + \|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2 \|\boldsymbol{v} - \boldsymbol{v}_0\|_2\right) \\
&\quad\leq (2K\log K + 10\log m + 1)(\varepsilon_1 + 2C_1 r\varepsilon_2),
\end{aligned}$$

where the last inequality arises from (B.66). Consequently,

$$\begin{aligned}
&\left|\boldsymbol{u}^{\mathsf{H}} \mathcal{T}^{\mathsf{debias}} \left(\boldsymbol{h}^\star (\boldsymbol{x} - \boldsymbol{x}^\star)^{\mathsf{H}}\right) \boldsymbol{v}\right| \\
&\quad= \left| \sum_{j=1}^m \left(\boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star (\boldsymbol{x} - \boldsymbol{x}^\star)^{\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{v} \, \mathbb{1}_{\left\{\left|(\boldsymbol{x}-\boldsymbol{x}^\star)^{\mathsf{H}} \boldsymbol{a}_j\right| \leq C_4 r\sqrt{\log m}\right\}} - \boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star (\boldsymbol{x} - \boldsymbol{x}^\star)^{\mathsf{H}} \boldsymbol{v} \right) \right| \\
&\quad\leq |f(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) - f(\boldsymbol{x}_0, \boldsymbol{u}_0, \boldsymbol{v}_0)| + |f(\boldsymbol{x}_0, \boldsymbol{u}_0, \boldsymbol{v}_0)| \\
&\quad\leq (2K\log K + 10\log m + 1)(\varepsilon_1 + 2C_1 r\varepsilon_2) + \frac{\lambda}{100} \\
&\quad\leq \frac{\lambda}{50},
\end{aligned}$$

where the last inequality is due to the definitions $r = \lambda + \sigma\sqrt{K\log m}$, $\varepsilon_1 = r/(m\log m)$, $\varepsilon_2 = 1/(m\log m)$ and $m \gg K$. Therefore, for any $(\boldsymbol{h}, \boldsymbol{x})$ satisfying (B.5), there holds

$$\left\|\mathcal{T}^{\mathsf{debias}}\left(\boldsymbol{h}^\star \boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\right)\right\| = \sup_{\boldsymbol{u}, \boldsymbol{v} \in \mathcal{S}^{K-1}} \boldsymbol{u}^{\mathsf{H}} \mathcal{T}^{\mathsf{debias}} \left(\boldsymbol{h}^\star (\boldsymbol{x} - \boldsymbol{x}^\star)^{\mathsf{H}}\right) \boldsymbol{v} \leq \frac{1}{50}\lambda \tag{B.47}$$

with probability exceeding $1 - O\left(m^{-10} + me^{-CK}\right)$.

2. We now move on to β_2, for which we have a similar decomposition as follows

$$\begin{aligned}
&\boldsymbol{u}^{\mathsf{H}}\mathcal{T}^{\mathsf{debias}}\left(\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{x}^{\star\mathsf{H}}\right)\boldsymbol{v}\\
&=\sum_{j=1}^{m}\left(\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}-\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{v}\right)\\
&=\sum_{j=1}^{m}\Bigg(\underbrace{\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{|\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j|\le 20\sqrt{\log m}\right\}}}_{=:y_j}-\mathbb{E}\left[y_j\right]\Bigg)\\
&\quad\underbrace{-\sum_{j=1}^{m}\mathbb{E}\left[\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{|\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j|> 20\sqrt{\log m}\right\}}\right]}_{=:\omega_4}+\sum_{j=1}^{m}\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{|\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j|> 20\sqrt{\log m}\right\}}.
\end{aligned}$$

- For ω_4, similar to (B.43) we have

$$\begin{aligned}
|\omega_4| &=\left|\sum_{j=1}^{m}\mathbb{E}\left[\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{|\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j|> 20\sqrt{\log m}\right\}}\right]\right|\\
&\overset{\text{(i)}}{\le}\sum_{j=1}^{m}\sqrt{\mathbb{E}\left[\left|\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\right|^2\right]\mathbb{P}\left(\left|\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\right|>20\sqrt{\log m}\right)}\\
&\overset{\text{(ii)}}{\le}\sum_{j=1}^{m}\left|\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\right|\sqrt{\left(2\left|\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{v}\right|^2+\left\|\boldsymbol{x}^{\star}\right\|_2^2\left\|\boldsymbol{v}\right\|_2^2\right)2\exp\left(-200\log m\right)}\\
&\le\sum_{j=1}^{m}\left|\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\right|\frac{4}{m^{100}}\\
&\overset{\text{(iii)}}{\le}\sum_{j=1}^{m}\left\|\boldsymbol{b}_j\right\|_2\times C_9\sigma\times\frac{4}{m^{100}} \qquad \text{(B.48)}\\
&\overset{\text{(iv)}}{\le}\sqrt{\frac{K}{m}}\times m\times C_9\sigma\times\frac{4}{m^{100}}\\
&\le\frac{\lambda}{m^{99}}, \qquad \text{(B.49)}
\end{aligned}$$

  where (i) follows from Cauchy-Schwarz inequality, (ii) comes from the property of sub-Gaussian variable $\left|\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\right|$ and (B.44), (iii) is due to the assumption (B.5b), and (iv) comes from the fact $\left\|\boldsymbol{b}_j\right\|_2=\sqrt{K/m}$.

- Regarding the term $\omega_3 \coloneqq \sum_{j=1}^{m}\left(y_j-\mathbb{E}\left[y_j\right]\right)$, we note that

$$\left\|\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{|\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j|\le 20\sqrt{\log m}\right\}}\right\|_{\psi_2}\le\frac{\mu\lambda}{\sqrt{m}}\log^2 m\times 20\sqrt{\log m}\left|\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\right|.$$

  Hoeffding's inequality Vershynin [2018, Theorem 2.6.3] tells us that

$$\mathbb{P}\left(\left|\sum_{j=1}^{m}\left(y_j-\mathbb{E}\left[y_j\right]\right)\right|\ge t\right)\le 2\exp\left(-\frac{ct^2}{400\frac{\mu^2\lambda^2}{m}\log^5 m\sum_{j=1}^{m}\left|\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\right|^2}\right)=2\exp\left(-\frac{ct^2}{400\frac{\mu^2\lambda^2}{m}\log^5 m}\right)$$

  for any $t\ge 0$. Setting $t=\frac{C\mu\lambda\sqrt{K}}{\sqrt{m}}\log^3 m$ for some sufficiently large constant $C>0$ yields

$$\mathbb{P}\left(\left|\sum_{j=1}^{m}\left(y_j-\mathbb{E}\left[y_j\right]\right)\right|\ge C\frac{\mu\lambda\sqrt{K}}{\sqrt{m}}\log^3 m\right)\le 2\exp\left(-10K\log m\right). \qquad \text{(B.50)}$$

Invoking a similar covering argument, we know that with probability exceeding $1 - O\left(m^{-10}\right)$,

$$\left|\sum_{j=1}^{m}\left(y_j - \mathbb{E}\left[y_j\right]\right)\right| \geq C\frac{\mu\lambda\sqrt{K}}{\sqrt{m}}\log^3 m$$

holds uniformly for any $\boldsymbol{h}$ over the ε_1-net $\mathcal{N}_{\boldsymbol{h}}$ of $\mathcal{B}_{\boldsymbol{h}}\left(\frac{C_5}{1-\rho}\eta r\right) \coloneqq \left\{\boldsymbol{h} : \left\|\boldsymbol{h} - \boldsymbol{h}^\star\right\|_2 \leq \frac{C_5}{1-\rho}\eta r\right\}$ and any $\boldsymbol{u}$, $\boldsymbol{v}$ over the ε_2-net $\mathcal{N}_0$ of the unit sphere $\mathcal{S}^{K-1}$. As a result, one has

$$\begin{aligned}
&\left|\sum_{j=1}^{m}\left(\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{\left|\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\right|\leq 20\sqrt{\log m}\right\}} - \boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{v}\right)\right| \\
&\qquad\leq \left|\sum_{j=1}^{m}\left(y_j - \mathbb{E}\left[y_j\right]\right)\right| + \left|\sum_{j=1}^{m}\mathbb{E}\left[\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{\left|\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\right|> 20\sqrt{\log m}\right\}}\right]\right| \\
&\qquad\leq C\frac{\mu\lambda\sqrt{K}}{\sqrt{m}}\log^3 m + \frac{\lambda}{m^{99}} \\
&\qquad\leq \frac{\lambda}{100},
\end{aligned} \tag{B.51}$$

where the penultimate inequality comes from (B.49) and (B.50). Next, let us define

$$g\left(\boldsymbol{h},\boldsymbol{u},\boldsymbol{v}\right) \coloneqq \sum_{j=1}^{m}\left(\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{h}-\boldsymbol{h}^\star\right)\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{\left|\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\right|\leq 20\sqrt{\log m}\right\}} - \boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{h}-\boldsymbol{h}^\star\right)\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{v}\right).$$

Since we can always find some $\boldsymbol{x}_0 \in \mathcal{N}_{\boldsymbol{x}}$, $\boldsymbol{u}_0$, $\boldsymbol{v}_0 \in \mathcal{N}_0$ such that $\left\|\boldsymbol{h}-\boldsymbol{h}_0\right\|_2 \leq \varepsilon_1$ and $\max\left\{\left\|\boldsymbol{u}-\boldsymbol{u}_0\right\|_2, \left\|\boldsymbol{v}-\boldsymbol{v}_0\right\|_2\right\} \leq \varepsilon_2$, this guarantees that

$$\begin{aligned}
&\left|g\left(\boldsymbol{h},\boldsymbol{u},\boldsymbol{v}\right) - g\left(\boldsymbol{h}_0,\boldsymbol{u}_0,\boldsymbol{v}_0\right)\right| \\
&\quad\leq \left|\sum_{j=1}^{m}\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{h}-\boldsymbol{h}_0\right)\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{\left|\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\right|\leq 20\sqrt{\log m}\right\}}\right| + \left|\sum_{j=1}^{m}\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{h}-\boldsymbol{h}_0\right)\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{v}\right| \\
&\qquad+ \left|\sum_{j=1}^{m}\left(\left(\boldsymbol{u}-\boldsymbol{u}_0\right)^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{h}-\boldsymbol{h}^\star\right)\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{\left|\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\right|\leq 20\sqrt{\log m}\right\}} - \left(\boldsymbol{u}-\boldsymbol{u}_0\right)^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{h}-\boldsymbol{h}^\star\right)\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{v}\right)\right| \\
&\qquad+ \left|\sum_{j=1}^{m}\left(\boldsymbol{u}_0^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{h}-\boldsymbol{h}^\star\right)\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\left(\boldsymbol{v}-\boldsymbol{v}_0\right)\mathbb{1}_{\left\{\left|\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\right|\leq 20\sqrt{\log m}\right\}} - \boldsymbol{u}_0^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{h}-\boldsymbol{h}^\star\right)\boldsymbol{x}^{\star\mathsf{H}}\left(\boldsymbol{v}-\boldsymbol{v}_0\right)\right)\right| \\
&\quad\leq \left(\left\|\mathcal{A}\right\|^2 + 1\right)\left(\left\|\boldsymbol{x}^\star\right\|_2\left\|\boldsymbol{h}-\boldsymbol{h}_0\right\|_2 + \left\|\left(\boldsymbol{h}-\boldsymbol{h}^\star\right)\right\|_2\left\|\boldsymbol{u}-\boldsymbol{u}_0\right\|_2 + \left\|\boldsymbol{h}-\boldsymbol{h}^\star\right\|_2\left\|\boldsymbol{v}-\boldsymbol{v}_0\right\|_2\right) \\
&\quad\leq \left(2K\log K + 10\log m + 1\right)\left(\varepsilon_1 + 2C_1 r\varepsilon_2\right),
\end{aligned}$$

where the last inequality comes from (B.66). Since $\mathbb{P}\left(\left|\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\right| > 20\sqrt{\log m}\right) \leq O\left(m^{-100}\right)$ (in view of (F.1)), we have, with probability exceeding $1 - O\left(m^{-10}\right)$, that

$$\begin{aligned}
\left\|\mathcal{T}^{\mathsf{debias}}\left(\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{x}^{\star\mathsf{H}}\right)\right\| &= \sup_{\boldsymbol{u},\boldsymbol{v}\in\mathcal{S}^{K-1}}\left|\sum_{j=1}^{m}\left(\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{\left|\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\right|\leq 20\sqrt{\log m}\right\}} - \boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{v}\right)\right| \\
&\leq \sup_{\boldsymbol{u},\boldsymbol{v}\in\mathcal{S}^{K-1}}\left|g\left(\boldsymbol{h},\boldsymbol{u},\boldsymbol{v}\right) - g\left(\boldsymbol{h}_0,\boldsymbol{u}_0,\boldsymbol{v}_0\right)\right| + \left|g\left(\boldsymbol{h}_0,\boldsymbol{u}_0,\boldsymbol{v}_0\right)\right| \\
&\leq \left(2K\log K + 10\log m + 1\right)\left(\varepsilon_1 + 2C_1 r\varepsilon_2\right) + \frac{\lambda}{100} \\
&\leq \frac{\lambda}{50}
\end{aligned} \tag{B.52}$$

holds uniformly over $\boldsymbol{h} \in \mathcal{B}_{\boldsymbol{h}}\left(C_1 r\right)$, where the last inequality is due to the choices $\varepsilon_1 = r/\left(m\log m\right)$, $\varepsilon_2 = 1/\left(m\log m\right)$ and $r = \lambda + \sigma\sqrt{K\log m}$.

3. Finally, we turn attention to β_3. Observe that for any fixed $\boldsymbol{h}$ and $\boldsymbol{x}$, one has

$$\mathcal{T}^{\mathsf{debias}}\left(\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\right)=\sum_{j=1}^{m}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\left(\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}-\boldsymbol{I}_K\right).$$

This indicates that for any fixed unit vectors $\boldsymbol{u}$, $\boldsymbol{v}\in\mathbb{C}^K$ we have

$$\boldsymbol{u}^{\mathsf{H}}\mathcal{T}^{\mathsf{debias}}\left(\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\right)\boldsymbol{v}=\sum_{j=1}^{m}\left(\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}-\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{v}\right),$$

which is a sum of independent variables. Letting $r\coloneqq\lambda+\sigma\sqrt{K\log m}$ and $C_4\coloneqq 10\max\{C_1,C_3,1\}$, we can demonstrate that

$$\begin{aligned}
&\sum_{j=1}^{m}\Bigg(\underbrace{\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{|\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j|\le C_4r\sqrt{\log m}\right\}}}_{=:s_j}-\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{v}\Bigg)\\
&\quad=\sum_{j=1}^{m}\left(s_j-\mathbb{E}\left[s_j\right]\right)+\sum_{j=1}^{m}\left(\mathbb{E}\left[\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{|\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j|\le C_4r\sqrt{\log m}\right\}}\right]-\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^{\star}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{v}\right)\\
&\quad=\underbrace{\sum_{j=1}^{m}\left(s_j-\mathbb{E}\left[s_j\right]\right)}_{=:\omega_5}-\underbrace{\sum_{j=1}^{m}\mathbb{E}\left[\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{|\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j|> C_4r\sqrt{\log m}\right\}}\right]}_{=:\omega_6}.
\end{aligned}$$

- With regards to ω_6, similar to (B.43) we have

$$\begin{aligned}
|\omega_6|&=\left|\sum_{j=1}^{m}\mathbb{E}\left[\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{|\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j|> C_4r\sqrt{\log m}\right\}}\right]\right|\\
&\overset{\text{(i)}}{\le}\sum_{j=1}^{m}\sqrt{\mathbb{E}\left[\left|\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\right|^2\right]\mathbb{P}\left[\mathbb{1}_{\left\{|\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j|> C_4r\sqrt{\log m}\right\}}\right]}\\
&\overset{\text{(ii)}}{\le}\sum_{j=1}^{m}\left|\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\right|\sqrt{\left(2\left|\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{v}\right|^2+\left\|\boldsymbol{\Delta}_{\boldsymbol{x}}\right\|_2^2\left\|\boldsymbol{v}\right\|_2^2\right)2\exp\left(-\frac{C_4^2r^2\log m}{2\left\|\boldsymbol{\Delta}_{\boldsymbol{x}}\right\|_2^2}\right)}\\
&\le\sum_{j=1}^{m}\left|\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\right|\sqrt{6\left\|\boldsymbol{\Delta}_{\boldsymbol{x}}\right\|_2^2\exp\left(-50\log m\right)}\\
&\le\sum_{j=1}^{m}\left\|\boldsymbol{b}_j\right\|_2\left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\right|\frac{\sqrt{6}\left\|\boldsymbol{\Delta}_{\boldsymbol{x}}\right\|_2}{m^{25}}\\
&\overset{\text{(iii)}}{\le}\frac{\lambda\left\|\boldsymbol{\Delta}_{\boldsymbol{x}}\right\|_2}{m^{24}},
\end{aligned}$$

  where (i) follows from Cauchy-Schwarz inequality, (ii) comes from the property of sub-Gaussian variable $\left|\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j\right|$ and (B.44), and (iii) is due to the fact $\|\boldsymbol{b}_j\|_2=\sqrt{K/m}$ and the assumption (B.5b).
- Regarding ω_5, we note that s_j is a sub-Gaussian random variable satisfying

$$\left\|s_j-\mathbb{E}\left[s_j\right]\right\|_{\psi_2}\lesssim C_4r\sqrt{\log m}\left|\left(\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\right)\left(\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\right)\right|\le C_4\frac{\mu\sqrt{\log^5 m}}{\sqrt{m}}r\left|\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\right|.$$

  Therefore, invoking Hoeffding's inequality (cf. Vershynin [2018, Theorem 2.6.3]) reveals that

$$\mathbb{P}\left(\left|\sum_{j=1}^{m}s_j-\mathbb{E}\left[s_j\right]\right|\ge t\right)\le 2\exp\left(-\frac{ct^2}{\frac{C_4^2\mu^2r^2\log^5 m}{m}\sum_{j=1}^{m}\left|\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\right|^2}\right)=2\exp\left(-\frac{ct^2}{\frac{C_4^2\mu^2r^2\log^5 m}{m}}\right)$$

for any $t \geq 0$. Setting $t = \frac{C\mu r\sqrt{K}\log^3 m}{\sqrt{m}}$ for some sufficiently large constant $C > 0$, we obtain

$$\mathbb{P}\left(\left|\sum_{j=1}^{m} s_j - \mathbb{E}\left[s_j\right]\right| \geq \frac{C\mu r\sqrt{K}\log^3 m}{\sqrt{m}}\right) \leq 2\exp\left(-10K\log m\right). \tag{B.53}$$

Let $\varepsilon_1 = r/(m\log m)$ and $\varepsilon_2 = 1/(m\log m)$, and set $\mathcal{N}_{\boldsymbol{h}}$ to be an ε_1-net of $\mathcal{B}_{\boldsymbol{h}}\left(\frac{C_5}{1-\rho}\eta r\right) := \left\{\boldsymbol{h} : \|\boldsymbol{h} - \boldsymbol{h}^\star\|_2 \leq \frac{C_5}{1-\rho}\eta r\right\}$, $\mathcal{N}_{\boldsymbol{x}}$ an ε_1-net of $\mathcal{B}_{\boldsymbol{x}}\left(\frac{C_5}{1-\rho}\eta r\right) := \left\{\boldsymbol{x} : \|\boldsymbol{x} - \boldsymbol{x}^\star\|_2 \leq \frac{C_5}{1-\rho}\eta r\right\}$, and $\mathcal{N}_0$ an ε_2-net of the unit sphere $\mathcal{S}^{K-1} = \left\{\boldsymbol{u} \in \mathbb{C}^K : \|\boldsymbol{u}\|_2 = 1\right\}$. In view of Vershynin [2018, Corollary 4.2.13], these epsilon nets can be chosen to satisfy the following cardinality bounds

$$|\mathcal{N}_{\boldsymbol{h}}| \leq \left(1 + \frac{2C_5\eta r}{(1-\rho)\,\varepsilon_1}\right)^{2K}, \quad |\mathcal{N}_{\boldsymbol{x}}| \leq \left(1 + \frac{2C_5\eta r}{(1-\rho)\,\varepsilon_1}\right)^{2K} \quad \text{and} \quad |\mathcal{N}_0| \leq \left(1 + \frac{2}{\varepsilon_2}\right)^{2K}.$$

By taking the union bound, we show that with probability at least

$$1 - \left(1 + \frac{2C_5\eta r}{(1-\rho)\,\varepsilon_1}\right)^{4K}\left(1 + \frac{2}{\varepsilon_2}\right)^{4K} e^{-10K\log m} \geq 1 - O\left(m^{-100}\right),$$

the following bound

$$\left|\sum_{j=1}^{m} s_j - \mathbb{E}\left[s_j\right]\right| \geq \frac{C\mu r\sqrt{K}\log^3 m}{\sqrt{m}}$$

holds uniformly for any $\boldsymbol{h}$ over $\mathcal{N}_{\boldsymbol{h}}$, any $\boldsymbol{x}$ over $\mathcal{N}_{\boldsymbol{x}}$, and any $\boldsymbol{u}$, $\boldsymbol{v}$ over $\mathcal{N}_0$. Consequently, with probability exceeding $1 - O(m^{-100})$, the inequality

$$\begin{aligned}&\left|\sum_{j=1}^{m}\left(\underbrace{\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{|\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j| \leq C_4 r\sqrt{\log m}\right\}}}_{=:s_j} - \boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{v}\right)\right| \\ &\quad \leq \frac{C\mu r\sqrt{K}\log^3 m}{\sqrt{m}} + \frac{\lambda\|\boldsymbol{\Delta}_{\boldsymbol{x}}\|_2}{m^{24}} \leq \frac{\lambda}{100}\end{aligned} \tag{B.54}$$

holds simultaneously for any $\boldsymbol{h}$ over $\mathcal{N}_{\boldsymbol{h}}$, any $\boldsymbol{x}$ over $\mathcal{N}_{\boldsymbol{x}}$, and any $\boldsymbol{u}$, $\boldsymbol{v}$ over $\mathcal{N}_0$. Additionally, for any $\boldsymbol{x}$ obeying $\max_{1\leq j\leq m}\left|(\boldsymbol{x} - \boldsymbol{x}^\star)^{\mathsf{H}}\boldsymbol{a}_j\right| \leq C_3 r\sqrt{\log m}$ and any $\boldsymbol{u}$, $\boldsymbol{v} \in \mathcal{S}^{K-1}$, we can find $\boldsymbol{h}_0 \in \mathcal{N}_{\boldsymbol{h}}$, $\boldsymbol{x}_0 \in \mathcal{N}_{\boldsymbol{x}}$, $\boldsymbol{u}_0 \in \mathcal{N}_0$ and $\boldsymbol{v}_0 \in \mathcal{N}_0$ satisfying $\max\left\{\|\boldsymbol{h} - \boldsymbol{h}_0\|_2, \|\boldsymbol{x} - \boldsymbol{x}_0\|_2\right\} \leq \varepsilon_1$ and $\max\left\{\|\boldsymbol{u} - \boldsymbol{u}_0\|_2, \|\boldsymbol{v} - \boldsymbol{v}_0\|_2\right\} \leq \varepsilon_2$. Recognizing that $\|\boldsymbol{a}_j\|_2 \leq 10\sqrt{K}$ with probability $1 - O\left(me^{-CK}\right)$ for some constant $C > 0$ (see (F.2)), we can guarantee that

$$\left|\boldsymbol{\Delta}_{\boldsymbol{x}_0}^{\mathsf{H}}\boldsymbol{a}_j\right| \leq \left|\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j\right| + 10\varepsilon_1\sqrt{K} \leq 2C_3\left(\lambda + \sigma\sqrt{K\log m}\right)\sqrt{\log m}.$$

Recalling that $C_4 \geq 10C_3$, we have

$$\left|\boldsymbol{\Delta}_{\boldsymbol{x}_0}^{\mathsf{H}}\boldsymbol{a}_j\right| \leq C_4\left(\lambda + \sigma\sqrt{K\log m}\right)\sqrt{\log m} = C_4 r\sqrt{\log m},$$

and hence $\mathbb{1}_{\left\{|\boldsymbol{\Delta}_{\boldsymbol{x}_0}^{\mathsf{H}}\boldsymbol{a}_j| \leq C_4 r\sqrt{\log m}\right\}} = 1$ for all $1 \leq j \leq m$. Therefore, if we take

$$r\left(\boldsymbol{h}, \boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}\right) := \sum_{j=1}^{m}\left(\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{h} - \boldsymbol{h}^\star\right)\left(\boldsymbol{x} - \boldsymbol{x}^\star\right)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{|\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j| \leq C_4 r\sqrt{\log m}\right\}} - \boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{h} - \boldsymbol{h}^\star\right)\left(\boldsymbol{x} - \boldsymbol{x}^\star\right)^{\mathsf{H}}\boldsymbol{v}\right),$$

then it follows that

$$\left|r\left(\boldsymbol{h}, \boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}\right) - r\left(\boldsymbol{h}_0, \boldsymbol{x}_0, \boldsymbol{u}_0, \boldsymbol{v}_0\right)\right|$$

$$
\begin{aligned}
&\leq \left|\sum_{j=1}^{m} \boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{h}-\boldsymbol{h}_0\right)\left(\boldsymbol{x}-\boldsymbol{x}^\star\right)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\right| + \left|\sum_{j=1}^{m} \boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{h}-\boldsymbol{h}_0\right)\left(\boldsymbol{x}-\boldsymbol{x}_0\right)^{\mathsf{H}}\boldsymbol{v}\right| \\
&\quad + \left|\sum_{j=1}^{m} \boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{h}_0-\boldsymbol{h}^\star\right)\left(\boldsymbol{x}-\boldsymbol{x}_0\right)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\right| + \left|\sum_{j=1}^{m} \boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{h}_0-\boldsymbol{h}^\star\right)\left(\boldsymbol{x}-\boldsymbol{x}_0\right)^{\mathsf{H}}\boldsymbol{v}\right| \\
&\quad + \left|\sum_{j=1}^{m} \Big(\left(\boldsymbol{u}-\boldsymbol{u}_0\right)^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{h}_0-\boldsymbol{h}^\star\right)\left(\boldsymbol{x}_0-\boldsymbol{x}^\star\right)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v} - \left(\boldsymbol{u}-\boldsymbol{u}_0\right)^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{h}_0-\boldsymbol{h}^\star\right)\left(\boldsymbol{x}_0-\boldsymbol{x}^\star\right)^{\mathsf{H}}\boldsymbol{v}\Big)\right| \\
&\quad + \left|\sum_{j=1}^{m} \Big(\boldsymbol{u}_0^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{h}_0-\boldsymbol{h}^\star\right)\left(\boldsymbol{x}_0-\boldsymbol{x}^\star\right)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\left(\boldsymbol{v}-\boldsymbol{v}_0\right) - \boldsymbol{u}_0^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{h}_0-\boldsymbol{h}^\star\right)\left(\boldsymbol{x}_0-\boldsymbol{x}^\star\right)^{\mathsf{H}}\left(\boldsymbol{v}-\boldsymbol{v}_0\right)\Big)\right| \\
&\leq \left(\|\mathcal{A}\|^2+1\right)\|\boldsymbol{h}-\boldsymbol{h}_0\|_2\|\boldsymbol{x}-\boldsymbol{x}_0\|_2 + \left(\|\mathcal{A}\|^2+1\right)\|\boldsymbol{h}_0-\boldsymbol{h}^\star\|_2\|\boldsymbol{x}-\boldsymbol{x}_0\|_2 \\
&\quad + \left(\|\mathcal{A}\|^2+1\right)\|\boldsymbol{x}_0-\boldsymbol{x}^\star\|_2\|\boldsymbol{u}-\boldsymbol{u}_0\|_2 + \left(\|\mathcal{A}\|^2+1\right)\|\boldsymbol{x}_0-\boldsymbol{x}^\star\|_2\|\boldsymbol{v}-\boldsymbol{v}_0\|_2 \\
&\leq \left(2K\log K + 10\log m + 1\right)\left(2\left(\varepsilon_1\right)^2 + 2C_1 r\varepsilon_2\right), \qquad \text{(B.55)}
\end{aligned}
$$

where the last inequality arises from (B.66). This further leads to

$$
\begin{aligned}
&\left|\boldsymbol{u}^{\mathsf{H}}\mathcal{T}^{\mathsf{debias}}\left(\left(\boldsymbol{h}-\boldsymbol{h}^\star\right)\left(\boldsymbol{x}-\boldsymbol{x}^\star\right)^{\mathsf{H}}\right)\boldsymbol{v}\right| \\
&\quad = \left|\sum_{j=1}^{m} \Big(\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{h}-\boldsymbol{h}^\star\right)\left(\boldsymbol{x}-\boldsymbol{x}^\star\right)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{|\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j|\leq C_4 r\sqrt{\log m}\right\}} - \boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{h}-\boldsymbol{h}^\star\right)\left(\boldsymbol{x}-\boldsymbol{x}^\star\right)^{\mathsf{H}}\boldsymbol{v}\Big)\right| \\
&\quad = \left|r\left(\boldsymbol{h},\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}\right) - r\left(\boldsymbol{h}_0,\boldsymbol{x}_0,\boldsymbol{u}_0,\boldsymbol{v}_0\right)\right| + \left|r\left(\boldsymbol{h}_0,\boldsymbol{x}_0,\boldsymbol{u}_0,\boldsymbol{v}_0\right)\right| \\
&\quad \leq \left(2K\log K + 10\log m + 1\right)\left(2\left(\varepsilon_1\right)^2 + 2C_1 r\varepsilon_2\right) + \frac{\lambda}{100} \\
&\quad \leq \frac{\lambda}{50},
\end{aligned}
$$

where the last inequality follows from (B.54) and (B.55). As a consequence, for any point $(\boldsymbol{h},\boldsymbol{x})$ satisfying (B.5), we have, with probability exceeding $1-O\left(m^{-10}+me^{-CK}\right)$, that

$$
\left\|\mathcal{T}^{\mathsf{debias}}\left(\left(\boldsymbol{h}-\boldsymbol{h}^\star\right)\left(\boldsymbol{x}-\boldsymbol{x}^\star\right)^{\mathsf{H}}\right)\right\| = \sup_{\boldsymbol{u},\boldsymbol{v}\in\mathcal{S}^{K-1}} \boldsymbol{u}^{\mathsf{H}}\mathcal{T}^{\mathsf{debias}}\left(\left(\boldsymbol{h}-\boldsymbol{h}^\star\right)\left(\boldsymbol{x}-\boldsymbol{x}^\star\right)^{\mathsf{H}}\right)\boldsymbol{v} \leq \frac{1}{50}\lambda. \qquad \text{(B.56)}
$$

To finish up, combining the bounds obtained in (B.47), (B.52) and (B.56), we arrive at

$$
\left\|\mathcal{T}^{\mathsf{debias}}\left(\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right)\right\| \leq \frac{\lambda}{50} + \frac{\lambda}{50} + \frac{\lambda}{50} < \frac{\lambda}{8}.
$$

B.6 Proof of Lemma 14

We intend to invoke Koltchinskii et al. [2011, Proposition 2] to bound the spectral norm of the random matrix of interest. Set $\boldsymbol{Z}_i = \xi_i\boldsymbol{b}_i\boldsymbol{a}_i^{\mathsf{H}}$. Letting $\|\cdot\|_{\psi_1}$ (resp. $\|\cdot\|_{\psi_2}$) denoting the sub-exponential norm of a random variable Vershynin [2018, Chapter 2], we have

$$
B_{\boldsymbol{Z}} := \left\|\left\|\xi_j\boldsymbol{b}_j\boldsymbol{a}_j^{\mathsf{H}}\right\|\right\|_{\psi_1} = \left\|\xi_j\|\boldsymbol{b}_j\|_2\|\boldsymbol{a}_j\|_2\right\|_{\psi_1} \leq \|\xi_j\|_{\psi_2}\left\|\|\boldsymbol{a}_j\|_2\right\|_{\psi_2}\sqrt{\frac{K}{m}} \lesssim \sigma\frac{K}{\sqrt{m}}.
$$

Here, we have used the assumption that $\|\xi_j\|_{\psi_2} \lesssim \sigma$, as well as the simple facts that $\|\boldsymbol{b}_j\|_2 = \sqrt{K/m}$ and $\left\|\|\boldsymbol{a}_j\|_2\right\|_{\psi_2} \lesssim \sqrt{K}$ (cf. Vershynin [2018, Theorem 3.1.1]). In addition, simple calculation yields

$$
\left\|\sum_{j=1}^{m}\mathbb{E}\left[\boldsymbol{Z}_j\boldsymbol{Z}_j^{\mathsf{H}}\right]\right\| = \left\|\sum_{j=1}^{m}\mathbb{E}\Big[\left|\xi_j\right|^2\boldsymbol{b}_i\boldsymbol{a}_i^{\mathsf{H}}\boldsymbol{a}_i\boldsymbol{b}_i^{\mathsf{H}}\Big]\right\| = \left\|\sum_{j=1}^{m}\mathbb{E}\left[|\xi_j|^2\right]\mathbb{E}\left[\|\boldsymbol{a}_j\|_2^2\right]\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\right\| \asymp K\sigma^2,
$$

$$\left\|\sum_{j=1}^{m}\mathbb{E}\big[\boldsymbol{Z}_j^{\mathsf{H}}\boldsymbol{Z}_j\big]\right\| = \left\|\sum_{j=1}^{m}\mathbb{E}\Big[\,|\xi_j|^2\,\boldsymbol{a}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{a}_j^{\mathsf{H}}\Big]\right\| = \left\|\sum_{j=1}^{m}\mathbb{E}\big[|\xi_j|^2\big]\,\|\boldsymbol{b}_j\|_2^2\,\mathbb{E}\big[\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\big]\right\| \asymp K\sigma^2,$$

which rely on the facts that $\mathbb{E}\big[|\xi_j|^2\big] \asymp \sigma^2$, $\|\boldsymbol{b}_j\|_2 = \sqrt{K/m}$, $\sum_{j=1}^{m}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}} = \boldsymbol{I}_k$ and $\mathbb{E}\big[\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\big] = \boldsymbol{I}_k$. As a result, by setting

$$\sigma_{\boldsymbol{Z}} := \max\left\{\left\|\sum\nolimits_{j=1}^{m}\mathbb{E}\big[\boldsymbol{Z}_j\boldsymbol{Z}_j^{\mathsf{H}}\big]\right\|^{1/2}, \left\|\sum\nolimits_{j=1}^{m}\mathbb{E}\big[\boldsymbol{Z}_j^{\mathsf{H}}\boldsymbol{Z}_j\big]\right\|^{1/2}\right\} \asymp \sigma\sqrt{K},$$

we can apply the matrix Bernstein inequality Koltchinskii et al. [2011, Proposition 2] to derive

$$\left\|\sum_{j=1}^{m}\xi_j\boldsymbol{b}_j\boldsymbol{a}_j^{\mathsf{H}}\right\| \lesssim \sigma_{\boldsymbol{Z}}\sqrt{\log m} + B_{\boldsymbol{Z}}\log\left(\frac{B_{\boldsymbol{Z}}}{\sigma_{\boldsymbol{Z}}}\right)\log m \lesssim \sigma\sqrt{K\log m} \tag{B.57}$$

with probability exceeding $1 - O(m^{-20})$, where the last inequality holds as long as $m \gtrsim K\log^3 m$.

B.7 Proof of Lemma 15

By the definition of T (cf. (B.28)), any $\boldsymbol{Z} \in T$ takes the following form

$$\boldsymbol{Z} = \boldsymbol{h}\boldsymbol{u}^{\mathsf{H}} + \boldsymbol{v}\boldsymbol{x}^{\mathsf{H}}$$

for some $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{C}^K$. Since this is an underdetermined system of equations, there might exist more than one possibilities of $(\boldsymbol{h}, \boldsymbol{x})$ that enable and are compatible with this decomposition. Here, we shall take a specific choice among them as follows

$$(\boldsymbol{h}, \boldsymbol{x}) := \arg\min_{(\widetilde{\boldsymbol{h}}, \widetilde{\boldsymbol{x}})}\left\{\frac{1}{2}\|\widetilde{\boldsymbol{h}}\|_2^2 + \frac{1}{2}\|\widetilde{\boldsymbol{x}}\|_2^2 \mid \boldsymbol{Z} = \widetilde{\boldsymbol{h}}\boldsymbol{u}^{\mathsf{H}} + \boldsymbol{v}\widetilde{\boldsymbol{x}}^{\mathsf{H}} \text{ for some } \boldsymbol{u} \text{ and } \boldsymbol{v}\right\}. \tag{B.58}$$

As can be straightforwardly verified, this special choice enjoys the following property

$$\boldsymbol{h}^{\mathsf{H}}\boldsymbol{v} = \boldsymbol{u}^{\mathsf{H}}\boldsymbol{x},$$

which plays a crucial role in the proof.

The proof consists of two steps: (1) showing that

$$\|\boldsymbol{Z}\|_{\mathrm{F}}^2 \le 8\left(\|\boldsymbol{u}\|_2^2 + \|\boldsymbol{v}\|_2^2\right), \tag{B.59}$$

and (2) demonstrating that

$$\|\mathcal{A}(\boldsymbol{Z})\|_2^2 \ge \frac{1}{2}\left(\|\boldsymbol{u}\|_2^2 + \|\boldsymbol{v}\|_2^2\right). \tag{B.60}$$

The first claim (B.59) can be justified in the same way as Chen et al. [2020, Equation (81)]; we thus omit this part here for brevity.

It then boils down to justifying the second claim (B.60), towards which we first decompose

$$\|\mathcal{A}(\boldsymbol{Z})\|_2^2 = \underbrace{\|\mathcal{A}(\boldsymbol{Z})\|_2^2 - \|\boldsymbol{Z}\|_2^2}_{=:\alpha_1} + \underbrace{\|\boldsymbol{Z}\|_2^2}_{=:\alpha_2}. \tag{B.61}$$

By repeating the same argument as in Chen et al. [2020, Appendix C.3.1, 2(a)], we can lower bound α_2 by

$$\alpha_2 \ge \left\|\boldsymbol{h}^\star\boldsymbol{u}^{\mathsf{H}}\right\|_{\mathrm{F}}^2 + \left\|\boldsymbol{v}\boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}}^2 - \frac{1}{50}\left(\|\boldsymbol{u}\|_2^2 + \|\boldsymbol{v}\|_2^2\right).$$

We then turn attention to controlling α_1. Letting $\boldsymbol{\Delta}_{\boldsymbol{h}} = \boldsymbol{h} - \boldsymbol{h}^\star$ and $\boldsymbol{\Delta}_{\boldsymbol{x}} = \boldsymbol{x} - \boldsymbol{x}^\star$, we can write

$$\boldsymbol{h}\boldsymbol{u}^{\mathsf{H}} + \boldsymbol{v}\boldsymbol{x}^{\mathsf{H}} = (\boldsymbol{h}^\star + \boldsymbol{\Delta}_{\boldsymbol{h}})\,\boldsymbol{u}^{\mathsf{H}} + \boldsymbol{v}\,(\boldsymbol{x}^\star + \boldsymbol{\Delta}_{\boldsymbol{x}})^{\mathsf{H}}$$

$$= \boldsymbol{h}^{\star}\boldsymbol{u}^{\mathsf{H}} + \boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{u}^{\mathsf{H}} + \boldsymbol{v}\boldsymbol{x}^{\star\mathsf{H}} + \boldsymbol{v}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}.$$

This implies that α_1 can be expanded as follows

$$\alpha_1 = \underbrace{\left\|\mathcal{A}\left(\boldsymbol{h}^{\star}\boldsymbol{u}^{\mathsf{H}} + \boldsymbol{v}\boldsymbol{x}^{\star\mathsf{H}}\right)\right\|_2^2 - \left\|\boldsymbol{h}^{\star}\boldsymbol{u}^{\mathsf{H}} + \boldsymbol{v}\boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}}^2}_{=:\gamma_1} + \underbrace{\left\|\mathcal{A}\left(\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{u}^{\mathsf{H}} + \boldsymbol{v}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\right)\right\|_2^2 - \left\|\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{u}^{\mathsf{H}} + \boldsymbol{v}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\right\|_{\mathrm{F}}^2}_{=:\gamma_2}$$
$$+ 2\underbrace{\left\langle \mathcal{A}\left(\boldsymbol{h}^{\star}\boldsymbol{u}^{\mathsf{H}} + \boldsymbol{v}\boldsymbol{x}^{\star\mathsf{H}}\right), \mathcal{A}\left(\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{u}^{\mathsf{H}} + \boldsymbol{v}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\right)\right\rangle - \left\langle \boldsymbol{h}^{\star}\boldsymbol{u}^{\mathsf{H}} + \boldsymbol{v}\boldsymbol{x}^{\star\mathsf{H}}, \boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{u}^{\mathsf{H}} + \boldsymbol{v}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\right\rangle}_{=:\gamma_3},$$

thereby motivating us to cope with these terms separately.

- Regarding γ_1, it is easily seen that

$$|\gamma_1| \le \left\|\mathcal{P}_T\mathcal{A}^*\mathcal{A}\mathcal{P}_T - \mathcal{P}_T\right\| \cdot \left\|\boldsymbol{h}^{\star}\boldsymbol{u}^{\mathsf{H}} + \boldsymbol{v}\boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}}^2 \le \frac{1}{100}\left(\|\boldsymbol{u}\|^2 + \|\boldsymbol{v}\|^2\right),$$

  where the last inequality is obtained by invoking Li et al. [2019, Lemma 5.12].

- When it comes to γ_2, we observe that

$$\gamma_2 \ge -\left\|\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{u}^{\mathsf{H}} + \boldsymbol{v}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\right\|_{\mathrm{F}}^2 \ge -\frac{1}{100}\left(\|\boldsymbol{u}\|_2^2 + \|\boldsymbol{v}\|_2^2\right)$$

  under our constraints on the sizes of $\boldsymbol{\Delta}_{\boldsymbol{h}}$ and $\boldsymbol{\Delta}_{\boldsymbol{x}}$.

- The term γ_3 can be further decomposed into four terms, which we control separately.

  1. First of all, observe that

$$\begin{aligned}
&\left|\left\langle \mathcal{A}\left(\boldsymbol{v}\boldsymbol{x}^{\star\mathsf{H}}\right), \mathcal{A}\left(\boldsymbol{v}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\right)\right\rangle - \left\langle \boldsymbol{v}\boldsymbol{x}^{\star\mathsf{H}}, \boldsymbol{v}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\right\rangle\right| \\
&\quad\le \left|\left\langle \mathcal{A}\left(\boldsymbol{v}\boldsymbol{x}^{\star\mathsf{H}}\right), \mathcal{A}\left(\boldsymbol{v}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\right)\right\rangle\right| + \left|\left\langle \boldsymbol{v}\boldsymbol{x}^{\star\mathsf{H}}, \boldsymbol{v}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\right\rangle\right| \\
&\quad\overset{\text{(i)}}{\le} \left\|\mathcal{A}\left(\boldsymbol{v}\boldsymbol{x}^{\star\mathsf{H}}\right)\right\|_2 \left\|\mathcal{A}\left(\boldsymbol{v}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\right)\right\|_2 + \|\boldsymbol{x}^{\star}\|_2 \left\|\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\right\|_2 \|\boldsymbol{v}\|_2^2 \\
&\quad\overset{\text{(ii)}}{\le} \sqrt{\sum_{j=1}^m \left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{v}\right|^2 \left|\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\right|^2}\sqrt{\sum_{j=1}^m \left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{v}\right|^2 \left|\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j\right|^2} + \frac{1}{200}\|\boldsymbol{v}\|_2^2 \\
&\quad\overset{\text{(iii)}}{\le} \sqrt{\|\boldsymbol{v}\|_2^2 \max_{1\le j\le m}\left|\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\right|^2} \cdot \sqrt{\|\boldsymbol{v}\|_2^2 \max_{1\le j\le m}\left|\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j\right|^2} + \frac{1}{200}\|\boldsymbol{v}\|_2^2 \\
&\quad\overset{\text{(iv)}}{\le} 20\sqrt{\log m} \cdot C\sqrt{\log m}\left(\lambda + \sigma\sqrt{K\log m}\right)\|\boldsymbol{v}\|_2^2 + \frac{1}{200}\|\boldsymbol{v}\|_2^2 \\
&\quad\le \frac{1}{100}\|\boldsymbol{v}\|_2^2,
\end{aligned} \tag{B.62}$$

     where the (i) and (ii) follow from the Cauchy-Schwarz inequality and (B.5a) that $\left\|\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\right\|_2 \lesssim \lambda + \sigma\sqrt{K\log m} \le 1/200$; (iii) comes from the fact that $\sum_{j=1}^m \boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}} = \boldsymbol{I}_K$ and thus $\sum_{j=1}^m \left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{v}\right|^2 = \sum_{j=1}^m \boldsymbol{v}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{v} = \boldsymbol{v}^{\mathsf{H}}\boldsymbol{v} = \|\boldsymbol{v}\|_2^2$; (iv) is due to Lemma 38 and (B.5b); and the last inequality holds true as long as $\sigma\sqrt{K\log^3 m} \ll 1$.

  2. Similarly, we can demonstrate that

$$\begin{aligned}
&\left|\left\langle \mathcal{A}\left(\boldsymbol{h}^{\star}\boldsymbol{u}^{\mathsf{H}}\right), \mathcal{A}\left(\boldsymbol{v}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\right)\right\rangle - \left\langle \boldsymbol{h}^{\star}\boldsymbol{u}^{\mathsf{H}}, \boldsymbol{v}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\right\rangle\right| \\
&\quad\overset{\text{(i)}}{\le} \sqrt{\sum_{j=1}^m \left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^{\star}\right|^2 \left|\boldsymbol{u}^{\mathsf{H}}\boldsymbol{a}_j\right|^2}\sqrt{\sum_{j=1}^m \left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{v}\right|^2 \left|\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j\right|^2} + \frac{1}{200}\|\boldsymbol{u}\|_2\|\boldsymbol{v}\|_2 \\
&\quad\overset{\text{(ii)}}{\le} \sqrt{\sum_{j=1}^m \left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^{\star}\right|^2 \left|\boldsymbol{u}^{\mathsf{H}}\boldsymbol{a}_j\right|^2} \cdot C\sqrt{\log m}\left(\lambda + \sigma\sqrt{K\log m}\right)\|\boldsymbol{v}\|_2 + \frac{1}{200}\|\boldsymbol{u}\|_2\|\boldsymbol{v}\|_2
\end{aligned}$$

$$\leq \frac{1}{100}\left\|\boldsymbol{u}\right\|_2\left\|\boldsymbol{v}\right\|_2,$$

where (i) holds for the same reason as Step (ii) in (B.62); (ii) arises due to the identity $\sum_{j=1}^m \left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{v}\right|^2 = \left\|\boldsymbol{v}\right\|_2^2$ and (B.5b); and the last inequality relies on the following claim.

Claim 4. With probability exceeding $1 - O\left(m^{-100}\right)$, the following inequality

$$\left|\sum_{j=1}^m \left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\right|^2\left|\boldsymbol{u}^{\mathsf{H}}\boldsymbol{a}_j\right|^2 - \left\|\boldsymbol{u}\right\|_2^2\right| \lesssim \sqrt{\frac{\mu^2 K\log m}{m}}\left\|\boldsymbol{u}\right\|_2^2 \tag{B.63}$$

holds uniformly for any $\boldsymbol{u}$.

Proof. See Appendix B.7.1. □

3. The next term we shall control is

$$\left\langle \mathcal{A}\left(\boldsymbol{h}^\star\boldsymbol{u}^{\mathsf{H}}\right), \mathcal{A}\left(\boldsymbol{\Delta_h}\boldsymbol{u}^{\mathsf{H}}\right)\right\rangle - \left\langle \boldsymbol{h}^\star\boldsymbol{u}^{\mathsf{H}}, \boldsymbol{\Delta_h}\boldsymbol{u}^{\mathsf{H}}\right\rangle = \sum_{j=1}^m \left(\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\right)\left(\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta_h}\right)\left(\left|\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{u}\right|^2 - \left\|\boldsymbol{u}\right\|_2^2\right).$$

By virtue of the Bernstein inequality Vershynin [2018, Theorem 2.8.2], we have

$$\begin{aligned}&\mathbb{P}\left(\left|\left\langle \mathcal{A}\left(\boldsymbol{h}^\star\boldsymbol{u}^{\mathsf{H}}\right), \mathcal{A}\left(\boldsymbol{\Delta_h}\boldsymbol{u}^{\mathsf{H}}\right)\right\rangle - \left\langle \boldsymbol{h}^\star\boldsymbol{u}^{\mathsf{H}}, \boldsymbol{\Delta_h}\boldsymbol{u}^{\mathsf{H}}\right\rangle\right| \geq \tau\left\|\boldsymbol{u}\right\|_2^2\right)\\ &\quad\leq 2\max\left\{\exp\left(-\frac{\tau^2}{4\left\|\boldsymbol{B}\boldsymbol{\Delta_h}\right\|_\infty^2}\right), \exp\left(-\frac{\tau}{4\left\|\boldsymbol{B}\boldsymbol{\Delta_h}\right\|_\infty\left\|\boldsymbol{B}\boldsymbol{h}^\star\right\|_\infty}\right)\right\}\end{aligned}$$

for any $\tau \geq 0$. Let us choose τ to be

$$\tau = 2\left\|\boldsymbol{B}\boldsymbol{\Delta_h}\right\|_\infty\sqrt{2K\log m} + 8\left\|\boldsymbol{B}\boldsymbol{\Delta_h}\right\|_\infty\left\|\boldsymbol{B}\boldsymbol{h}^\star\right\|_\infty K\log m.$$

In view of (B.5b) and (7), this quantity is bounded above by

$$\tau \lesssim 2\sigma\sqrt{2K\log m} + 8\sigma\frac{\mu}{\sqrt{m}}K\log m \leq \frac{1}{100}.$$

It then follows that

$$\mathbb{P}\left(\left|\left\langle \mathcal{A}\left(\boldsymbol{h}^\star\boldsymbol{u}^{\mathsf{H}}\right), \mathcal{A}\left(\boldsymbol{\Delta_h}\boldsymbol{u}^{\mathsf{H}}\right)\right\rangle - \left\langle \boldsymbol{h}^\star\boldsymbol{u}^{\mathsf{H}}, \boldsymbol{\Delta_h}\boldsymbol{u}^{\mathsf{H}}\right\rangle\right| \geq \frac{1}{100}\left\|\boldsymbol{u}\right\|_2^2\right) \leq 2\exp\left(-2K\log m\right). \tag{B.64}$$

Additionally, define $r := \lambda + \sigma\sqrt{K\log m}$, and let $\mathcal{N}_{\boldsymbol{h}}$ be an ε_1-net of $\mathcal{B}_{\boldsymbol{h}}\left(\frac{C_5}{1-\rho}\eta r\right) := \left\{\boldsymbol{h} : \left\|\boldsymbol{h} - \boldsymbol{h}^\star\right\|_2 \leq \frac{C_5}{1-\rho}\eta r\right\}$ and $\mathcal{N}_0$ an ε_2-net of the unit sphere $\mathcal{S}^{K-1} = \left\{\boldsymbol{u} \in \mathbb{C}^K : \left\|\boldsymbol{u}\right\|_2 = 1\right\}$. Let $\varepsilon_1 = r/\left(m\log m\right)$ and $\varepsilon_2 = 1/\left(m\log m\right)$. In view of Vershynin [2018, Corollary 4.2.13], it is seen that

$$\left|\mathcal{N}_{\boldsymbol{h}}\right| \leq \left(1 + \frac{2C_5\eta r}{\left(1-\rho\right)\varepsilon_1}\right)^{2K} \quad \text{and} \quad \left|\mathcal{N}_0\right| \leq \left(1 + \frac{2}{\varepsilon_2}\right)^{2K}.$$

Taking the union bound indicates that with probability at least

$$1 - \left(1 + \frac{2C_5\eta r}{\left(1-\rho\right)\varepsilon_1}\right)^{2K}\left(1 + \frac{2}{\varepsilon_2}\right)^{4K}\cdot 2e^{-2K\log m} \geq 1 - O\left(m^{-100}\right),$$

the following inequality

$$\left|\left\langle \mathcal{A}\left(\boldsymbol{h}^\star\boldsymbol{u}^{\mathsf{H}}\right), \mathcal{A}\left(\boldsymbol{\Delta_h}\boldsymbol{u}^{\mathsf{H}}\right)\right\rangle - \left\langle \boldsymbol{h}^\star\boldsymbol{u}^{\mathsf{H}}, \boldsymbol{\Delta_h}\boldsymbol{u}^{\mathsf{H}}\right\rangle\right| \geq \frac{1}{100}\left\|\boldsymbol{u}\right\|_2^2$$

holds uniformly for all $(\boldsymbol{h},\boldsymbol{u}) \in \mathcal{N}_{\boldsymbol{h}} \times \mathcal{N}_0$. As a result, for any $(\boldsymbol{h},\boldsymbol{u}) \in \mathcal{N}_{\boldsymbol{h}} \times \mathcal{N}_0$, there holds

$$\left|\left\langle \mathcal{A}\left(\boldsymbol{h}^{\star}\boldsymbol{u}^{\mathsf{H}}\right), \mathcal{A}\left(\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{u}^{\mathsf{H}}\right)\right\rangle - \left\langle \boldsymbol{h}^{\star}\boldsymbol{u}^{\mathsf{H}}, \boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{u}^{\mathsf{H}}\right\rangle\right| \geq \frac{1}{100}\left\|\boldsymbol{u}\right\|_2^2$$

with probability exceeding $1 - O\left(m^{-100}\right)$. Furthermore, if we let

$$F\left(\boldsymbol{h},\boldsymbol{u}\right) \coloneqq \left\langle \mathcal{A}\left(\boldsymbol{h}^{\star}\boldsymbol{u}^{\mathsf{H}}\right), \mathcal{A}\left(\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{u}^{\mathsf{H}}\right)\right\rangle - \left\langle \boldsymbol{h}^{\star}\boldsymbol{u}^{\mathsf{H}}, \boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{u}^{\mathsf{H}}\right\rangle,$$

then for any $\boldsymbol{h} \in \mathcal{B}_{\boldsymbol{h}}\left(\frac{C_5}{1-\rho}\eta r\right)$ and $\boldsymbol{u} \in \mathcal{S}^{K-1}$, we can find a point $(\boldsymbol{h}_0,\boldsymbol{u}_0) \in \mathcal{N}_{\boldsymbol{h}} \times \mathcal{N}_0$ satisfying $\left\|\boldsymbol{h}-\boldsymbol{h}_0\right\|_2 \leq \varepsilon_1$ and $\left\|\boldsymbol{u}-\boldsymbol{u}_0\right\|_2 \leq \varepsilon_2$. Consequently, one can deduce that

$$\begin{aligned}
&\left|F\left(\boldsymbol{h},\boldsymbol{u}\right) - F\left(\boldsymbol{h}_0,\boldsymbol{u}_0\right)\right| \\
&\leq \left|\left\langle \mathcal{A}\left(\boldsymbol{h}^{\star}\left(\boldsymbol{u}-\boldsymbol{u}_0\right)^{\mathsf{H}}\right), \mathcal{A}\left(\left(\boldsymbol{h}-\boldsymbol{h}^{\star}\right)\boldsymbol{u}^{\mathsf{H}}\right)\right\rangle - \left\langle \boldsymbol{h}^{\star}\left(\boldsymbol{u}-\boldsymbol{u}_0\right)^{\mathsf{H}}, \left(\boldsymbol{h}-\boldsymbol{h}^{\star}\right)\boldsymbol{u}^{\mathsf{H}}\right\rangle\right| \\
&\quad + \left|\left\langle \mathcal{A}\left(\boldsymbol{h}^{\star}\boldsymbol{u}_0^{\mathsf{H}}\right), \mathcal{A}\left(\left(\boldsymbol{h}-\boldsymbol{h}_0\right)\boldsymbol{u}^{\mathsf{H}}\right)\right\rangle - \left\langle \boldsymbol{h}^{\star}\boldsymbol{u}_0^{\mathsf{H}}, \left(\boldsymbol{h}-\boldsymbol{h}_0\right)\boldsymbol{u}^{\mathsf{H}}\right\rangle\right| \\
&\quad + \left|\left\langle \mathcal{A}\left(\boldsymbol{h}^{\star}\boldsymbol{u}_0^{\mathsf{H}}\right), \mathcal{A}\left(\left(\boldsymbol{h}-\boldsymbol{h}_0\right)\left(\boldsymbol{u}-\boldsymbol{u}_0\right)^{\mathsf{H}}\right)\right\rangle - \left\langle \boldsymbol{h}^{\star}\boldsymbol{u}_0^{\mathsf{H}}, \left(\boldsymbol{h}-\boldsymbol{h}_0\right)\left(\boldsymbol{u}-\boldsymbol{u}_0\right)^{\mathsf{H}}\right\rangle\right| \\
&\leq \left(\left\|\mathcal{A}\right\|^2+1\right)\left\|\boldsymbol{h}^{\star}\right\|\left\|\boldsymbol{u}\right\|_2\left\|\boldsymbol{u}-\boldsymbol{u}_0\right\|_2\left\|\boldsymbol{h}-\boldsymbol{h}^{\star}\right\|_2 + \left(\left\|\mathcal{A}\right\|^2+1\right)\left\|\boldsymbol{h}^{\star}\right\|_2\left\|\boldsymbol{u}\right\|_2\left\|\boldsymbol{u}_0\right\|_2\left\|\boldsymbol{h}-\boldsymbol{h}_0\right\|_2 \\
&\quad + \left(\left\|\mathcal{A}\right\|^2+1\right)\left\|\boldsymbol{h}^{\star}\right\|_2\left\|\boldsymbol{u}_0\right\|_2\left\|\boldsymbol{u}-\boldsymbol{u}_0\right\|_2\left\|\boldsymbol{h}-\boldsymbol{h}_0\right\|_2 \\
&\leq \left(2K\log K + 10\log m + 1\right)\left(\frac{C_5}{1-\rho}\eta r\epsilon_2 + \epsilon_1 + \epsilon_1\epsilon_2\right) \\
&\leq \frac{1}{100}\left\|\boldsymbol{u}\right\|_2^2
\end{aligned}$$

as long as $m \gg K$, where the above bound on $\left\|\mathcal{A}\right\|$ relies on Lemma 1. Hence, with probability exceeding $1 - O\left(m^{-10}\right)$ we have

$$\begin{aligned}
\left|\left\langle \mathcal{A}\left(\boldsymbol{h}^{\star}\boldsymbol{u}^{\mathsf{H}}\right), \mathcal{A}\left(\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{u}^{\mathsf{H}}\right)\right\rangle - \left\langle \boldsymbol{h}^{\star}\boldsymbol{u}^{\mathsf{H}}, \boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{u}^{\mathsf{H}}\right\rangle\right| &\leq \left|F\left(\boldsymbol{h},\boldsymbol{u}\right) - F\left(\boldsymbol{h}_0,\boldsymbol{u}_0\right)\right| + \left|F\left(\boldsymbol{h}_0,\boldsymbol{u}_0\right)\right| \\
&\leq \frac{1}{100}\left\|\boldsymbol{u}\right\|_2^2 + \frac{1}{100}\left\|\boldsymbol{u}\right\|_2^2 \leq \frac{1}{50}\left\|\boldsymbol{u}\right\|_2^2,
\end{aligned}$$

which holds uniformly over all $\boldsymbol{h} \in \mathcal{B}_{\boldsymbol{h}}\left(\frac{C_5}{1-\rho}\eta r\right)$ and $\boldsymbol{u} \in \mathcal{S}^{K-1}$.

4. The bound on $\left\langle \mathcal{A}\left(\boldsymbol{v}\boldsymbol{x}^{\star\mathsf{H}}\right), \mathcal{A}\left(\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{u}^{\mathsf{H}}\right)\right\rangle - \left\langle \boldsymbol{v}\boldsymbol{x}^{\star\mathsf{H}}, \boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{u}^{\mathsf{H}}\right\rangle$ can be obtained in a similar manner; we thus omit it here for simplicity.
5. The above bounds on four terms taken collectively demonstrate that

$$\left|\gamma_3\right| \leq \frac{1}{100}\left\|\boldsymbol{v}\right\|_2^2 + \frac{1}{100}\left\|\boldsymbol{u}\right\|_2\left\|\boldsymbol{v}\right\|_2 + \frac{1}{50}\left\|\boldsymbol{u}\right\|_2^2 + \frac{1}{100}\left\|\boldsymbol{u}\right\|_2\left\|\boldsymbol{v}\right\|_2 \leq \frac{1}{25}\left(\left\|\boldsymbol{u}\right\|_2^2 + \left\|\boldsymbol{v}\right\|_2^2\right).$$

Combining the above results, we can continue the relation (B.61) to conclude that

$$\begin{aligned}
\left\|\mathcal{A}\left(\boldsymbol{Z}\right)\right\|_2^2 &= \alpha_2 + \alpha_1 \\
&\geq \left\|\boldsymbol{h}^{\star}\boldsymbol{u}^{\mathsf{H}}\right\|_{\mathrm{F}}^2 + \left\|\boldsymbol{v}\boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}}^2 - \frac{1}{50}\left(\left\|\boldsymbol{u}\right\|_2^2 + \left\|\boldsymbol{v}\right\|_2^2\right) - \left|\gamma_1\right| - \frac{1}{100}\left(\left\|\boldsymbol{u}\right\|_2^2 + \left\|\boldsymbol{v}\right\|_2^2\right) - \frac{1}{25}\left(\left\|\boldsymbol{u}\right\|_2^2 + \left\|\boldsymbol{v}\right\|_2^2\right) \\
&\geq \frac{1}{2}\left(\left\|\boldsymbol{u}\right\|_2^2 + \left\|\boldsymbol{v}\right\|_2^2\right)
\end{aligned}$$

as claimed.

B.7.1 Proof of Claim 4

We start by defining

$$\eta := \sum_{j=1}^{m} \left| \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \right|^2 \left(\left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{u} \right|^2 - \|\boldsymbol{u}\|_2^2 \right),$$

which is the sum of sub-exponential variables with zero mean $\mathbb{E}\left[\left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{u} \right|^2 - \|\boldsymbol{u}\|_2^2 \right] = 0$.

Concentration. In view of the Bernstein inequality (cf. Vershynin [2018, Theorem 2.8.2]), we have

$$\begin{aligned}
&\mathbb{P}\left(\left| \sum_{j=1}^{m} \left| \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \right|^2 \left(\left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{u} \right|^2 - \|\boldsymbol{u}\|_2^2 \right) \right| \geq \tau \|\boldsymbol{u}\|_2^2 \right) \\
&\qquad \leq 2 \max \left\{ \exp\left(- \frac{\tau^2}{4 \|\boldsymbol{B}\boldsymbol{h}^\star\|_\infty^2 \|\boldsymbol{u}\|_2^2} \right), \exp\left(- \frac{\tau}{4 \|\boldsymbol{B}\boldsymbol{\Delta}_{\boldsymbol{h}}\|_\infty^2 \|\boldsymbol{u}\|_2} \right) \right\}
\end{aligned}$$

for any $\tau \geq 0$. Set

$$\tau = 4 \|\boldsymbol{B}\boldsymbol{h}^\star\|_\infty \|\boldsymbol{u}\|_2 \sqrt{2K \log m} + 16 \|\boldsymbol{B}\boldsymbol{h}^\star\|_\infty^2 \|\boldsymbol{u}\|_2 K \log m,$$

then there holds

$$\mathbb{P}\left(\left| \sum_{j=1}^{m} \left| \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \right|^2 \left(\left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{u} \right|^2 - \|\boldsymbol{u}\|_2^2 \right) \right| \geq \tau \|\boldsymbol{u}\|_2^2 \right) \leq 2 \exp\left(-4K \log m \right). \tag{B.65}$$

Union bound. Next, define $\mathcal{N}_0$ to be an ϵ_0-net of the unit sphere $\mathcal{S}^{K-1} := \left\{ \boldsymbol{u} \in \mathbb{C}^K : \|\boldsymbol{u}\|_2 = 1 \right\}$, which can be chosen to obey Vershynin [2018, Corollary 4.2.13]

$$|\mathcal{N}_0| \leq \left(1 + \frac{2}{\epsilon_0} \right)^{2K}.$$

By taking the union bound over $\mathcal{N}_0$, we reach

$$\left| \sum_{j=1}^{m} \left| \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \right|^2 \left(\left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{u} \right|^2 - \|\boldsymbol{u}\|_2^2 \right) \right| \geq 4 \|\boldsymbol{B}\boldsymbol{h}^\star\|_\infty \sqrt{2K \log m} + 16 \|\boldsymbol{B}\boldsymbol{h}^\star\|_\infty^2 K \log m, \qquad \forall \boldsymbol{u} \in \mathcal{N}_0$$

with probability at least

$$1 - \left(1 + \frac{2}{\epsilon_0} \right)^{2K} e^{-4K \log m} \geq 1 - O\left(m^{-10} \right).$$

Approximation. Our goal is then to extend the above concentration result to cover all $\boldsymbol{h} \in \mathcal{B}_{\boldsymbol{h}}$, $\boldsymbol{u} \in \mathcal{S}^{K-1}$ simultaneously, towards which we invoke the standard epsilon-net argument. For any $\boldsymbol{u} \in \mathcal{S}^{K-1}$, let $\boldsymbol{u}_0 \in \mathcal{N}_0$ be a point satisfying $\|\boldsymbol{u} - \boldsymbol{u}_0\|_2 \leq \epsilon_0$. Then straightforward calculation gives

$$\begin{aligned}
&\left| \left(\sum_{j=1}^{m} \left| \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \right|^2 \left(\left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{u} \right|^2 - \|\boldsymbol{u}\|_2^2 \right) \right) - \left(\sum_{j=1}^{m} \left| \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \right|^2 \left(\left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{u}_0 \right|^2 - \|\boldsymbol{u}_0\|_2^2 \right) \right) \right| \\
&\quad \overset{\text{(i)}}{=} \left| \sum_{j=1}^{m} \left| \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \right|^2 \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{u} \right|^2 - \|\boldsymbol{h}^\star\|_2^2 \|\boldsymbol{u}\|_2^2 - \sum_{j=1}^{m} \left| \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \right|^2 \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{u}_0 \right|^2 + \|\boldsymbol{h}^\star\|_2^2 \|\boldsymbol{u}_0\|_2^2 \right| \\
&\quad = \left| \left\| \mathcal{A}\left(\boldsymbol{h}^\star \boldsymbol{u}^{\mathsf{H}} \right) \right\|_2^2 - \left\| \mathcal{A}\left(\boldsymbol{h}^\star \boldsymbol{u}_0^{\mathsf{H}} \right) \right\|_2^2 + \|\boldsymbol{u}_0\|_2^2 - \|\boldsymbol{u}\|_2^2 \right| \\
&\quad \overset{\text{(ii)}}{\leq} \left| \left\| \mathcal{A}\left(\boldsymbol{h}^\star \boldsymbol{u}^{\mathsf{H}} \right) \right\|_2^2 - \left\| \mathcal{A}\left(\boldsymbol{h}^\star \boldsymbol{u}_0^{\mathsf{H}} \right) \right\|_2^2 \right| + \|\boldsymbol{u}_0 - \boldsymbol{u}\|_2 \left(\|\boldsymbol{u}_0\|_2 + \|\boldsymbol{u}\|_2 \right)
\end{aligned}$$

$$\begin{aligned}
&\overset{\text{(iii)}}{\leq} \left|\left(\left\|\mathcal{A}\left(\boldsymbol{h}^\star\boldsymbol{u}^{\mathsf{H}}\right)\right\|_2 + \left\|\mathcal{A}\left(\boldsymbol{h}^\star\boldsymbol{u}_0^{\mathsf{H}}\right)\right\|_2\right)\left\|\mathcal{A}\left(\boldsymbol{h}^\star\boldsymbol{u}^{\mathsf{H}}\right) - \mathcal{A}\left(\boldsymbol{h}^\star\boldsymbol{u}_0^{\mathsf{H}}\right)\right\|_2\right| + \epsilon_0 \\
&\lesssim \|\mathcal{A}\|^2\left(\|\boldsymbol{h}^\star\|_2\|\boldsymbol{u}\|_2 + \|\boldsymbol{h}^\star\|_2\|\boldsymbol{u}_0\|_2\right)\|\boldsymbol{h}^\star\|_2\|\boldsymbol{u}-\boldsymbol{u}_0\|_2 + \epsilon_0 \\
&\overset{\text{(iv)}}{\leq} (4K\log K + 20\log m + 1)\,\epsilon_0,
\end{aligned}$$

where (i) comes from $\sum_{j=1}^m \left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\right|^2 = \|\boldsymbol{h}^\star\|_2^2$; (ii) and (iii) are due to triangle inequality; (iv) follows from the following bound

$$\|\mathcal{A}\| \leq \sqrt{2K\log K + 10\log m}, \tag{B.66}$$

which holds with probability at least $1 - O\left(m^{-10}\right)$ according to Lemma 1. Letting $\epsilon_0 = r/\left(m\log m\right)$ with $r = \lambda + \sigma\sqrt{K\log m}$, we note it satisfies

$$1 - \left(1 + \frac{2}{\epsilon_0}\right)^{2K} e^{-4K\log m} \geq 1 - O\left(m^{-10}\right).$$

Putting all this together. Therefore, we conclude that: with probability at least $1 - O\left(m^{-10}\right)$, one has

$$\begin{aligned}
|\eta| &\leq 4\|\boldsymbol{B}\boldsymbol{h}^\star\|_\infty\sqrt{2K\log m} + 16\|\boldsymbol{B}\boldsymbol{h}^\star\|_\infty^2 K\log m + (4K\log K + 20\log m + 1)\,\epsilon_0 \\
&\lesssim \sqrt{\frac{\mu^2 K\log m}{m}}
\end{aligned}$$

uniformly for all $\boldsymbol{h} \in \mathcal{B}_{\boldsymbol{h}}$ and $\boldsymbol{u} \in \mathcal{S}^{K-1}$, with the proviso that $m \geq C\mu^2 K\log m$. Here, the second inequality arises from (7).

C Analysis: Nonconvex formulation under Gaussian design

We consider the loss function

$$\underset{\boldsymbol{Z}\in\mathbb{C}^{K\times K}}{\text{minimize}} \quad f(\boldsymbol{h},\boldsymbol{x}) = \sum_{j=1}^m \left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}\boldsymbol{a}_j - y_j\right|^2 + \lambda\|\boldsymbol{h}\|_2^2 + \lambda\|\boldsymbol{x}\|_2^2. \tag{C.1}$$

The main idea similar to the one presented in Appendix A, although the proof for Gaussian design is easier due to the presence of more randomness. We shall also assume $\|\boldsymbol{h}^\star\|_2 = \|\boldsymbol{x}^\star\|_2 = 1$ for the sake of simplicity and adopt the same notation as (A.2a)-(A.5b). The main part of the analysis lies in demonstrating the following set of hypotheses by induction:

$$\mathsf{dist}\left(\boldsymbol{z}^t,\boldsymbol{z}^\star\right) \leq \left\|\widehat{\boldsymbol{z}}^{t-1/2} - \boldsymbol{z}^\star\right\|_2 \leq \rho\,\mathsf{dist}\left(\boldsymbol{z}^{t-1},\boldsymbol{z}^\star\right) + C_{11}\eta\left(\lambda + \sigma\sqrt{mK\log m}\right) \tag{C.2a}$$

$$\mathsf{dist}\big(\boldsymbol{z}^{t,(l)},\widetilde{\boldsymbol{z}}^t\big) \leq C_{12}\left(\frac{\sqrt{K\log^3 m}}{m} + \frac{\sigma\sqrt{K\log^2 m}}{m}\right) \tag{C.2b}$$

$$\max_{1\leq l\leq m}\left|\boldsymbol{a}_l^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^t - \boldsymbol{x}^\star\right)\right| \leq C_{13}\left(\frac{\sqrt{mK\log^3 m}}{m} + \frac{\sigma\sqrt{mK\log^2 m}}{m}\right) \tag{C.2c}$$

$$\max_{1\leq l\leq m}\left|\boldsymbol{b}_l^{\mathsf{H}}\big(\widetilde{\boldsymbol{h}}^t - \boldsymbol{h}^\star\big)\right| \leq C_{13}\left(\frac{\sqrt{mK\log^3 m}}{m} + \frac{\sigma\sqrt{mK\log^2 m}}{m}\right) \tag{C.2d}$$

for some constants $C_{11}, C_{12}, C_{13} > 0$. Additionally, to complete the induction argument for the base case, we are in need of the following results of initialization:

$$\text{dist}\left(\boldsymbol{z}^0,\boldsymbol{z}^\star\right) \lesssim \frac{\sqrt{mK\log^2 m}}{m} + \frac{\sigma\sqrt{mK\log m}}{m}, \tag{C.3a}$$

$$\mathsf{dist}\big(\boldsymbol{z}^{0,(l)}, \widetilde{\boldsymbol{z}}^{0}\big) \le C_{13}\left(\frac{\sqrt{K\log^3 m}}{m} + \frac{\sigma\sqrt{K\log^2 m}}{m}\right), \tag{C.3b}$$

$$\max_{1\le l\le m}\left|\boldsymbol{a}_l^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^0 - \boldsymbol{x}^\star\right)\right| \le C_{12}\left(\frac{\sqrt{mK\log^3 m}}{m} + \frac{\sigma\sqrt{mK\log^2 m}}{m}\right), \tag{C.3c}$$

$$\max_{1\le l\le m}\left|\boldsymbol{b}_l^{\mathsf{H}}\big(\widetilde{\boldsymbol{h}}^0 - \boldsymbol{h}^\star\big)\right| \le C_{13}\left(\frac{\sqrt{mK\log^3 m}}{m} + \frac{\sigma\sqrt{mK\log^2 m}}{m}\right). \tag{C.3d}$$

C.1 Induction analysis

Before embarking on the analysis, we state below a useful lemma which is direct consequence of the hypotheses (C.2) and (C.3).

Lemma 19. *Instate the notation and assumptions in Theorem 3. For $t \ge 0$, suppose that the hypotheses (C.2) and (A.14) hold in the first t iterations. Then there exist some constants $C, C' > 0$ such that for any $1 \le l \le m$,*

$$\mathsf{dist}\left(\boldsymbol{z}^t, \boldsymbol{z}^\star\right) \le C\left(\frac{\sqrt{mK\log^2 m}}{m} + \frac{\lambda + \sigma\sqrt{mK\log m}}{m}\right), \tag{C.4a}$$

$$\left\|\boldsymbol{h}^t\left(\boldsymbol{x}^t\right)^{\mathsf{H}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right\| \le C'\left(\frac{\sqrt{mK\log^2 m}}{m} + \frac{\lambda + \sigma\sqrt{mK\log m}}{m}\right), \tag{C.4b}$$

$$\left\|\widetilde{\boldsymbol{z}}^{t,(l)} - \boldsymbol{z}^\star\right\|_2 \le 2C\left(\frac{\sqrt{mK\log^2 m}}{m} + \frac{\lambda + \sigma\sqrt{mK\log m}}{m}\right), \tag{C.4c}$$

$$\frac{1}{2} \le \left\|\widetilde{\boldsymbol{x}}^t\right\|_2 \le \frac{3}{2}, \qquad \frac{1}{2} \le \left\|\widetilde{\boldsymbol{h}}^t\right\|_2 \le \frac{3}{2}, \tag{C.4d}$$

$$\frac{1}{2} \le \left\|\widetilde{\boldsymbol{x}}^{t,(l)}\right\|_2 \le \frac{3}{2}, \qquad \frac{1}{2} \le \left\|\widetilde{\boldsymbol{h}}^{t,(l)}\right\|_2 \le \frac{3}{2}, \tag{C.4e}$$

$$\frac{1}{2} \le \left\|\widehat{\boldsymbol{x}}^{t,(l)}\right\|_2 \le \frac{3}{2}, \qquad \frac{1}{2} \le \left\|\widehat{\boldsymbol{h}}^{t,(l)}\right\|_2 \le \frac{3}{2}. \tag{C.4f}$$

In addition, if $t > 0$, then one also has

$$\left\|\widehat{\boldsymbol{z}}^{t-1/2} - \boldsymbol{z}^\star\right\|_2 \le C\left(\frac{\sqrt{mK\log^2 m}}{m} + \frac{\lambda + \sigma\sqrt{K\log m}}{m}\right). \tag{C.4g}$$

This lemma can be proved in the same manner as Lemma 3 and hence we omit the proof here for brevity.

C.1.1 Characterizing local geometry

Similar to the nonconvex analysis of blind deconvolution, our first step is to establish some kind of restricted strong convexity and smoothness as described in the following lemma. The proof can be found in Appendix C.2.

Lemma 20. *Let $\delta := c/\log^2 m$ for some sufficiently small constant $c > 0$. Suppose that $m \ge CK\log^6 m$ for some sufficiently large constant $C > 0$ and that $\sigma\sqrt{K\log^3 m/m} \le c_1$ for some sufficiently small constant $c_1 > 0$. Then with probability $1 - O\left(m^{-10} + e^{-K}\log m\right)$, one has*

$$\boldsymbol{u}^{\mathsf{H}}\left[\boldsymbol{D}\nabla^2 f\left(\boldsymbol{z}\right) + \nabla^2 f\left(\boldsymbol{z}\right)\boldsymbol{D}\right]\boldsymbol{u} \ge \frac{m}{4}\left\|\boldsymbol{u}\right\|_2^2 \quad \text{and}$$

$$\left\|\nabla^2 f\left(\boldsymbol{z}\right)\right\| \le 3m$$

simultaneously for all points

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

obeying the following properties:

- $\boldsymbol{z}$ *satisfies*

$$\max\left\{\|\boldsymbol{h} - \boldsymbol{h}^\star\|_2, \|\boldsymbol{x} - \boldsymbol{x}^\star\|_2\right\} \leq \delta, \tag{C.5a}$$

$$\max_{1\leq j\leq m}\left\{\left|\boldsymbol{a}_j^{\mathsf{H}}\left(\boldsymbol{x} - \boldsymbol{x}^\star\right)\right|, \left|\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{h} - \boldsymbol{h}^\star\right)\right|\right\} \leq C_{13}\frac{1}{\log^{3/2} m}, \tag{C.5b}$$

- $\boldsymbol{z}_1 := (\boldsymbol{h}_1, \boldsymbol{x}_1)$ *is aligned with* $\boldsymbol{z}_2 := (\boldsymbol{h}_2, \boldsymbol{x}_2)$ *in the sense that* $\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|_2 = \mathsf{dist}(\boldsymbol{z}_1, \boldsymbol{z}_2)$*; in addition, they satisfy*

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

- $\gamma_1, \gamma_2 \in \mathbb{R}$ *and obey*

$$\max\left\{|\gamma_1 - 1|, |\gamma_2 - 1|\right\} \leq \delta.$$

C.1.2 ℓ_2 error contraction

Next, by employing the established restricted strong convexity and smoothness in Lemma 20, we can prove the hypothesis (C.2a) holds inductively. Our result is this:

Lemma 21. *Set* $\lambda = C_\lambda \sigma\sqrt{mK\log m}$ *for some sufficiently large constant* $C_\lambda > 0$ *and the stepsize* $\eta = c_\eta/m$ *for some sufficiently small constant* $c_\eta > 0$*. Suppose the sample complexity satisfies* $m \geq CK\log^3 m$ *for some sufficiently large constant* $C > 0$*. Then if the hypotheses (C.2) hold true at* t*th iteration, we have for some constant* $C_{11} > 0$*,*

$$\mathsf{dist}\left(\boldsymbol{z}^{t+1}, \boldsymbol{z}^\star\right) \leq (1 - c_\eta/16)\,\mathsf{dist}\left(\boldsymbol{z}^t, \boldsymbol{z}^\star\right) + C_{11}\eta\left(\lambda + \sigma\sqrt{mK\log m}\right),$$

holds with probability exceeding $1 - O(m^{-100})$.

Proof. The proof is the same as the analysis for Lemma 5 with the help of Lemma 20 and thus omitted here for simplicity. □

Before moving on to the next step, we provide a corollary to guarantee that the alignment parameters does not change much between adjacent iterates.

Corollary 2. *Instate the notation and assumptions in Theorem 3. For an integer* $t > 0$*, suppose that the hypotheses (A.6) and (A.14) hold in the first* $t - 1$ *iterations. Then there exists some constant* $C > 0$ *such that for any* $1 \leq l \leq m$*, one has*

$$\left|\left|\alpha^t\right| - 1\right| \lesssim \mathsf{dist}\left(\tilde{\boldsymbol{z}}^t, \boldsymbol{z}^\star\right) \lesssim \frac{\sqrt{mK\log^2 m}}{m} + \frac{\sigma\sqrt{mK\log m}}{m}, \tag{C.6a}$$

$$\left|\frac{\alpha^{t-1/2}}{\alpha^{t-1}} - 1\right| \lesssim c_\eta\left(\frac{\sqrt{mK\log^2 m}}{m} + \frac{\sigma\sqrt{mK\log m}}{m}\right), \tag{C.6b}$$

$$\left|\left|\alpha_{\text{mutual}}^{t,(l)}\right| - 1\right| \lesssim \left\|\tilde{\boldsymbol{z}}^{t,(l)} - \boldsymbol{z}^\star\right\|_2 \lesssim \frac{\sqrt{mK\log^2 m}}{m} + \frac{\sigma\sqrt{mK\log m}}{m}, \tag{C.6c}$$

$$\frac{1}{2} \leq \left\|\boldsymbol{x}^t\right\|_2 \leq \frac{3}{2}, \qquad \frac{1}{2} \leq \left\|\boldsymbol{h}^t\right\|_2 \leq \frac{3}{2}, \tag{C.6d}$$

$$\frac{1}{2} \leq \left\|\boldsymbol{x}^{t,(l)}\right\|_2 \leq \frac{3}{2}, \qquad \frac{1}{2} \leq \left\|\boldsymbol{h}^{t,(l)}\right\|_2 \leq \frac{3}{2} \tag{C.6e}$$

with probability at least $1 - O\left(m^{-100} + e^{-CK}\log m\right)$.

This corollary can be proved in the same way as Corollary 1 and hence we omit it here for simplicity.

C.1.3 Leave-one-out proximity

The next step is to control the discrepancy between the leave-one-out sequence and the original sequence. The formal statement is given in the lemma below.

Lemma 22. *Suppose the sample size obeys $m \geq CK \log^3 m$ for some large enough constant $C > 0$. If the hypotheses (C.2) hold true for the tth iteration, then with probability exceeding $1 - O(m^{-10})$, we have*

$$\max_{1 \leq l \leq m} \mathsf{dist}\left(\boldsymbol{z}^{t+1,(l)}, \widetilde{\boldsymbol{z}}^{t+1}\right) \leq C_{12}\left(\frac{\sqrt{K \log^3 m}}{m} + \frac{\sigma\sqrt{K} \log m}{m}\right), \tag{C.7}$$

$$\max_{1 \leq l \leq m} \left\|\widetilde{\boldsymbol{z}}^{t+1,(l)} - \widetilde{\boldsymbol{z}}^{t+1}\right\|_2 \lesssim C_{12}\left(\frac{\sqrt{K \log^3 m}}{m} + \frac{\sigma\sqrt{K} \log m}{m}\right). \tag{C.8}$$

The proof can be found in Appendix C.3.

C.1.4 Establishing incoherence

Then we proceed to prove the incoherence hypotheses, i.e. (C.2c) and (C.2d). They are much easier to handle than the Fourier designs. We actually only need to prove the incoherence of $\boldsymbol{a}_l$ and $\boldsymbol{x}^{t+1}$. Then the other follows immediately by the symmetry between $\{\boldsymbol{a}_j\}_{j=1}^m$ and $\{\boldsymbol{b}_j\}_{j=1}^m$ under Assumption 2. Similar to (A.21), the triangle inequality and Cauchy-Schwarz inequality yield

$$\begin{aligned}
\left|\boldsymbol{a}_l^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^{t+1} - \boldsymbol{x}^{\star}\right)\right| &\leq \left|\boldsymbol{a}_l^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^{t+1} - \widetilde{\boldsymbol{x}}^{t+1,(l)}\right)\right| + \left|\boldsymbol{a}_l^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^{t+1,(l)} - \boldsymbol{x}^{\star}\right)\right| \\
&\leq \|\boldsymbol{a}_l\|_2 \left\|\widetilde{\boldsymbol{x}}^{t+1} - \widetilde{\boldsymbol{x}}^{t+1,(l)}\right\|_2 + \left|\boldsymbol{a}_l^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^{t+1,(l)} - \boldsymbol{x}^{\star}\right)\right| \\
&\leq 10\sqrt{K} \cdot C_{12}\left(\frac{\sqrt{K \log^3 m}}{m} + \frac{\sigma\sqrt{K \log^2 m}}{m}\right) \\
&\quad + 20\sqrt{\log m} \cdot 2C_{11}\left(\frac{\sqrt{mK \log^2 m}}{m} + \frac{\sigma\sqrt{mK \log m}}{m}\right) \\
&\leq C_{13}\left(\frac{\sqrt{mK \log^3 m}}{m} + \frac{\sigma\sqrt{mK \log^2 m}}{m}\right),
\end{aligned} \tag{C.9}$$

where the penultimate inequality follows from (F.1), (F.2) and (C.8). This establishes the hypothesis (C.2c) for the $(t+1)$-th iteration.

The incoherence of $\boldsymbol{b}_l$ and $\boldsymbol{h}^{t+1}$ (as stated in the hypothesis (C.2d)) follows from the symmetry between $\{\boldsymbol{a}_j\}_{j=1}^m$ and $\{\boldsymbol{b}_j\}_{j=1}^m$. We summarize the results in the following lemma.

Lemma 23. *Suppose the sample complexity obeys $m \geq CK \log m$ for some sufficiently large constant $C > 0$ and $\lambda = C_\lambda \sigma \sqrt{mK \log m}$ for some absolute constant $C_\lambda > 0$. If the hypotheses (C.2a)-(C.2d) hold for the tth iteration, then with probability exceeding $1 - O\left(m^{-100}\right)$ for some constant $C_{13} > 0$, one has*

$$\max_{1 \leq l \leq m} \left|\boldsymbol{a}_l^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^{t+1} - \boldsymbol{x}^{\star}\right)\right| \leq C_{13}\left(\frac{\sqrt{mK \log^3 m}}{m} + \frac{\sigma\sqrt{mK \log^2 m}}{m}\right),$$

$$\max_{1 \leq l \leq m} \left|\boldsymbol{b}_l^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^{t+1} - \boldsymbol{h}^{\star}\right)\right| \leq C_{13}\left(\frac{\sqrt{mK \log^3 m}}{m} + \frac{\sigma\sqrt{mK \log^2 m}}{m}\right),$$

as long as $C_{13} > 0$ is some sufficiently large constant and $\eta > 0$ is taken to be some sufficiently small constant.

C.1.5 The base case: Spectral initialization

The last step of the proof is to establish the induction hypotheses for the base case. The following three lemmas justify (C.3a)-(C.3d) respectively.

Lemma 24. *Suppose the sample size satisfies $m \geq CK \log^5 m$ for some large enough constant $C > 0$. Then with probability exceeding $1 - O(m^{-10})$, one has*

$$\mathsf{dist}\left(\boldsymbol{z}^0, \boldsymbol{z}^\star\right) \lesssim \sqrt{\frac{K \log^2 m}{m}} + \sigma \sqrt{\frac{K \log m}{m}},$$
$$\mathsf{dist}\left(\boldsymbol{z}^{0,(l)}, \boldsymbol{z}^\star\right) \lesssim \sqrt{\frac{K \log^2 m}{m}} + \sigma \sqrt{\frac{K \log m}{m}}, \qquad 1 \leq l \leq m,$$

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

Proof. With the aid of Lemma 40, the proof is essentially identical to Ma et al. [2018, Eqn (94)] and thus omitted here for brevity. □

Lemma 25. *Suppose $m \geq CK \log^5 m$ for some sufficiently large constant $C_{12} > 0$. Then with probability at least $1 - O(m^{-1})$, one has*

$$\max_{1 \leq l \leq m} \mathsf{dist}\left(\boldsymbol{z}^{0,(l)}, \widetilde{\boldsymbol{z}}^0\right) \leq \frac{C_{12}\sqrt{K \log^3 m}}{m}.$$

Proof. The proof of this lemma is deferred to Appendix C.4. □

Lemma 26. *Suppose that $m \geq CK \log^6 m$ for some large enough constant $C > 0$. Then with probability at least $1 - O(m^{-1})$, we have*

$$\max_{1 \leq j \leq m} \left|\boldsymbol{a}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^0 - \boldsymbol{x}^\star\right)\right| \leq C_{13}\left(\sqrt{\frac{K \log^3 m}{m}} + \sigma \sqrt{\frac{K \log^2 m}{m}}\right),$$
$$\max_{1 \leq j \leq m} \left|\boldsymbol{b}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^0 - \boldsymbol{h}^\star\right)\right| \leq C_{13}\left(\sqrt{\frac{K \log^3 m}{m}} + \sigma \sqrt{\frac{K \log^2 m}{m}}\right).$$

Proof. The first inequality can be established by the same derivation as for Ma et al. [2018, Lemma 21], which is omitted here for brevity. The second inequality follows immediately since $\{\boldsymbol{a}_j\}_{j=1}^m$ and $\{\boldsymbol{b}_j\}_{j=1}^m$ have the same distributions. □

C.2 Proof of Lemma 20

To begin with, we decompose $\nabla^2 f(\boldsymbol{z})$ as follows

$$\nabla^2 f(\boldsymbol{z}) = \lambda \boldsymbol{I}_{4K} + \mathbb{E}\left[\nabla^2 f_{\mathsf{reg\text{-}free}}(\boldsymbol{z}^\star)\right] + \left(\nabla^2 f(\boldsymbol{z}) - \mathbb{E}\left[\nabla^2 f_{\mathsf{reg\text{-}free}}(\boldsymbol{z}^\star)\right] - \lambda \boldsymbol{I}_{4K}\right),$$

where

$$f_{\mathsf{reg\text{-}free}}(\boldsymbol{z}) = \sum_{j=1}^m \left|\boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h} \boldsymbol{x}^{\mathsf{H}} \boldsymbol{a}_j - y_j\right|^2.$$

The following two lemmas allow us to control the two terms on the right-hand side of the above identity separately.

Lemma 27. *Instate the notation and conditions of Lemma 20. One has*

$$\left\|\mathbb{E}\left[\nabla^2 f_{\mathsf{reg\text{-}free}}(\boldsymbol{z}^\star)\right]\right\| = 2m \qquad \text{and} \qquad \boldsymbol{u}^{\mathsf{H}}\left[\boldsymbol{D}\mathbb{E}\left[\nabla^2 f_{\mathsf{reg\text{-}free}}(\boldsymbol{z}^\star)\right] + \mathbb{E}\left[\nabla^2 f_{\mathsf{reg\text{-}free}}(\boldsymbol{z}^\star)\right]\boldsymbol{D}\right]\boldsymbol{u} \geq m\|\boldsymbol{u}\|_2^2.$$

Proof. Note that the expression of $\mathbb{E}[\nabla^2 f_{\mathsf{reg\text{-}free}}(\boldsymbol{z}^\star)]/m$ is the same as that of $\nabla^2 F(\boldsymbol{z}^\star)$ in Ma et al. [2018, Lemma 26]. Hence the proof there can be straightforwardly adapted to our case and thus omitted here. □

Lemma 28. *Suppose the sample size obeys $m \geq CK \log^3 m$ for some large enough constant $C > 0$. Then with probability at least $1 - O(m^{-10})$, one has*

$$\left\| \nabla^2 f(\boldsymbol{z}) - \mathbb{E}\left[\nabla^2 f(\boldsymbol{z}^\star)\right] \right\| \leq \frac{1}{4} m$$

holds uniformly for all $\boldsymbol{z}$ satisfying (C.5).

Proof. See Appendix C.2.1. □

With these two lemmas in hand, we have, for any $(\boldsymbol{h}, \boldsymbol{x})$ obeying (C.5), that

$$\begin{aligned}
\left\| \nabla^2 f(\boldsymbol{z}) \right\| &\leq \left\| \mathbb{E}\left[\nabla^2 f(\boldsymbol{z}^\star)\right] \right\| + \left\| \nabla^2 f(\boldsymbol{z}) - \mathbb{E}\left[\nabla^2 f(\boldsymbol{z}^\star)\right] \right\| \\
&\leq \left\| \mathbb{E}\left[\nabla^2 f_{\textsf{reg-free}}(\boldsymbol{z}^\star)\right] \right\| + \lambda + \left\| \nabla^2 f(\boldsymbol{z}) - \mathbb{E}\left[\nabla^2 f(\boldsymbol{z}^\star)\right] \right\| \\
&\leq 2m + \lambda + \frac{1}{4} m \\
&\leq 3m.
\end{aligned}$$

Furthermore, it is readily seen that

$$\begin{aligned}
&\boldsymbol{u}^{\mathsf{H}} \left[\boldsymbol{D} \nabla^2 f(\boldsymbol{z}) + \nabla^2 f(\boldsymbol{z}) \boldsymbol{D}\right] \boldsymbol{u} \\
&\quad = \boldsymbol{u}^{\mathsf{H}} \left\{\boldsymbol{D} \mathbb{E}\left[\nabla^2 f_{\textsf{reg-free}}(\boldsymbol{z}^\star)\right] + \mathbb{E}\left[\nabla^2 f_{\textsf{reg-free}}(\boldsymbol{z}^\star)\right] \boldsymbol{D}\right\} \boldsymbol{u} + 2\lambda \boldsymbol{u}^{\mathsf{H}} \boldsymbol{D} \boldsymbol{u} \\
&\qquad + \boldsymbol{u}^{\mathsf{H}} \boldsymbol{D} \left\{\nabla^2 f(\boldsymbol{z}) - \mathbb{E}\left[\nabla^2 f(\boldsymbol{z}^\star)\right]\right\} \boldsymbol{u} + \boldsymbol{u}^{\mathsf{H}} \left\{\nabla^2 f(\boldsymbol{z}) - \mathbb{E}\left[\nabla^2 f(\boldsymbol{z}^\star)\right]\right\} \boldsymbol{D} \boldsymbol{u} \\
&\quad \overset{\text{(i)}}{\geq} m \|\boldsymbol{u}\|_2^2 + 2\lambda (1 - \delta) \|\boldsymbol{u}\|_2^2 - 2 \|\boldsymbol{D}\| \left\| \nabla^2 f(\boldsymbol{z}) - \mathbb{E}\left[\nabla^2 f(\boldsymbol{z}^\star)\right] \right\| \|\boldsymbol{u}\|_2^2 \\
&\quad \overset{\text{(ii)}}{\geq} m \|\boldsymbol{u}\|_2^2 + 2\lambda (1 - \delta) \|\boldsymbol{u}\|_2^2 - 2(1 + \delta) \cdot \frac{1}{4} m \|\boldsymbol{u}\|_2^2 \\
&\quad \overset{\text{(iii)}}{\geq} \frac{1}{4} m \|\boldsymbol{u}\|_2^2,
\end{aligned}$$

where (i) is due to Lemma 27 and the fact that $\boldsymbol{u}^{\mathsf{H}} \boldsymbol{D} \boldsymbol{u} \geq (1 - \delta) \|\boldsymbol{u}\|_2^2$; (ii) relies on the bound $\|\boldsymbol{D}\| \leq 1 + \delta$ and Lemma 28; and (iii) holds as long as $\delta \leq 1/4$. We have thus finished the proof for the desired smoothness and restricted strong convexity conditions.

C.2.1 Proof of Lemma 28

The idea of the proof is similar to that of Ma et al. [2018, Lemma 27] except that the design of $\{\boldsymbol{b}_j\}_{j=1}^m$ is different. By triangle inequality, we can upper bound the quantity of interest as

$$\left\| \nabla^2 f(\boldsymbol{z}) - \mathbb{E}\left[\nabla^2 f(\boldsymbol{z}^\star)\right] \right\| \leq 2\alpha_1 + 2\alpha_2 + 4\alpha_3 + 4\alpha_4, \tag{C.10}$$

where

$$\begin{aligned}
\alpha_1 &= \left\| \sum_{j=1}^m \left|\boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}\right|^2 \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} - m \boldsymbol{I}_K \right\|, \qquad &\alpha_2 &= \left\| \sum_{j=1}^m \left|\boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}\right|^2 \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} - m \boldsymbol{I}_K \right\|, \\
\alpha_3 &= \left\| \sum_{j=1}^m \left(\boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h} \boldsymbol{x}^{\mathsf{H}} \boldsymbol{a}_j - y_j\right) \boldsymbol{b}_j \boldsymbol{a}_j^{\mathsf{H}} \right\|, \qquad &\alpha_4 &= \left\| \sum_{j=1}^m \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h} \left(\boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}\right)^{\mathsf{H}} - m \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}} \right\|.
\end{aligned}$$

We will control these four terms separately as follows.

Controlling α_1. In terms of α_1, by the triangle inequality, one has

$$\alpha_1 \le \underbrace{\left\| \sum_{j=1}^{m} \left(\left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x} \right|^2 - \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 \right) \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \right\|}_{=: \gamma_1} + \underbrace{\left\| \sum_{j=1}^{m} \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} - m \boldsymbol{I}_K \right\|}_{=: \gamma_2}.$$

1. Regarding γ_1, we have

$$\begin{aligned}
\gamma_1 &\le \left\| \sum_{j=1}^{m} \left| \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x} \right|^2 - \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 \right| \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \right\| \\
&\le \max_{1 \le j \le m} \left| \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x} \right|^2 - \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 \right| \left\| \sum_{j=1}^{m} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \right\| \\
&\le \max_{1 \le j \le m} \left(\left| \boldsymbol{a}_j^{\mathsf{H}} \left(\boldsymbol{x} - \boldsymbol{x}^\star \right) \right|^2 + 2 \left| \boldsymbol{a}_j^{\mathsf{H}} \left(\boldsymbol{x} - \boldsymbol{x}^\star \right) \right| \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right| \right) \cdot \left\| \sum_{j=1}^{m} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \right\|. \qquad \text{(C.11)}
\end{aligned}$$

It is first seen that

$$\begin{aligned}
&\max_{1 \le j \le m} \left(\left| \boldsymbol{a}_j^{\mathsf{H}} \left(\boldsymbol{x} - \boldsymbol{x}^\star \right) \right|^2 + 2 \left| \boldsymbol{a}_j^{\mathsf{H}} \left(\boldsymbol{x} - \boldsymbol{x}^\star \right) \right| \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right| \right) \\
&\quad \le \left(C_{13} \tfrac{1}{\log^{3/2} m} \right)^2 + 2 \cdot C_{13} \tfrac{1}{\log^{3/2} m} \cdot 20 \sqrt{\log m} \\
&\quad \lesssim C_{13} \frac{1}{\log m}.
\end{aligned}$$

When it comes to $\| \sum_{j=1}^{m} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \|$, one has

$$\left\| \sum_{j=1}^{m} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \right\| \le \left\| \sum_{j=1}^{m} \left(\boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} - \boldsymbol{I}_K \right) \right\| + m. \qquad \text{(C.12)}$$

We intend to invoke the matrix Bernstein inequality Koltchinskii et al. [2011, Proposition 2] to control $\| \sum_{j=1}^{m} (\boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} - \boldsymbol{I}_K) \|$. Observe that

$$B_{\boldsymbol{Z}} := \left\| \left\| \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} - \boldsymbol{I}_K \right\| \right\|_{\psi_1} = \left\| \max \left\{ \left| \| \boldsymbol{b}_j \|_2^2 - 1 \right|, 1 \right\} \right\|_{\psi_1} \le \left\| \| \boldsymbol{b}_j \|_2 \right\|_{\psi_2}^2 + 1 \lesssim K..$$

Here, we have used $\left\| \| \boldsymbol{b}_j \|_2 \right\|_{\psi_2} \lesssim \sqrt{K}$ (cf. Vershynin [2018, Theorem 3.1.1]). In addition, simple calculation yields

$$\left\| \sum_{j=1}^{m} \mathbb{E}\left[\left(\boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} - \boldsymbol{I}_K \right) \left(\boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} - \boldsymbol{I}_K \right)^{\mathsf{H}} \right] \right\| = \left\| \sum_{j=1}^{m} \mathbb{E}\left[\boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} - \boldsymbol{I}_K \right] \right\| = (K+1)\, m,$$

and

$$\left\| \sum_{j=1}^{m} \mathbb{E}\left[\left(\boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} - \boldsymbol{I}_K \right)^{\mathsf{H}} \left(\boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} - \boldsymbol{I}_K \right) \right] \right\| = \left\| \sum_{j=1}^{m} \mathbb{E}\left[\left(\boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} - \boldsymbol{I}_K \right) \left(\boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} - \boldsymbol{I}_K \right)^{\mathsf{H}} \right] \right\| = (K+1)\, m.$$

As a result, by setting

$$\sigma_{\boldsymbol{Z}} := \max \left\{ \left\| \sum_{j=1}^{m} \mathbb{E}\left[\left(\boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} - \boldsymbol{I}_K \right) \left(\boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} - \boldsymbol{I}_K \right)^{\mathsf{H}} \right] \right\|^{1/2}, \left\| \sum_{j=1}^{m} \mathbb{E}\left[\left(\boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} - \boldsymbol{I}_K \right)^{\mathsf{H}} \left(\boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} - \boldsymbol{I}_K \right) \right] \right\|^{1/2} \right\}$$

$$= \sqrt{(K+1)\,m},$$

we are ready to apply the matrix Bernstein inequality Koltchinskii et al. [2011, Proposition 2] to derive

$$\left\| \sum_{j=1}^{m} \left(\boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} - \boldsymbol{I}_K \right) \right\| \lesssim \sigma_{\boldsymbol{Z}} \sqrt{\log m} + B_{\boldsymbol{Z}} \log \left(\frac{B_{\boldsymbol{Z}} \sqrt{m}}{\sigma_{\boldsymbol{Z}}} \right) \log m \lesssim \sqrt{mK \log m} \tag{C.13}$$

with high probability. Substitution of (C.13) into (C.12) yields

$$\left\| \sum_{j=1}^{m} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \right\| \le 2m, \tag{C.14}$$

as long as $m \gg K \log m$. Plugging this inequality into (C.11) gives

$$\gamma_1 \lesssim C_{13} \frac{m}{\log m}. \tag{C.15}$$

2. The second term γ_2 can be further decomposed as follows

$$\gamma_2 \le \left\| \sum_{j=1}^{m} \left(\left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 - 1 \right) \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \right\| + \left\| \sum_{j=1}^{m} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} - m \boldsymbol{I}_K \right\|. \tag{C.16}$$

The second term on the right-hand side of (C.16) has already been considered in (C.13). We are therefore left to control the first term. Let

$$\boldsymbol{W}_j := \left(\left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 \mathbb{1}_{\left\{ \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right| \le 20\sqrt{\log m} \right\}} - \mathbb{E}\left[\left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 \mathbb{1}_{\left\{ \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right| \le 20\sqrt{\log m} \right\}} \right] \right) \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}}.$$

We make the observation that

$$\begin{aligned} & \left\| \sum_{j=1}^{m} \left(\left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 - 1 \right) \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \right\| \\ & \quad \le \left\| \sum_{j=1}^{m} \left(\left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 \mathbb{1}_{\left\{ \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right| \le 20\sqrt{\log m} \right\}} - 1 \right) \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \right\| + \left\| \sum_{j=1}^{m} \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 \mathbb{1}_{\left\{ \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right| > 20\sqrt{\log m} \right\}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \right\|. \end{aligned} \tag{C.17}$$

Regarding the second term of (C.17), due to (F.2) we have

$$\left\| \sum_{j=1}^{m} \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 \mathbb{1}_{\left\{ \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right| > 20\sqrt{\log m} \right\}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \right\| = 0$$

holds with probability over $1 - O(m^{-100})$. For the first term of (C.17), one can derive

$$\begin{aligned} & \left\| \sum_{j=1}^{m} \left(\left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 \mathbb{1}_{\left\{ \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right| \le 20\sqrt{\log m} \right\}} - 1 \right) \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \right\| \\ & \quad \le \left\| \sum_{j=1}^{m} \boldsymbol{W}_j \right\| + \left\| \sum_{j=1}^{m} \left(\mathbb{E}\left[\left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 \mathbb{1}_{\left\{ \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right| \le 20\sqrt{\log m} \right\}} \right] - 1 \right) \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \right\| \\ & \quad \le \left\| \sum_{j=1}^{m} \boldsymbol{W}_j \right\| + \max_{1 \le j \le m} \left| \mathbb{E}\left[\left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 \mathbb{1}_{\left\{ \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right| \le 20\sqrt{\log m} \right\}} \right] - 1 \right| \cdot \left\| \sum_{j=1}^{m} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \right\| \end{aligned}$$

$$= \left\| \sum_{j=1}^{m} \boldsymbol{W}_j \right\| + \max_{1\le j\le m} \mathbb{E}\left[\left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 \mathbb{1}_{\left\{ \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right| > 20\sqrt{\log m} \right\}} \right] \cdot \left\| \sum_{j=1}^{m} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \right\|,$$

where the first inequality holds due to the triangle inequality. Invoking the Cauchy-Schwartz inequality yields

$$\begin{aligned} \mathbb{E}\left[\left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 \mathbb{1}_{\left\{ \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right| > 20\sqrt{\log m} \right\}} \right] &\le \sqrt{ \mathbb{E}\left[\left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right|^4 \right] \cdot \mathbb{P}\left(\left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right| > 20\sqrt{\log m} \right) } \\ &\le O\left(m^{-100} \right), \end{aligned}$$

which taken collectively with (C.14) gives

$$\max_{1\le j\le m} \mathbb{E}\left[\left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 \mathbb{1}_{\left\{ \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right| > 20\sqrt{\log m} \right\}} \right] \cdot \left\| \sum_{j=1}^{m} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \right\| \le O\left(m^{-100} \right) \cdot 2m = O\left(m^{-98} \right). \tag{C.18}$$

We can then invoke the matrix Bernstein inequality Koltchinskii et al. [2011, Proposition 2] to control $\| \sum_{j=1}^{m} \boldsymbol{W}_j \|$. To this end, note that

$$B_{\boldsymbol{Z}} := \left\| \left\| \boldsymbol{W}_j \right\| \right\|_{\psi_1} \le \left(20\sqrt{\log m} \right)^2 \cdot \left\| \left\| \boldsymbol{b}_j \right\|_2 \right\|_{\psi_2}^2 \lesssim K \log m,$$

where we have used $\left\| \|\boldsymbol{b}_j\|_2 \right\|_{\psi_2} \lesssim \sqrt{K}$ (cf. Vershynin [2018, Theorem 3.1.1]). In addition, simple calculation yields

$$\left\| \sum_{j=1}^{m} \mathbb{E}\left[\boldsymbol{W}_j \boldsymbol{W}_j^{\mathsf{H}} \right] \right\| = \left\| \sum_{j=1}^{m} \mathrm{Var}\left(\left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 \mathbb{1}_{\left\{ \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right| \le 20\sqrt{\log m} \right\}} \right) \mathbb{E}\left[\boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \right] \right\| \le 3\left(K+2 \right) m,$$

and

$$\left\| \sum_{j=1}^{m} \mathbb{E}\left[\boldsymbol{W}_j^{\mathsf{H}} \boldsymbol{W}_j \right] \right\| = \left\| \sum_{j=1}^{m} \mathbb{E}\left[\boldsymbol{W}_j \boldsymbol{W}_j^{\mathsf{H}} \right] \right\| \le 3\left(K+2 \right) m.$$

As a result, by setting

$$\sigma_{\boldsymbol{Z}} := \max\left\{ \left\| \sum_{j=1}^{m} \mathbb{E}\left[\boldsymbol{W}_j \boldsymbol{W}_j^{\mathsf{H}} \right] \right\|^{1/2}, \left\| \sum_{j=1}^{m} \mathbb{E}\left[\boldsymbol{W}_j^{\mathsf{H}} \boldsymbol{W}_j \right] \right\|^{1/2} \right\} \le \sqrt{3\left(K+2 \right) m},$$

we can apply the matrix Bernstein inequality Koltchinskii et al. [2011, Proposition 2] to derive

$$\left\| \sum_{j=1}^{m} \boldsymbol{W}_j \right\| \lesssim \sigma_{\boldsymbol{Z}} \sqrt{\log m} + B_{\boldsymbol{Z}} \log\left(\frac{B_{\boldsymbol{Z}} \sqrt{m}}{\sigma_{\boldsymbol{Z}}} \right) \log m \lesssim \sqrt{mK \log m} \tag{C.19}$$

with high probability, where the last inequality holds as long as $m \gg K \log^5 m$. Plugging (C.18) and (C.19) into (C.17) gives

$$\left\| \sum_{j=1}^{m} \left(\left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 - 1 \right) \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \right\| \lesssim \sqrt{mK \log m}. \tag{C.20}$$

Substitution of (C.13) and (C.20) into (C.16) yields

$$\gamma_2 \lesssim \sqrt{mK \log m}. \tag{C.21}$$

As a consequence, taking (C.15) and (C.21) collectively yields

$$\alpha_1 \lesssim \frac{m}{\log m} + \sqrt{mK \log m}. \tag{C.22}$$

Controlling α_2. Regarding α_2, since the roles played by $\{\boldsymbol{a}_j\}_{j=1}^m$ and $\{\boldsymbol{b}_j\}_{j=1}^m$ are symmetric in this problem, it is easily seen that α_2 admits the same bound as that of α_1.

Controlling α_3. When it comes to the third term α_3, one makes the observation that

$$\alpha_3 \le \left\| \sum_{j=1}^m \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \left(\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} - \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}} \right) \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \right\| + \left\| \sum_{j=1}^m \xi_j \boldsymbol{b}_j \boldsymbol{a}_j^{\mathsf{H}} \right\|. \tag{C.23}$$

The second term on the right-hand side of this relation has already been bounded by Lemma 36. Regarding the first term on the right-hand side of (C.23), one can further decompose

$$\begin{aligned}
&\left\| \sum_{j=1}^m \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \left(\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} - \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}} \right) \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \right\| \\
&\quad \le \left\| m \left(\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} - \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}} \right) \right\| + \left\| \sum_{j=1}^m \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \left(\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} - \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}} \right) \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} - m \left(\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} - \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}} \right) \right\| \\
&\quad \le \left\| m \left(\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} - \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}} \right) \right\| + \left\| \sum_{j=1}^m \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \left(\boldsymbol{x} - \boldsymbol{x}^\star \right)^{\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} - m \boldsymbol{h}^\star \left(\boldsymbol{x} - \boldsymbol{x}^\star \right)^{\mathsf{H}} \right\| \\
&\qquad + \left\| \sum_{j=1}^m \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \left(\boldsymbol{h} - \boldsymbol{h}^\star \right) \boldsymbol{x}^{\star\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} - m \left(\boldsymbol{h} - \boldsymbol{h}^\star \right) \boldsymbol{x}^{\star\mathsf{H}} \right\| \\
&\qquad + \left\| \sum_{j=1}^m \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \left(\boldsymbol{h} - \boldsymbol{h}^\star \right) \left(\boldsymbol{x} - \boldsymbol{x}^\star \right)^{\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} - m \left(\boldsymbol{h} - \boldsymbol{h}^\star \right) \left(\boldsymbol{x} - \boldsymbol{x}^\star \right)^{\mathsf{H}} \right\|.
\end{aligned} \tag{C.24}$$

To bound the last three terms of (C.24), we resort to the following two lemmas, whose proofs can be found in Appendix C.2.2 and Appendix C.2.3.

Lemma 29. *With probability at least $1 - O(m^{-100} + me^{-CK})$ for some constant $C > 0$, one has*

$$\left\| \sum_{j=1}^m \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \left(\boldsymbol{x} - \boldsymbol{x}^\star \right)^{\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} - m \boldsymbol{h}^\star \left(\boldsymbol{x} - \boldsymbol{x}^\star \right)^{\mathsf{H}} \right\| \le 2\delta m \tag{C.25}$$

holds uniformly for any $\boldsymbol{x}$ satisfying (C.5).

Lemma 30. *With probability at least $1 - 2\exp(-CK \log m)$ for some constant $C > 0$, one has*

$$\left\| \sum_{j=1}^m \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \left(\boldsymbol{h} - \boldsymbol{h}^\star \right) \left(\boldsymbol{x} - \boldsymbol{x}^\star \right)^{\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} - m \left(\boldsymbol{h} - \boldsymbol{h}^\star \right) \left(\boldsymbol{x} - \boldsymbol{x}^\star \right)^{\mathsf{H}} \right\| \le \delta^2 m + 4C' \sqrt{mK} \tag{C.26}$$

holds uniformly for any $(\boldsymbol{h}, \boldsymbol{x})$ obeying (C.5) for some sufficiently large constant $C' > 0$.

By the symmetry between $\{\boldsymbol{a}_j\}_{j=1}^m$ and $\{\boldsymbol{b}_j\}_{j=1}^m$ and Lemma 29, one arrives at

$$\sup_{\boldsymbol{x} \in \mathcal{S}} \left\| \sum_{j=1}^m \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \left(\boldsymbol{h} - \boldsymbol{h}^\star \right) \boldsymbol{x}^{\star\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} - m \left(\boldsymbol{h} - \boldsymbol{h}^\star \right) \boldsymbol{x}^{\star\mathsf{H}} \right\| \le 2\delta m \tag{C.27}$$

with probability over $1 - 2\exp(-CK \log m)$. Plugging (C.25), (C.26) and (C.27) into (C.24) yields

$$\left\| \sum_{j=1}^m \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \left(\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} - \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}} \right) \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \right\| \le 6\delta m. \tag{C.28}$$

Substitution of (C.28) and (D.3) into (C.23) thus gives

$$\alpha_3 \le 6\delta m + C\sigma\sqrt{mK\log m} \tag{C.29}$$

for some large enough constant $C > 0$.

Controlling α_4. With regards to the last term α_4, we have

$$\alpha_4 \le \underbrace{\left\|\sum_{j=1}^{m} \boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} - \boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\right)\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\right\|}_{=:\theta_1} + \underbrace{\left\|\sum_{j=1}^{m} \boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}} - m\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\right\|}_{=:\theta_2}.$$

These two terms have already been bounded by (C.28) and (F.9) respectively. Combining this inequality with (C.28) gives

$$\alpha_4 \le 6\delta m + 4C_t\sqrt{mK}\log m. \tag{C.30}$$

Putting all this together. Finally, by plugging (C.22), (C.29) and (C.30) into (C.10), we arrive at

$$\left\|\nabla^2 f(\boldsymbol{z}) - \nabla^2 F(\boldsymbol{z}^{\star})\right\| \lesssim \sigma\sqrt{mK\log m} + \frac{m}{\log m} \le \frac{1}{4}m$$

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

C.2.2 Proof of Lemma 29

Consider the event

$$\mathcal{E} := \left\{\max_{1\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^{\star}\right| \le 20\sqrt{\log m},\ \max_{1\le j\le m}\left\|\boldsymbol{a}_j\right\|_2 \le 10\sqrt{K}\right\}. \tag{C.31}$$

(F.1) and (F.2) suggest that event $\mathcal{E}$ holds with probability at least $1 - O(m^{-100} + me^{-CK})$. The proof thereafter will be developed on this event.

Due to the assumptions (C.5), we have — for any given unit vectors $\boldsymbol{u}$, $\boldsymbol{v} \in \mathbb{C}^K$ — that

$$\sum_{j=1}^{m} \boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^{\star}\left(\boldsymbol{x}-\boldsymbol{x}^{\star}\right)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v} = \sum_{j=1}^{m}\underbrace{\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^{\star}\left(\boldsymbol{x}-\boldsymbol{x}^{\star}\right)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{\left|(\boldsymbol{x}-\boldsymbol{x}^{\star})^{\mathsf{H}}\boldsymbol{a}_j\right| \le 20C_{13}\frac{1}{\log^{3/2}m}\right\}}}_{=:X_j}.$$

In what follows, we shall first establish concentration inequalities for this quantity for a given point $(\boldsymbol{u}, \boldsymbol{v})$, and then establish uniform concentration that holds for simultaneously for all points of interest.

Concentration. Consider any fixed unit vectors $\boldsymbol{u}$ and $\boldsymbol{v}$). We seek to invoke the Bernstein inequality Vershynin [2018, Theorem 2.8.2] to control $\sum_{j=1}^{m}(X_j - \mathbb{E}[X_j])$. We observe that

$$\begin{aligned}
\left\|X_j - \mathbb{E}\left[X_j\right]\right\|_{\psi_1} &\le C\left\|X_j\right\|_{\psi_1} \le C\left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^{\star}\left(\boldsymbol{x}-\boldsymbol{x}^{\star}\right)^{\mathsf{H}}\boldsymbol{a}_j\,\mathbb{1}_{\left\{\left|(\boldsymbol{x}-\boldsymbol{x}^{\star})^{\mathsf{H}}\boldsymbol{a}_j\right| \le 20C_{13}\frac{1}{\log^{3/2}m}\right\}}\right|\left\|\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\right\|_{\psi_2}\left\|\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\right\|_{\psi_2} \\
&= C\left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^{\star}\right| \cdot \left|\left(\boldsymbol{x}-\boldsymbol{x}^{\star}\right)^{\mathsf{H}}\boldsymbol{a}_j\,\mathbb{1}_{\left\{\left|(\boldsymbol{x}-\boldsymbol{x}^{\star})^{\mathsf{H}}\boldsymbol{a}_j\right| \le 20C_{13}\frac{1}{\log^{3/2}m}\right\}}\right| \\
&\le 400CC_{13}\frac{1}{\log m},
\end{aligned}$$

where the first inequality comes from the fact that $\|X-\mathbb{E}[X]\|_{\psi_1} \leq C\|X\|_{\psi_1}$ (cf. Vershynin [2018, Section 2.7]) and the last inequality is due to the event $\mathcal{E}$. Hence, the Bernstein inequality Vershynin [2018, Theorem 2.8.2] reveals that

$$\mathbb{P}\left(\left|\sum_{j=1}^{m}\left(X_j-\mathbb{E}\left[X_j\right]\right)\right| \geq t\right) \leq 2\exp\left(-c\min\left(\frac{t^2\log^2 m}{m}, t\log m\right)\right).$$

Letting $t=C_t\sqrt{mK}$ for some large enough constant $C_t>0$, we obtain

$$\left|\sum_{j=1}^{m}\left(X_j-\mathbb{E}\left[X_j\right]\right)\right| \leq C_t\sqrt{mK}, \tag{C.32}$$

with probability exceeding $1-2\exp(-cC_t^2K\log m)$.

Union bound over epsilon-nets. Next, we intend to show that (C.32) holds uniformly for any unit vectors $\boldsymbol{u}$ and $\boldsymbol{v}$. Define $\mathcal{N}_{\boldsymbol{x}}$ to be an ϵ_1-net of $\mathcal{B}_{\boldsymbol{x}}(\delta) := \{\boldsymbol{x} : \|\boldsymbol{x}-\boldsymbol{x}^\star\|_2 \leq \delta\}$ and $\mathcal{N}_0$ an ϵ_2-net of the unit sphere $\mathcal{S}^{K-1}$. In view of Vershynin [2018, Corollary 4.2.13], we can choose these nets to guarantee that

$$|\mathcal{N}_{\boldsymbol{x}}| \leq \left(1+\frac{2\delta}{\epsilon_1}\right)^{4K} \qquad \text{and} \qquad |\mathcal{N}_0| \leq \left(1+\frac{2}{\epsilon_2}\right)^{2K}.$$

Taking these collectively with the union bound reveals that (C.32) holds uniformly for all $\boldsymbol{x} \in \mathcal{N}_{\boldsymbol{x}}$ and $\boldsymbol{u}$, $\boldsymbol{v} \in \mathcal{N}_0$ with probability exceeding

$$1-\left(1+\frac{2\delta}{\epsilon_1}\right)^{4K}\left(1+\frac{2}{\epsilon_2}\right)^{4K}\cdot 2\exp\left(-cC_t^2K\log m\right) \geq 1-2\exp\left(-CK\log m\right).$$

Approximation. We then turn to the following quantity

$$g\left(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x}\right) := \sum_{j=1}^{m}\left[\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\left(\boldsymbol{x}-\boldsymbol{x}^\star\right)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\mathbb{1}_{\left\{\left|(\boldsymbol{x}-\boldsymbol{x}^\star)^{\mathsf{H}}\boldsymbol{a}_j\right| \leq 20C_{13}\frac{1}{\log^{3/2} m}\right\}}\boldsymbol{v} - m\boldsymbol{h}^\star\left(\boldsymbol{x}-\boldsymbol{x}^\star\right)^{\mathsf{H}}\right].$$

For any $\boldsymbol{x}$ satisfying the assumptions (C.5) and any $\boldsymbol{u}$, $\boldsymbol{v} \in \mathcal{S}^{K-1}$, one can choose $\boldsymbol{x}_0 \in \mathcal{N}_{\boldsymbol{x}}$, $\boldsymbol{u}_0 \in \mathcal{N}_0$ and $\boldsymbol{v}_0 \in \mathcal{N}_0$ satisfying $\|\boldsymbol{x}-\boldsymbol{x}_0\|_2 \leq \epsilon_1$ and $\max\{\|\boldsymbol{u}-\boldsymbol{u}_0\|_2, \|\boldsymbol{v}-\boldsymbol{v}_0\|_2\} \leq \epsilon_2$. Set $\epsilon_1 = \delta/K$ and $\epsilon_2 = 1/4$. The triangle inequality gives

$$\begin{aligned}
&\left|g\left(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x}\right)-g\left(\boldsymbol{u}_0,\boldsymbol{v}_0,\boldsymbol{x}_0\right)\right| \\
&\quad\leq \left|g\left(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x}\right)-g\left(\boldsymbol{u}_0,\boldsymbol{v},\boldsymbol{x}\right)\right| + \left|g\left(\boldsymbol{u}_0,\boldsymbol{v},\boldsymbol{x}\right)-g\left(\boldsymbol{u}_0,\boldsymbol{v}_0,\boldsymbol{x}\right)\right| \\
&\qquad + \left|g\left(\boldsymbol{u}_0,\boldsymbol{v}_0,\boldsymbol{x}\right)-g\left(\boldsymbol{u}_0,\boldsymbol{v}_0,\boldsymbol{x}_0\right)\right| \\
&\quad\leq 2\left\|\sum_{j=1}^{m}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\left(\boldsymbol{x}-\boldsymbol{x}^\star\right)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\mathbb{1}_{\left\{\left|(\boldsymbol{x}-\boldsymbol{x}^\star)^{\mathsf{H}}\boldsymbol{a}_j\right| \leq 20C_{13}\frac{1}{\log^{3/2} m}\right\}} - m\boldsymbol{h}^\star\left(\boldsymbol{x}-\boldsymbol{x}^\star\right)^{\mathsf{H}}\right\|\epsilon_2 \\
&\qquad + \left|g\left(\boldsymbol{u}_0,\boldsymbol{v}_0,\boldsymbol{x}\right)-g\left(\boldsymbol{u}_0,\boldsymbol{v}_0,\boldsymbol{x}_0\right)\right|.
\end{aligned} \tag{C.33}$$

To simplify the second term above, we notice that on event $\mathcal{E}$ (cf. (C.31)),

$$\left|\left(\boldsymbol{x}-\boldsymbol{x}_0\right)^{\mathsf{H}}\boldsymbol{a}_j\right| \leq \max_{1\leq j\leq m}\|\boldsymbol{a}_j\|_2 \cdot \|\boldsymbol{x}-\boldsymbol{x}_0\|_2 \leq 10\sqrt{K}\cdot\epsilon_1 \leq 2C_{13}\frac{1}{\log^{3/2} m}, \tag{C.34}$$

and hence

$$\left|\left(\boldsymbol{x}_0-\boldsymbol{x}^\star\right)^{\mathsf{H}}\boldsymbol{a}_j\right| \leq \left|\left(\boldsymbol{x}-\boldsymbol{x}^\star\right)^{\mathsf{H}}\boldsymbol{a}_j\right| + \left|\left(\boldsymbol{x}-\boldsymbol{x}_0\right)^{\mathsf{H}}\boldsymbol{a}_j\right|$$

$$\leq 4C_{13}\frac{1}{\log^{3/2}m}. \tag{C.35}$$

As a result, one has the following identity

$$\mathbb{1}_{\left\{\left|(\boldsymbol{x}-\boldsymbol{x}_0)^{\mathsf{H}}\boldsymbol{a}_j\right|\leq 20C_{13}\frac{1}{\log^{3/2}m}\right\}}=\mathbb{1}_{\left\{\left|(\boldsymbol{x}_0-\boldsymbol{x}^{\star})^{\mathsf{H}}\boldsymbol{a}_j\right|\leq 20C_{13}\frac{1}{\log^{3/2}m}\right\}}=\mathbb{1}_{\left\{\left|(\boldsymbol{x}-\boldsymbol{x}^{\star})^{\mathsf{H}}\boldsymbol{a}_j\right|\leq 20C_{13}\frac{1}{\log^{3/2}m}\right\}}=1. \tag{C.36}$$

It then follows that

$$\begin{aligned}&\left|g\left(\boldsymbol{u}_0,\boldsymbol{v}_0,\boldsymbol{x}\right)-g\left(\boldsymbol{u}_0,\boldsymbol{v}_0,\boldsymbol{x}_0\right)\right|\\&\quad=\left\|\sum_{j=1}^{m}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^{\star}\left(\boldsymbol{x}-\boldsymbol{x}_0\right)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\,\mathbb{1}_{\left\{\left|(\boldsymbol{x}-\boldsymbol{x}_0)^{\mathsf{H}}\boldsymbol{a}_j\right|\leq 20C_{13}\frac{1}{\log^{3/2}m}\right\}}-m\boldsymbol{h}^{\star}\left(\boldsymbol{x}-\boldsymbol{x}_0\right)^{\mathsf{H}}\right\|.\end{aligned} \tag{C.37}$$

Plugging (C.37) into (C.33) yields

$$\begin{aligned}&\left|g\left(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x}\right)-g\left(\boldsymbol{u}_0,\boldsymbol{v}_0,\boldsymbol{x}_0\right)\right|\\&\quad\leq 2\left\|\sum_{j=1}^{m}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^{\star}\left(\boldsymbol{x}-\boldsymbol{x}^{\star}\right)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\,\mathbb{1}_{\left\{\left|(\boldsymbol{x}-\boldsymbol{x}^{\star})^{\mathsf{H}}\boldsymbol{a}_j\right|\leq 20C_{13}\frac{1}{\log^{3/2}m}\right\}}-m\boldsymbol{h}^{\star}\left(\boldsymbol{x}-\boldsymbol{x}^{\star}\right)^{\mathsf{H}}\right\|\epsilon_2\\&\qquad+\left\|\sum_{j=1}^{m}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^{\star}\left(\boldsymbol{x}-\boldsymbol{x}_0\right)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\,\mathbb{1}_{\left\{\left|(\boldsymbol{x}-\boldsymbol{x}_0)^{\mathsf{H}}\boldsymbol{a}_j\right|\leq 20C_{13}\frac{1}{\log^{3/2}m}\right\}}-m\boldsymbol{h}^{\star}\left(\boldsymbol{x}-\boldsymbol{x}_0\right)^{\mathsf{H}}\right\|.\end{aligned} \tag{C.38}$$

Next, we look at $g(\boldsymbol{u}_0,\boldsymbol{v}_0,\boldsymbol{x}_0)$, and notice that (C.32) holds for

$$X_j=\boldsymbol{u}_0^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^{\star}\left(\boldsymbol{x}_0-\boldsymbol{x}^{\star}\right)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}_0\,\mathbb{1}_{\left\{\left|(\boldsymbol{x}_0-\boldsymbol{x}^{\star})^{\mathsf{H}}\boldsymbol{a}_j\right|\leq 20C_{13}\frac{1}{\log^{3/2}m}\right\}},$$

due to $\boldsymbol{x}_0\in\mathcal{N}_{\boldsymbol{x}}$, $\boldsymbol{u}_0\in\mathcal{N}_0$ and $\boldsymbol{v}_0\in\mathcal{N}_0$. By virtue of the triangle inequality, one has

$$\begin{aligned}&\left|g\left(\boldsymbol{u}_0,\boldsymbol{v}_0,\boldsymbol{x}_0\right)\right|\\&\quad\leq\left|\sum_{j=1}^{m}\left(X_j-\mathbb{E}\left[X_j\right]\right)\right|+\left|\sum_{j=1}^{m}\left(\mathbb{E}\left[X_j\right]-m\boldsymbol{u}_0^{\mathsf{H}}\boldsymbol{h}^{\star}\left(\boldsymbol{x}_0-\boldsymbol{x}^{\star}\right)^{\mathsf{H}}\boldsymbol{v}_0\right)\right|\\&\quad\leq C_t\sqrt{mK}+\left|\sum_{j=1}^{m}\mathbb{E}\left[\boldsymbol{u}_0^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^{\star}\left(\boldsymbol{x}_0-\boldsymbol{x}^{\star}\right)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{\left|(\boldsymbol{x}_0-\boldsymbol{x}^{\star})^{\mathsf{H}}\boldsymbol{a}_j\right|>20C_{13}\frac{1}{\log^{3/2}m}\right\}}\right]\right|\\&\quad\leq C_t\sqrt{mK},\end{aligned} \tag{C.39}$$

where

$$\left|\mathbb{E}\left[\boldsymbol{u}_0^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^{\star}\left(\boldsymbol{x}_0-\boldsymbol{x}^{\star}\right)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}_0\,\mathbb{1}_{\left\{\left|(\boldsymbol{x}_0-\boldsymbol{x}^{\star})^{\mathsf{H}}\boldsymbol{a}_j\right|>20C_{13}\frac{1}{\log^{3/2}m}\right\}}\right]\right|=0,$$

which is a consequence of (C.36).

Putting all this together. Let us define

$$\mathcal{S}:=\left\{\boldsymbol{x}:\left|\left(\boldsymbol{x}-\boldsymbol{x}^{\star}\right)^{\mathsf{H}}\boldsymbol{a}_j\right|\leq 20C_{13}\frac{1}{\log^{3/2}m},\ \left\|\boldsymbol{x}-\boldsymbol{x}^{\star}\right\|_2\leq\delta\right\}.$$

Taking (C.38) and (C.39) collectively gives rise to

$$\left|g\left(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x}\right)\right|\leq\left|g\left(\boldsymbol{u}_0,\boldsymbol{v}_0,\boldsymbol{x}_0\right)\right|+\left|g\left(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x}\right)-g\left(\boldsymbol{u}_0,\boldsymbol{v}_0,\boldsymbol{x}_0\right)\right|$$

$$\leq C_t\sqrt{mK} + 2\left\|\sum_{j=1}^m \boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\left(\boldsymbol{x}-\boldsymbol{x}^\star\right)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\mathbb{1}_{\left\{\left|(\boldsymbol{x}-\boldsymbol{x}^\star)^{\mathsf{H}}\boldsymbol{a}_j\right|\leq 20C_{13}\frac{1}{\log^{3/2}m}\right\}} - m\boldsymbol{h}^\star\left(\boldsymbol{x}-\boldsymbol{x}^\star\right)^{\mathsf{H}}\right\|\epsilon_2$$

$$+\left\|\sum_{j=1}^m \boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\left(\boldsymbol{x}-\boldsymbol{x}_0\right)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\mathbb{1}_{\left\{\left|(\boldsymbol{x}-\boldsymbol{x}_0)^{\mathsf{H}}\boldsymbol{a}_j\right|\leq 20C_{13}\frac{1}{\log^{3/2}m}\right\}} - m\boldsymbol{h}^\star\left(\boldsymbol{x}-\boldsymbol{x}_0\right)^{\mathsf{H}}\right\|. \tag{C.40}$$

A key observation is that $\boldsymbol{x}' := 5(\boldsymbol{x}-\boldsymbol{x}_0)+\boldsymbol{x}^\star \in \mathcal{S}$ by $\|\boldsymbol{x}-\boldsymbol{x}_0\|_2 \leq \epsilon_1$ and (C.35). Hence, the last term in (C.40) satisfies

$$\begin{aligned}
&\left\|\sum_{j=1}^m \boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\left(\boldsymbol{x}-\boldsymbol{x}_0\right)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\mathbb{1}_{\left\{\left|(\boldsymbol{x}-\boldsymbol{x}_0)^{\mathsf{H}}\boldsymbol{a}_j\right|\leq 20C_{13}\frac{1}{\log^{3/2}m}\right\}} - m\boldsymbol{h}^\star\left(\boldsymbol{x}-\boldsymbol{x}_0\right)^{\mathsf{H}}\right\| \\
&= \frac{1}{5}\left\|\sum_{j=1}^m \boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\left(\boldsymbol{x}'-\boldsymbol{x}_\star\right)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\mathbb{1}_{\left\{\left|(\boldsymbol{x}'-\boldsymbol{x}_\star)^{\mathsf{H}}\boldsymbol{a}_j\right|\leq 20C_{13}\frac{1}{\log^{3/2}m}\right\}} - m\boldsymbol{h}^\star\left(\boldsymbol{x}'-\boldsymbol{x}_0\right)^{\mathsf{H}}\right\| \\
&\leq \frac{1}{5}\sup_{\widetilde{\boldsymbol{x}}\in\mathcal{S}}\left\|\sum_{j=1}^m \boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\left(\widetilde{\boldsymbol{x}}-\boldsymbol{x}^\star\right)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\mathbb{1}_{\left\{\left|(\widetilde{\boldsymbol{x}}-\boldsymbol{x}^\star)^{\mathsf{H}}\boldsymbol{a}_j\right|\leq 20C_{13}\frac{1}{\log^{3/2}m}\right\}} - m\boldsymbol{h}^\star\left(\widetilde{\boldsymbol{x}}-\boldsymbol{x}^\star\right)^{\mathsf{H}}\right\|,
\end{aligned}$$

where the first equality comes from (C.34). Plugging this inequality into (C.40), taking the maximum over $\boldsymbol{u}$ and $\boldsymbol{v}$ on the left-hand side of (C.40) and rearranging terms yield

$$\begin{aligned}
&(1-2\epsilon_2)\left\|\sum_{j=1}^m \boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\left(\boldsymbol{x}-\boldsymbol{x}^\star\right)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\mathbb{1}_{\left\{\left|(\boldsymbol{x}-\boldsymbol{x}^\star)^{\mathsf{H}}\boldsymbol{a}_j\right|\leq 20C_{13}\frac{1}{\log^{3/2}m}\right\}} - m\boldsymbol{h}^\star\left(\boldsymbol{x}-\boldsymbol{x}^\star\right)^{\mathsf{H}}\right\| \\
&\leq 2C_t\sqrt{mK} + \frac{1}{5}\sup_{\widetilde{\boldsymbol{x}}\in\mathcal{S}}\left\|\sum_{j=1}^m \boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\left(\widetilde{\boldsymbol{x}}-\boldsymbol{x}^\star\right)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\mathbb{1}_{\left\{\left|(\widetilde{\boldsymbol{x}}-\boldsymbol{x}^\star)^{\mathsf{H}}\boldsymbol{a}_j\right|\leq 20C_{13}\frac{1}{\log^{3/2}m}\right\}} - m\boldsymbol{h}^\star\left(\widetilde{\boldsymbol{x}}-\boldsymbol{x}^\star\right)^{\mathsf{H}}\right\|.
\end{aligned}$$

Further, taking the maximum over $\boldsymbol{x}\in\mathcal{S}$ on the left-hand side of the above inequality gives

$$\left(1-2\epsilon_2-\frac{1}{5}\right)\sup_{\boldsymbol{x}\in\mathcal{S}}\left\|\sum_{j=1}^m \boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\left(\boldsymbol{x}-\boldsymbol{x}^\star\right)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\mathbb{1}_{\left\{\left|(\boldsymbol{x}-\boldsymbol{x}^\star)^{\mathsf{H}}\boldsymbol{a}_j\right|\leq 20C_{13}\frac{1}{\log^{3/2}m}\right\}} - m\boldsymbol{h}^\star\left(\boldsymbol{x}-\boldsymbol{x}^\star\right)^{\mathsf{H}}\right\| \leq 2C_t\sqrt{mK},$$

and, consequently,

$$\begin{aligned}
&\sup_{\boldsymbol{x}\in\mathcal{S}}\left\|\sum_{j=1}^m \boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\left(\boldsymbol{x}-\boldsymbol{x}^\star\right)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\mathbb{1}_{\left\{\left|(\boldsymbol{x}-\boldsymbol{x}^\star)^{\mathsf{H}}\boldsymbol{a}_j\right|\leq 20C_{13}\frac{1}{\log^{3/2}m}\right\}}\right\| \\
&= \sup_{\boldsymbol{x}\in\mathcal{S}}\left\|\sum_{j=1}^m \boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\left(\boldsymbol{x}-\boldsymbol{x}^\star\right)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\right\| \\
&\leq m\left\|\boldsymbol{h}^\star\left(\boldsymbol{x}-\boldsymbol{x}^\star\right)^{\mathsf{H}}\right\| + 4C_t\sqrt{mK} \\
&\leq 2\delta m,
\end{aligned}$$

as long as $m \gg K\log^4 m$.

C.2.3 Proof of Lemma 30

Similar to proof of Lemma 29, we consider the event

$$\mathcal{E} := \left\{\max_{1\leq j\leq m}\left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\right| \leq 20\sqrt{\log m},\ \max_{1\leq j\leq m}\|\boldsymbol{a}_j\|_2 \leq 10\sqrt{K}\right\}, \tag{C.41}$$

which holds with probability at least $1 - O(m^{-100} + me^{-CK})$. The proof thereafter will be developed on this event. For any fixed unit vectors $\boldsymbol{u}$, $\boldsymbol{v} \in \mathbb{C}^K$ and $(\boldsymbol{h}, \boldsymbol{x})$ obeying the assumptions (C.5), one has

$$\sum_{j=1}^{m} \boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} (\boldsymbol{h} - \boldsymbol{h}^\star)(\boldsymbol{x} - \boldsymbol{x}^\star)^{\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{v}$$
$$= \sum_{j=1}^{m} \underbrace{\boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} (\boldsymbol{h} - \boldsymbol{h}^\star)(\boldsymbol{x} - \boldsymbol{x}^\star)^{\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{v}\, \mathbb{1}_{\left\{\max\left\{\left|\boldsymbol{b}_j^{\mathsf{H}}(\boldsymbol{h}-\boldsymbol{h}^\star)\right|, \left|(\boldsymbol{x}-\boldsymbol{x}^\star)^{\mathsf{H}}\boldsymbol{a}_j\right|\right\} \le 20C_{13}\frac{1}{\log^{3/2} m}\right\}}}_{=:W_j}.$$

Concentration. Consider any fixed vectors $\boldsymbol{u}$, $\boldsymbol{v}$ and $(\boldsymbol{h}, \boldsymbol{x})$. We seek to invoke the Bernstein inequality Vershynin [2018, Theorem 2.8.2] to control $\sum_{j=1}^{m} W_j$. We observe that

$$\begin{aligned}
\left\|W_j - \mathbb{E}\left[W_j\right]\right\|_{\psi_1} &\le C \left\|W_j\right\|_{\psi_1} \\
&\le C \left| \boldsymbol{b}_j^{\mathsf{H}} (\boldsymbol{h} - \boldsymbol{h}^\star)(\boldsymbol{x} - \boldsymbol{x}^\star)^{\mathsf{H}} \boldsymbol{a}_j\, \mathbb{1}_{\left\{\max\left\{\left|\boldsymbol{b}_j^{\mathsf{H}}(\boldsymbol{h}-\boldsymbol{h}^\star)\right|, \left|(\boldsymbol{x}-\boldsymbol{x}^\star)^{\mathsf{H}}\boldsymbol{a}_j\right|\right\} \le 20C_{13}\frac{1}{\log^{3/2} m}\right\}} \right| \left\|\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\right\|_{\psi_2} \left\|\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\right\|_{\psi_2} \\
&= C \left| \boldsymbol{b}_j^{\mathsf{H}} (\boldsymbol{h} - \boldsymbol{h}^\star)(\boldsymbol{x} - \boldsymbol{x}^\star)^{\mathsf{H}} \boldsymbol{a}_j\, \mathbb{1}_{\left\{\max\left\{\left|\boldsymbol{b}_j^{\mathsf{H}}(\boldsymbol{h}-\boldsymbol{h}^\star)\right|, \left|(\boldsymbol{x}-\boldsymbol{x}^\star)^{\mathsf{H}}\boldsymbol{a}_j\right|\right\} \le 20C_{13}\frac{1}{\log^{3/2} m}\right\}} \right| \\
&\le 400 C C_{13}^2 \frac{1}{\log^3 m},
\end{aligned}$$

where the first inequality comes from the fact that $\|X - \mathbb{E}[X]\|_{\psi_1} \le C\|X\|_{\psi_1}$ (cf. Vershynin [2018, Section 2.7]), the second one is due to Vershynin [2018, Lemma 2.7.7] and the last inequality is due to the event $\mathcal{E}$. Hence, the Bernstein inequality Vershynin [2018, Theorem 2.8.2] reveals that

$$\mathbb{P}\left(\left|\sum_{j=1}^{m}\left(W_j - \mathbb{E}\left[W_j\right]\right)\right| \ge t\right) \le 2\exp\left(-c\min\left(\frac{t^2\log^6 m}{m}, t\log^3 m\right)\right).$$

Letting $t = C_t\sqrt{mK}$ for some large enough constant $C_t > 0$, we obtain that

$$\left|\sum_{j=1}^{m}\left(W_j - \mathbb{E}\left[W_j\right]\right)\right| \le C_t\sqrt{mK}, \tag{C.42}$$

holds with probability exceeding $1 - 2\exp(-cC_t^2 K\log m)$.

Union bound. Next, we define $\mathcal{N}_{\boldsymbol{z}}$ to be an ϵ_1-net of $\mathcal{B}_{\boldsymbol{z}}(\delta) := \{(\boldsymbol{h}, \boldsymbol{x}) : \max\{\|\boldsymbol{h} - \boldsymbol{h}^\star\|_2, \|\boldsymbol{x} - \boldsymbol{x}^\star\|_2\} \le \delta\}$ and $\mathcal{N}_0$ an ϵ_2-net of the unit sphere $\mathcal{S}^{K-1}$. In view of Vershynin [2018, Corollary 4.2.13], we have

$$|\mathcal{N}_{\boldsymbol{z}}| \le \left(1 + \frac{2\delta}{\epsilon_1}\right)^{4K} \qquad \text{and} \qquad |\mathcal{N}_0| \le \left(1 + \frac{2}{\epsilon_2}\right)^{2K}.$$

Taking this collectively with the union bound yields that (C.42) holds uniformly for any $(\boldsymbol{h}, \boldsymbol{x}) \in \mathcal{N}_{\boldsymbol{z}}$ and $\boldsymbol{u}$, $\boldsymbol{v} \in \mathcal{N}_0$ with probability over

$$1 - \left(1 + \frac{2\delta}{\epsilon_1}\right)^{4K}\left(1 + \frac{2}{\epsilon_2}\right)^{4K} \cdot 2\exp\left(-CK\log m\right) \ge 1 - 2\exp(-CK\log m).$$

Approximation. Define

$$\boldsymbol{H}_j(\boldsymbol{h}, \boldsymbol{x}) := \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} (\boldsymbol{h} - \boldsymbol{h}^\star)(\boldsymbol{x} - \boldsymbol{x}^\star)^{\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}}\, \mathbb{1}_{\left\{\max\left\{\left|\boldsymbol{b}_j^{\mathsf{H}}(\boldsymbol{h}-\boldsymbol{h}^\star)\right|, \left|(\boldsymbol{x}-\boldsymbol{x}^\star)^{\mathsf{H}}\boldsymbol{a}_j\right|\right\} \le 20C_{13}\frac{1}{\log^{3/2} m}\right\}}.$$

For any $(\boldsymbol{h}, \boldsymbol{x})$ satisfying the assumptions (C.5) and any $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{S}^{K-1}$, one can choose $(\boldsymbol{h}_0, \boldsymbol{x}_0) \in \mathcal{N}_{\boldsymbol{z}}$, $\boldsymbol{u}_0 \in \mathcal{N}_0$ and $\boldsymbol{v}_0 \in \mathcal{N}_0$ satisfying $\max\{\|\boldsymbol{h}-\boldsymbol{h}_0\|_2, \|\boldsymbol{x}-\boldsymbol{x}_0\|_2\} \le \epsilon_1$ and $\max\{\|\boldsymbol{u}-\boldsymbol{u}_0\|_2, \|\boldsymbol{v}-\boldsymbol{v}_0\|_2\} \le \epsilon_2$. Let

$$g(\boldsymbol{u},\boldsymbol{v},\boldsymbol{h},\boldsymbol{x}) \coloneqq \sum_{j=1}^{m} \boldsymbol{u}^{\mathsf{H}} \boldsymbol{H}_j(\boldsymbol{h},\boldsymbol{x})\,\boldsymbol{v} - m(\boldsymbol{h}-\boldsymbol{h}^\star)(\boldsymbol{x}-\boldsymbol{x}^\star)^{\mathsf{H}}.$$

Set $\epsilon_1 = \delta/K$ and $\epsilon_2 = 1/4$. In view of the triangle inequality, one has

$$\begin{aligned}
&|g(\boldsymbol{u},\boldsymbol{v},\boldsymbol{h},\boldsymbol{x}) - g(\boldsymbol{u}_0,\boldsymbol{v}_0,\boldsymbol{h}_0,\boldsymbol{x}_0)| \\
&\quad\le |g(\boldsymbol{u},\boldsymbol{v},\boldsymbol{h},\boldsymbol{x}) - g(\boldsymbol{u}_0,\boldsymbol{v},\boldsymbol{h},\boldsymbol{x})| + |g(\boldsymbol{u}_0,\boldsymbol{v},\boldsymbol{h},\boldsymbol{x}) - g(\boldsymbol{u}_0,\boldsymbol{v}_0,\boldsymbol{h},\boldsymbol{x})| \\
&\qquad + |g(\boldsymbol{u}_0,\boldsymbol{v}_0,\boldsymbol{h},\boldsymbol{x}) - g(\boldsymbol{u}_0,\boldsymbol{v}_0,\boldsymbol{h}_0,\boldsymbol{x})| + |g(\boldsymbol{u}_0,\boldsymbol{v}_0,\boldsymbol{h}_0,\boldsymbol{x}) - g(\boldsymbol{u}_0,\boldsymbol{v}_0,\boldsymbol{h}_0,\boldsymbol{x}_0)| \\
&\quad\le 2\epsilon_2 \left\| \sum_{j=1}^{m} \boldsymbol{H}_j(\boldsymbol{h},\boldsymbol{x}) - m(\boldsymbol{h}-\boldsymbol{h}^\star)(\boldsymbol{x}-\boldsymbol{x}^\star)^{\mathsf{H}} \right\| \\
&\qquad + |g(\boldsymbol{u}_0,\boldsymbol{v}_0,\boldsymbol{h},\boldsymbol{x}) - g(\boldsymbol{u}_0,\boldsymbol{v}_0,\boldsymbol{h}_0,\boldsymbol{x})| + |g(\boldsymbol{u}_0,\boldsymbol{v}_0,\boldsymbol{h}_0,\boldsymbol{x}) - g(\boldsymbol{u}_0,\boldsymbol{v}_0,\boldsymbol{h}_0,\boldsymbol{x}_0)|. \qquad \text{(C.43)}
\end{aligned}$$

To simplify the last two terms, we observe that

$$\left|(\boldsymbol{x}-\boldsymbol{x}_0)^{\mathsf{H}}\boldsymbol{a}_j\right| \le \max_{1\le j\le m}\|\boldsymbol{a}_j\|_2 \cdot \|\boldsymbol{x}-\boldsymbol{x}_0\|_2 \le 10\sqrt{K}\epsilon_1 \le C_{13}\frac{1}{\log^{3/2} m}, \qquad \text{(C.44)}$$

and furthermore,

$$\begin{aligned}
\left|(\boldsymbol{x}_0-\boldsymbol{x}^\star)^{\mathsf{H}}\boldsymbol{a}_j\right| &\le \left|(\boldsymbol{x}-\boldsymbol{x}^\star)^{\mathsf{H}}\boldsymbol{a}_j\right| + \left|(\boldsymbol{x}-\boldsymbol{x}_0)^{\mathsf{H}}\boldsymbol{a}_j\right| \\
&\le \left|(\boldsymbol{x}-\boldsymbol{x}^\star)^{\mathsf{H}}\boldsymbol{a}_j\right| + C_{13}\frac{1}{\log^{3/2} m} \\
&\le 3C_{13}\frac{1}{\log^{3/2} m}.
\end{aligned}$$

Similarly the same bounds also hold for $|\boldsymbol{b}_j^{\mathsf{H}}(\boldsymbol{h}_0-\boldsymbol{h}^\star)|$. It follows that

$$\mathbb{1}_{\left\{\left|(\boldsymbol{x}-\boldsymbol{x}_0)^{\mathsf{H}}\boldsymbol{a}_j\right| \le 20C_{13}\frac{1}{\log^{3/2} m}\right\}} = \mathbb{1}_{\left\{\left|(\boldsymbol{x}_0-\boldsymbol{x}^\star)^{\mathsf{H}}\boldsymbol{a}_j\right| \le 20C_{13}\frac{1}{\log^{3/2} m}\right\}} = \mathbb{1}_{\left\{\left|(\boldsymbol{x}-\boldsymbol{x}^\star)^{\mathsf{H}}\boldsymbol{a}_j\right| \le 20C_{13}\frac{1}{\log^{3/2} m}\right\}} = 1, \qquad \text{(C.45)}$$

$$\mathbb{1}_{\left\{\left|\boldsymbol{b}_j^{\mathsf{H}}(\boldsymbol{h}-\boldsymbol{h}_0)\right| \le 20C_{13}\frac{1}{\log^{3/2} m}\right\}} = \mathbb{1}_{\left\{\left|\boldsymbol{b}_j^{\mathsf{H}}(\boldsymbol{h}_0-\boldsymbol{h}^\star)\right| \le 20C_{13}\frac{1}{\log^{3/2} m}\right\}} = \mathbb{1}_{\left\{\left|\boldsymbol{b}_j^{\mathsf{H}}(\boldsymbol{h}-\boldsymbol{h}^\star)\right| \le 20C_{13}\frac{1}{\log^{3/2} m}\right\}} = 1. \qquad \text{(C.46)}$$

Then, we can bound the last two term in (C.43) as follows

$$\begin{aligned}
&|g(\boldsymbol{u}_0,\boldsymbol{v}_0,\boldsymbol{h},\boldsymbol{x}) - g(\boldsymbol{u}_0,\boldsymbol{v}_0,\boldsymbol{h}_0,\boldsymbol{x})| + |g(\boldsymbol{u}_0,\boldsymbol{v}_0,\boldsymbol{h}_0,\boldsymbol{x}) - g(\boldsymbol{u}_0,\boldsymbol{v}_0,\boldsymbol{h}_0,\boldsymbol{x}_0)| \\
&\quad\le \left\| \sum_{j=1}^{m} \boldsymbol{H}_j(\boldsymbol{h}-\boldsymbol{h}_0+\boldsymbol{h}^\star,\boldsymbol{x}) - m(\boldsymbol{h}-\boldsymbol{h}_0)(\boldsymbol{x}-\boldsymbol{x}^\star)^{\mathsf{H}} \right\| \\
&\qquad + \left\| \sum_{j=1}^{m} \boldsymbol{H}_j(\boldsymbol{h},\boldsymbol{x}-\boldsymbol{x}_0+\boldsymbol{x}^\star) - m(\boldsymbol{h}_0-\boldsymbol{h}^\star)(\boldsymbol{x}-\boldsymbol{x}_0)^{\mathsf{H}} \right\|.
\end{aligned}$$

Considering $g(\boldsymbol{u}_0,\boldsymbol{v}_0,\boldsymbol{h}_0,\boldsymbol{x}_0)$, one has

$$\begin{aligned}
&|g(\boldsymbol{u}_0,\boldsymbol{v}_0,\boldsymbol{h}_0,\boldsymbol{x}_0)| \\
&\quad\le \left|\sum_{j=1}^{m}(W_j - \mathbb{E}[W_j])\right| + \left|\sum_{j=1}^{m}\left(\mathbb{E}[W_j] - m\boldsymbol{u}^{\mathsf{H}}(\boldsymbol{h}_0-\boldsymbol{h}^\star)(\boldsymbol{x}_0-\boldsymbol{x}^\star)^{\mathsf{H}}\boldsymbol{v}\right)\right|
\end{aligned}$$

$$\le C_t\sqrt{mK} + \left|\sum_{j=1}^{m}\mathbb{E}\left[\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{h}_0-\boldsymbol{h}^{\star}\right)\left(\boldsymbol{x}_0-\boldsymbol{x}^{\star}\right)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{\max\left\{\left|\boldsymbol{b}_j^{\mathsf{H}}(\boldsymbol{h}_0-\boldsymbol{h}^{\star})\right|,\left|(\boldsymbol{x}_0-\boldsymbol{x}^{\star})^{\mathsf{H}}\boldsymbol{a}_j\right|\right\}>20C_{13}\frac{1}{\log^{3/2}m}\right\}}\right]\right|$$
$$= C_t\sqrt{mK}, \tag{C.47}$$

where the first inequality is due to triangle inequality; the second comes from (C.42) and the last is because of (C.46).

Putting all this together. Let

$$\mathcal{S}' := \left\{(\boldsymbol{h},\boldsymbol{x}) : \max\left\{\left|\left(\boldsymbol{x}-\boldsymbol{x}^{\star}\right)^{\mathsf{H}}\boldsymbol{a}_j\right|,\left|\left(\boldsymbol{h}-\boldsymbol{h}^{\star}\right)^{\mathsf{H}}\boldsymbol{b}_j\right|\right\} \le 20C_{13}\frac{1}{\log^{3/2}m}, \max\left\{\left\|\boldsymbol{h}-\boldsymbol{h}^{\star}\right\|_2,\left\|\boldsymbol{x}-\boldsymbol{x}^{\star}\right\|_2\right\}\le\delta\right\}.$$

It is easy to check that $(\boldsymbol{h}, 5(\boldsymbol{x}-\boldsymbol{x}_0)+\boldsymbol{x}^{\star}) \in \mathcal{S}$ by using the facts that $\|\boldsymbol{x}-\boldsymbol{x}_0\|_2 \le \epsilon_1$ and (C.44). Hence, we have

$$\left\|\sum_{j=1}^{m}\boldsymbol{H}_j\left(\boldsymbol{h},\boldsymbol{x}-\boldsymbol{x}_0+\boldsymbol{x}^{\star}\right) - m\left(\boldsymbol{h}-\boldsymbol{h}^{\star}\right)\left(\boldsymbol{x}-\boldsymbol{x}_0\right)^{\mathsf{H}}\right\|$$
$$\le \frac{1}{5}\sup_{(\boldsymbol{h},\boldsymbol{x})\in\mathcal{S}'}\left\|\sum_{j=1}^{m}\boldsymbol{H}_j\left(\boldsymbol{h},\boldsymbol{x}\right) - m\left(\boldsymbol{h}-\boldsymbol{h}^{\star}\right)\left(\boldsymbol{x}-\boldsymbol{x}^{\star}\right)^{\mathsf{H}}\right\|.$$

Similarly, one has $(5(\boldsymbol{h}-\boldsymbol{h}_0)+\boldsymbol{h}^{\star},\boldsymbol{x}) \in \mathcal{S}$ and therefore,

$$\left\|\sum_{j=1}^{m}\boldsymbol{H}_j\left(\boldsymbol{h}-\boldsymbol{h}_0+\boldsymbol{h}^{\star},\boldsymbol{x}\right) - m\left(\boldsymbol{h}-\boldsymbol{h}_0\right)\left(\boldsymbol{x}-\boldsymbol{x}^{\star}\right)^{\mathsf{H}}\right\|$$
$$\le \frac{1}{5}\sup_{(\boldsymbol{h},\boldsymbol{x})\in\mathcal{S}'}\left\|\sum_{j=1}^{m}\boldsymbol{H}_j\left(\boldsymbol{h},\boldsymbol{x}\right) - m\left(\boldsymbol{h}-\boldsymbol{h}^{\star}\right)\left(\boldsymbol{x}-\boldsymbol{x}^{\star}\right)^{\mathsf{H}}\right\|.$$

Hence, combining the above two inequalities with (C.43) and (C.47) reveals that

$$\left|g\left(\boldsymbol{u},\boldsymbol{v},\boldsymbol{h},\boldsymbol{x}\right)\right| \le \left|g\left(\boldsymbol{u}_0,\boldsymbol{v}_0,\boldsymbol{h}_0,\boldsymbol{x}_0\right)\right| + \left|g\left(\boldsymbol{u},\boldsymbol{v},\boldsymbol{h},\boldsymbol{x}\right)-g\left(\boldsymbol{u}_0,\boldsymbol{v}_0,\boldsymbol{h}_0,\boldsymbol{x}_0\right)\right|$$
$$\le C_t\sqrt{mK} + 2\epsilon_2\left\|\sum_{j=1}^{m}\boldsymbol{H}_j\left(\boldsymbol{h},\boldsymbol{x}\right) - m\left(\boldsymbol{h}-\boldsymbol{h}^{\star}\right)\left(\boldsymbol{x}-\boldsymbol{x}^{\star}\right)^{\mathsf{H}}\right\|$$
$$+\frac{2}{5}\sup_{(\boldsymbol{h},\boldsymbol{x})\in\mathcal{S}'}\left\|\sum_{j=1}^{m}\boldsymbol{H}_j\left(\boldsymbol{h},\boldsymbol{x}\right) - m\left(\boldsymbol{h}-\boldsymbol{h}^{\star}\right)\left(\boldsymbol{x}-\boldsymbol{x}^{\star}\right)^{\mathsf{H}}\right\|.$$

Taking the maximum over $\boldsymbol{u}$ and $\boldsymbol{v}$ on the left-hand side of the above inequality and rearranging terms yield

$$\left(1-2\epsilon_2\right)\left\|\sum_{j=1}^{m}\boldsymbol{H}_j\left(\boldsymbol{h},\boldsymbol{x}\right) - m\left(\boldsymbol{h}-\boldsymbol{h}^{\star}\right)\left(\boldsymbol{x}-\boldsymbol{x}^{\star}\right)^{\mathsf{H}}\right\|$$
$$\le C_t\sqrt{mK} + \frac{2}{5}\sup_{(\boldsymbol{h},\boldsymbol{x})\in\mathcal{S}'}\left\|\sum_{j=1}^{m}\boldsymbol{H}_j\left(\boldsymbol{h},\boldsymbol{x}\right) - m\left(\boldsymbol{h}-\boldsymbol{h}^{\star}\right)\left(\boldsymbol{x}-\boldsymbol{x}^{\star}\right)^{\mathsf{H}}\right\|.$$

Further taking the maximum over $(\boldsymbol{h},\boldsymbol{x})$ on $\mathcal{S}'$ gives

$$\left(1-2\epsilon_2-\frac{2}{5}\right)\sup_{(\boldsymbol{h},\boldsymbol{x})\in\mathcal{S}'}\left\|\sum_{j=1}^{m}\boldsymbol{H}_j\left(\boldsymbol{h},\boldsymbol{x}\right) - m\left(\boldsymbol{h}-\boldsymbol{h}^{\star}\right)\left(\boldsymbol{x}-\boldsymbol{x}^{\star}\right)^{\mathsf{H}}\right\| \le C_t\sqrt{mK},$$

and then rearranging terms yields

$$\sup_{(\boldsymbol{h},\boldsymbol{x})\in\mathcal{S}'}\left\|\sum_{j=1}^{m}\boldsymbol{H}_j\left(\boldsymbol{h},\boldsymbol{x}\right)\right\| \le \sup_{(\boldsymbol{h},\boldsymbol{x})\in\mathcal{S}'}\left\|m\left(\boldsymbol{h}-\boldsymbol{h}^\star\right)\left(\boldsymbol{x}-\boldsymbol{x}^\star\right)^{\mathsf{H}}\right\| + 4C_t\sqrt{mK}$$
$$\le \delta^2 m + 4C_t\sqrt{mK}.$$

Recognizing that

$$\sup_{(\boldsymbol{h},\boldsymbol{x})\in\mathcal{S}'}\left\|\sum_{j=1}^{m}\boldsymbol{H}_j\left(\boldsymbol{h},\boldsymbol{x}\right)\right\| = \sup_{(\boldsymbol{h},\boldsymbol{x})\in\mathcal{S}'}\left\|\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{h}-\boldsymbol{h}^\star\right)\left(\boldsymbol{x}-\boldsymbol{x}^\star\right)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\right\|$$

and that the set of all $(\boldsymbol{h},\boldsymbol{x})$ obeying (C.5) is a subset of $\mathcal{S}'$, we have established the desired result.

C.3 Proof of Lemma 22

The proof is very much the same as that of Lemma 6, except that the contraction coefficient in the expression $\boldsymbol{\nu}_1$ in (A.34) is $(1-c_\eta)$ rather than $(1-\eta)$ and the bound on $\boldsymbol{\nu}_3$ is different. In what follows, we shall only describe how to bound $\boldsymbol{\nu}_3$ here, for the sake of brevity.

The proof proceeds by bounding $\boldsymbol{\nu}_3$ via the four terms as indicated by (A.35a), which we discuss as follows.

1. For the first term ν_{31}, one has

$$\begin{aligned}\nu_{31} &\le \left|\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\widehat{\boldsymbol{x}}^{t,(l)\mathsf{H}}\boldsymbol{a}_l - \boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_l\right|\left\|\boldsymbol{b}_l\right\|_2\left|\boldsymbol{a}_l^{\mathsf{H}}\widehat{\boldsymbol{x}}^{t,(l)}\right| \\ &\le \left|\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\widehat{\boldsymbol{x}}^{t,(l)\mathsf{H}}\boldsymbol{a}_l - \boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_l\right|\cdot 10\sqrt{K}\cdot 20\sqrt{\log m}\cdot\left\|\widehat{\boldsymbol{x}}^{t,(l)}\right\|_2 \\ &\le 400\sqrt{K\log m}\left|\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\widehat{\boldsymbol{x}}^{t,(l)\mathsf{H}}\boldsymbol{a}_l - \boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_l\right|,\end{aligned} \tag{C.48}$$

where the penultimate inequality follows from (F.1) and (F.2); the last inequality is due to (C.4f).

2. Regarding ν_{32}, one has

$$\begin{aligned}\nu_{32} &\le \left|\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\widehat{\boldsymbol{x}}^{t,(l)\mathsf{H}}\boldsymbol{a}_l - \boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_l\right|\left\|\boldsymbol{a}_l\right\|_2\left|\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\right| \\ &\le \left|\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\widehat{\boldsymbol{x}}^{t,(l)\mathsf{H}}\boldsymbol{a}_l - \boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_l\right|\cdot 10\sqrt{K}\cdot 20\sqrt{\log m}\left\|\widehat{\boldsymbol{h}}^{t,(l)}\right\|_2 \\ &\le 400\sqrt{K\log m}\left|\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\widehat{\boldsymbol{x}}^{t,(l)\mathsf{H}}\boldsymbol{a}_l - \boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_l\right|,\end{aligned} \tag{C.49}$$

where the second line follows from (F.1) and (F.2); the last inequality is due to (C.4f). Further for some sufficiently large constant $C>0$, there holds

$$\begin{aligned}\left|\boldsymbol{b}_l^{\mathsf{H}}\left(\widehat{\boldsymbol{h}}^{t,(l)}-\boldsymbol{h}^\star\right)\right| &\le 20\sqrt{\log m}\left\|\widehat{\boldsymbol{h}}^{t,(l)}-\boldsymbol{h}^\star\right\|_2 \\ &\le 20\sqrt{\log m}\left(\left\|\widehat{\boldsymbol{h}}^{t,(l)}-\widetilde{\boldsymbol{h}}^{t}\right\|_2+\left\|\widetilde{\boldsymbol{h}}^{t}-\boldsymbol{h}^\star\right\|_2\right) \\ &\le C\left(\frac{\sqrt{mK\log^3 m}}{m}+\frac{\sigma\sqrt{K\log^2 m}}{m}\right),\end{aligned} \tag{C.50}$$

where the last inequality comes from (C.2b) and (C.4a). Similarly we can see this bound also holds for $|(\widehat{\boldsymbol{x}}^{t,(l)}-\boldsymbol{x}^\star)^{\mathsf{H}}\boldsymbol{a}_l|$. Therefore,

$$\left|\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\widehat{\boldsymbol{x}}^{t,(l)\mathsf{H}}\boldsymbol{a}_l - \boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_l\right|$$

$$
\begin{aligned}
&\le \left| \boldsymbol{b}_l^{\mathsf{H}} \widehat{\boldsymbol{h}}^{t,(l)} \left(\widehat{\boldsymbol{x}}^{t,(l)} - \boldsymbol{x}^{\star} \right)^{\mathsf{H}} \boldsymbol{a}_l \right| + \left| \boldsymbol{b}_l^{\mathsf{H}} \big(\widehat{\boldsymbol{h}}^{t,(l)} - \boldsymbol{h}^{\star} \big) \boldsymbol{x}^{\star \mathsf{H}} \boldsymbol{a}_l \right| \\
&\le \left(\left| \boldsymbol{b}_l^{\mathsf{H}} \big(\widehat{\boldsymbol{h}}^{t,(l)} - \boldsymbol{h}^{\star} \big) \right| + \left| \boldsymbol{b}_l^{\mathsf{H}} \boldsymbol{h}^{\star} \right| \right) \cdot \left| \left(\widehat{\boldsymbol{x}}^{t,(l)} - \boldsymbol{x}^{\star} \right)^{\mathsf{H}} \boldsymbol{a}_l \right| + \left| \boldsymbol{b}_l^{\mathsf{H}} \big(\widehat{\boldsymbol{h}}^{t,(l)} - \boldsymbol{h}^{\star} \big) \right| \cdot \left| \boldsymbol{x}^{\star \mathsf{H}} \boldsymbol{a}_l \right| \\
&\le \left(C \left(\frac{\sqrt{mK \log^3 m}}{m} + \frac{\sigma \sqrt{mK \log^2 m}}{m} \right) + 20 \sqrt{\log m} \right) \cdot C \left(\frac{\sqrt{mK \log^3 m}}{m} + \frac{\sigma \sqrt{K \log^2 m}}{m} \right) \\
&\quad + C \left(\frac{\sqrt{mK \log^3 m}}{m} + \frac{\sigma \sqrt{K \log^2 m}}{m} \right) \cdot 20 \sqrt{\log m} \qquad \text{(C.51)} \\
&\lesssim \frac{\sqrt{mK \log^4 m}}{m} + \frac{\sigma \sqrt{mK \log^3 m}}{m}, \qquad \text{(C.52)}
\end{aligned}
$$

where the penultimate inequality follows from (F.1) and (C.50). Substituting (C.52) into (C.48) and (C.49), we reach

$$
\begin{aligned}
\nu_{31} + \nu_{32} &\lesssim \sqrt{K \log m} \cdot \left(\frac{\sqrt{mK \log^4 m}}{m} + \frac{\sigma \sqrt{mK \log^3 m}}{m} \right) \\
&\le \frac{K \sqrt{m \log^5 m}}{m} + \frac{\sigma K \sqrt{m \log^4 m}}{m}. \qquad \text{(C.53a)}
\end{aligned}
$$

3. Regarding ν_{33} and ν_{34}, it can be seen that

$$
\nu_{33} = \left\| \xi_l \boldsymbol{b}_l \boldsymbol{a}_l^{\mathsf{H}} \widehat{\boldsymbol{x}}^{t,(l)} \right\|_2 \le |\xi_l| \, \|\boldsymbol{b}_l\|_2 \left| \boldsymbol{a}_l^{\mathsf{H}} \widehat{\boldsymbol{x}}^{t,(l)} \right| \overset{\text{(i)}}{\lesssim} \sigma \sqrt{K} \big\| \widehat{\boldsymbol{x}}^{t,(l)} \big\|_2 \log m \le 2\sigma \sqrt{K} \log m, \qquad \text{(C.53b)}
$$

$$
\nu_{34} = \left\| \overline{\xi_l} \boldsymbol{a}_l \boldsymbol{b}_l^{\mathsf{H}} \widehat{\boldsymbol{h}}^{t,(l)} \right\|_2 \le |\xi_l| \, \|\boldsymbol{a}_l\|_2 \left| \boldsymbol{b}_l^{\mathsf{H}} \widehat{\boldsymbol{h}}^{t,(l)} \right| \overset{\text{(i)}}{\lesssim} \sigma \sqrt{K} \big\| \widehat{\boldsymbol{h}}^{t,(l)} \big\|_2 \log m \le 2\sigma \sqrt{K} \log m, \qquad \text{(C.53c)}
$$

where (i) holds by (F.1), (F.2) and the independence between $\xi_l, \boldsymbol{a}_l$, $\boldsymbol{b}_l$ and $\widehat{\boldsymbol{x}}^{t,(l)}$.

Consequently, by (C.48) and (C.53a)-(C.53c) we have

$$
\|\boldsymbol{\nu}_3\|_2 \lesssim \frac{K \sqrt{m \log^5 m}}{m} + \sigma \sqrt{K} \log m, \qquad \text{(C.54)}
$$

as long as $m \gg K \log^2 m$. Then the proof follows the same line of idea as Appendix A.8, resulting in a similar inequality as (A.38) as follows:

$$
\begin{aligned}
\mathsf{dist}\big(\boldsymbol{z}^{t+1,(l)}, \widetilde{\boldsymbol{z}}^{t+1} \big) &\le (1 - c_\eta) \, \mathsf{dist}\big(\boldsymbol{z}^{t,(l)}, \widetilde{\boldsymbol{z}}^{t} \big) + \eta C \left(\frac{K \sqrt{m \log^5 m}}{m} + \sigma \sqrt{K} \log m \right) \\
&\le C \left(\frac{\sqrt{K \log^3 m}}{m} + \frac{\sigma \sqrt{K \log^2 m}}{m} \right),
\end{aligned}
$$

provided that $\eta = c_\eta / m$ with $c_\eta > 0$ being some sufficiently small constant. The proof for (C.8) follows from the same argument leading to (A.40) and is thus omitted here for simplicity.

C.4 Proof of Lemma 25

Recall the definition of $\boldsymbol{M}$ and $\boldsymbol{M}^{(l)}$ under the Gaussian design:

$$
\boldsymbol{M} := \frac{1}{m} \sum_{j=1}^{m} y_j \boldsymbol{b}_j \boldsymbol{a}_j^{\mathsf{H}}, \qquad \text{and} \qquad \boldsymbol{M}^{(l)} := \frac{1}{m} \sum_{j \neq l} y_j \boldsymbol{b}_j \boldsymbol{a}_j^{\mathsf{H}}.
$$

Applying Wedin's sinΘ theorem Dopico [2000, Theorem 2.1] gives that for some universal constant $C' > 0$, there holds

$$\min_{\alpha\in\mathbb{C},|\alpha|=1}\left\{\left\|\alpha\check{\boldsymbol{h}}^{0}-\check{\boldsymbol{h}}^{0,(l)}\right\|_{2}+\left\|\alpha\check{\boldsymbol{x}}^{0}-\check{\boldsymbol{x}}^{0,(l)}\right\|_{2}\right\}\leq C'\frac{\left\|\left(\boldsymbol{M}-\boldsymbol{M}^{(l)}\right)\check{\boldsymbol{x}}^{0,(l)}\right\|_{2}+\left\|\check{\boldsymbol{h}}^{0,(l)\mathsf{H}}\left(\boldsymbol{M}-\boldsymbol{M}^{(l)}\right)\right\|_{2}}{\sigma_{1}\left(\boldsymbol{M}^{(l)}\right)-\sigma_{2}\left(\boldsymbol{M}\right)}.$$

By invoking Weyl's inequality, we obtain

$$\begin{aligned}\sigma_{1}\left(\boldsymbol{M}^{(l)}\right)-\sigma_{2}\left(\boldsymbol{M}\right)&\geq\sigma_{1}\left(\mathbb{E}\left[\boldsymbol{M}^{(l)}\right]\right)-\left\|\boldsymbol{M}^{(l)}-\mathbb{E}\left[\boldsymbol{M}^{(l)}\right]\right\|-\sigma_{2}\left(\mathbb{E}\left[\boldsymbol{M}\right]\right)-\left\|\boldsymbol{M}-\mathbb{E}\left[\boldsymbol{M}\right]\right\|\\&\overset{\text{(i)}}{\geq}\frac{3}{4}-\left\|\boldsymbol{M}^{(l)}-\mathbb{E}\left[\boldsymbol{M}^{(l)}\right]\right\|-\left\|\boldsymbol{M}-\mathbb{E}\left[\boldsymbol{M}\right]\right\|\overset{\text{(ii)}}{\geq}\frac{1}{2},\end{aligned}$$

where (i) is due to the facts that

$$\sigma_{1}\left(\mathbb{E}\left[\boldsymbol{M}^{(l)}\right]\right)=\sigma_{1}\left(\frac{m-1}{m}\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\right)\geq\frac{3}{4},\qquad\text{and}\qquad\sigma_{2}\left(\mathbb{E}\left[\boldsymbol{M}\right]\right)=\sigma_{2}\left(\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\right)=0,$$

and (ii) comes from Lemma 40. Hence, one has

$$\min_{\alpha\in\mathbb{C},|\alpha|=1}\left\{\left\|\alpha\check{\boldsymbol{h}}^{0}-\check{\boldsymbol{h}}^{0,(l)}\right\|_{2}+\left\|\alpha\check{\boldsymbol{x}}^{0}-\check{\boldsymbol{x}}^{0,(l)}\right\|_{2}\right\}\leq 2C'\left(\left\|\left(\boldsymbol{M}-\boldsymbol{M}^{(l)}\right)\check{\boldsymbol{x}}^{0,(l)}\right\|_{2}+\left\|\check{\boldsymbol{h}}^{0,(l)\mathsf{H}}\left(\boldsymbol{M}-\boldsymbol{M}^{(l)}\right)\right\|_{2}\right).\tag{C.55}$$

We are left with bounding the two terms on the right-hand side of (C.55).

- Regarding the first term on the right-hand side of (C.55), we have

$$\begin{aligned}\left\|\left(\boldsymbol{M}-\boldsymbol{M}^{(l)}\right)\check{\boldsymbol{x}}^{0,(l)}\right\|_{2}&=\left\|\frac{1}{m}\boldsymbol{b}_{l}\left(\boldsymbol{b}_{l}^{\mathsf{H}}\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_{l}+\xi_{l}\right)\boldsymbol{a}_{l}^{\mathsf{H}}\check{\boldsymbol{x}}^{0,(l)}\right\|_{2}\\&\leq\left\|\frac{1}{m}\boldsymbol{b}_{l}\boldsymbol{b}_{l}^{\mathsf{H}}\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_{l}\boldsymbol{a}_{l}^{\mathsf{H}}\check{\boldsymbol{x}}^{0,(l)}\right\|_{2}+\left\|\frac{1}{m}\xi_{l}\boldsymbol{b}_{l}\boldsymbol{a}_{l}^{\mathsf{H}}\check{\boldsymbol{x}}^{0,(l)}\right\|_{2}\\&=\frac{1}{m}\left\|\boldsymbol{b}_{l}\right\|_{2}\left|\boldsymbol{b}_{l}^{\mathsf{H}}\boldsymbol{h}^{\star}\right|\left|\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_{l}\right|\left|\boldsymbol{a}_{l}^{\mathsf{H}}\check{\boldsymbol{x}}^{0,(l)}\right|+\frac{1}{m}\left|\xi_{l}\right|\left|\boldsymbol{a}_{l}^{\mathsf{H}}\check{\boldsymbol{x}}^{0,(l)}\right|\left\|\boldsymbol{b}_{l}\right\|_{2}\\&\leq\frac{1}{m}\cdot 10\sqrt{K}\cdot\left(20\sqrt{\log m}\right)^{2}\cdot 20\sqrt{\log m}+\frac{1}{m}\cdot 20\sigma\sqrt{\log m}\cdot 20\sqrt{\log m}\cdot 10\sqrt{K}\\&\lesssim\frac{\sqrt{K\log^{3}m}}{m}+\frac{\sigma\sqrt{K\log^{2}m}}{m},\end{aligned}$$

  where the second inequality is due to the triangle inequality; the penultimate inequality comes from (F.1), (F.2) and the fact that with probability exceeding $1-O(m^{-100})$,

$$\max_{1\leq l\leq m}\left|\boldsymbol{a}_{l}^{\mathsf{H}}\check{\boldsymbol{x}}^{0,(l)}\right|\leq 20\sqrt{\log m},$$

  due to the independence between $\check{\boldsymbol{x}}^{0,(l)}$ and $\boldsymbol{a}_{l}$.

- The second term on the right-hand side of (C.55) can be bounded in a similar fashion as follows

$$\begin{aligned}\left\|\check{\boldsymbol{h}}^{0,(l)\mathsf{H}}\left(\boldsymbol{M}-\boldsymbol{M}^{(l)}\right)\right\|_{2}&=\frac{1}{m}\left\|\check{\boldsymbol{h}}^{0,(l)\mathsf{H}}\boldsymbol{b}_{l}\left(\boldsymbol{b}_{l}^{\mathsf{H}}\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_{l}+\xi_{l}\right)\boldsymbol{a}_{l}^{\mathsf{H}}\right\|_{2}\\&\leq\frac{1}{m}\left\|\check{\boldsymbol{h}}^{0,(l)\mathsf{H}}\boldsymbol{b}_{l}\boldsymbol{b}_{l}^{\mathsf{H}}\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_{l}\boldsymbol{a}_{l}^{\mathsf{H}}\right\|_{2}+\frac{1}{m}\left\|\xi_{l}\check{\boldsymbol{h}}^{0,(l)\mathsf{H}}\boldsymbol{b}_{l}\boldsymbol{a}_{l}^{\mathsf{H}}\right\|_{2}\\&=\frac{1}{m}\cdot\left|\check{\boldsymbol{h}}^{0,(l)\mathsf{H}}\boldsymbol{b}_{l}\right|\left|\boldsymbol{b}_{l}^{\mathsf{H}}\boldsymbol{h}^{\star}\right|\left|\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_{l}\right|\left\|\boldsymbol{a}_{l}^{\mathsf{H}}\right\|_{2}+\frac{1}{m}\left|\xi_{l}\right|\left|\check{\boldsymbol{h}}^{0,(l)\mathsf{H}}\boldsymbol{b}_{l}\right|\left\|\boldsymbol{a}_{l}^{\mathsf{H}}\right\|_{2}\\&\leq\frac{1}{m}\cdot 20\sqrt{\log m}\cdot\left(20\sqrt{\log m}\right)^{2}\cdot 10\sqrt{K}+\frac{1}{m}\cdot 20\sigma\sqrt{\log m}\cdot 20\sqrt{\log m}\cdot 10\sqrt{K}\\&\lesssim\frac{\sqrt{K\log^{3}m}}{m}+\frac{\sigma\sqrt{K\log^{2}m}}{m},\end{aligned}$$

where the penultimate inequality comes from (F.1), (F.2) and the fact that

$$\max_{1\le l\le m}\left|\check{\boldsymbol{h}}^{0,(l)\mathsf{H}}\boldsymbol{b}_l\right|\le 20\sqrt{\log m}$$

holds with probability exceeding $1-O(m^{-100})$ (due to the independence between $\check{\boldsymbol{h}}^{0,(l)}$ and $\boldsymbol{b}_l$). Plugging the above two bounds into (C.55) leads to

$$\min_{\alpha\in\mathbb{C},|\alpha|=1}\left\{\left\|\alpha\check{\boldsymbol{h}}^0-\check{\boldsymbol{h}}^{0,(l)}\right\|_2+\left\|\alpha\check{\boldsymbol{x}}^0-\check{\boldsymbol{x}}^{0,(l)}\right\|_2\right\}\le\widetilde{C}\left(\frac{\sqrt{K\log^3 m}}{m}+\frac{\sigma\sqrt{K\log^2 m}}{m}\right),$$

for some universal constant $\widetilde{C}>0$. To convert this bound into the desired version, we can employ the same argument connecting Ma et al. [2018, Eqn (240)] to Ma et al. [2018, Eqn (245)]. The details are omitted here for brevity.

D Analysis under Gaussian design: connections between convex and nonconvex solutions

D.1 Preliminaries

Here, we state below a few elementary technical lemmas that prove useful in the proof. To begin with, we show that the operator $\mathcal{A}$ is well-controlled in this case, whose counterpart in the Fourier design is Lemma 1.

Lemma 31. *For the operator $\mathcal{A}$ defined under the Gaussian setting, we have, with probability at least $1-O(m^{-10})$, that*

$$\|\mathcal{A}\|\le 10\sqrt{mK\log m}.$$

Proof. Denote

$$\boldsymbol{A}\coloneqq\begin{bmatrix}\boldsymbol{a}_1^\top\\ \vdots\\ \boldsymbol{a}_m^\top\end{bmatrix}\in\mathbb{C}^{m\times K},\qquad \boldsymbol{B}\coloneqq\begin{bmatrix}\boldsymbol{b}_1^\top\\ \vdots\\ \boldsymbol{b}_m^\top\end{bmatrix}\in\mathbb{C}^{m\times K}.$$

We can rewrite $\mathcal{A}$ in matrix form as follows

$$\mathcal{A}(\boldsymbol{Z})=\left\{\boldsymbol{b}_j^\mathsf{H}\boldsymbol{Z}\boldsymbol{a}_j\right\}_{j=1}^m=\begin{bmatrix}\mathrm{diag}(\boldsymbol{A}_{:,1})\boldsymbol{B} & \mathrm{diag}(\boldsymbol{A}_{:,2})\boldsymbol{B} & \cdots & \mathrm{diag}(\boldsymbol{A}_{:,K})\boldsymbol{B}\end{bmatrix}\mathrm{vec}(\boldsymbol{Z}).$$

This allows one to express and obtain

$$\begin{aligned}\|\mathcal{A}\|^2&=\left\|\begin{bmatrix}\mathrm{diag}(\boldsymbol{A}_{:,1})\boldsymbol{B} & \mathrm{diag}(\boldsymbol{A}_{:,2})\boldsymbol{B} & \cdots & \mathrm{diag}(\boldsymbol{A}_{:,K})\boldsymbol{B}\end{bmatrix}\right\|^2\\&\le\|\boldsymbol{B}\|^2\cdot\sum_{i=1}^K\|\mathrm{diag}(\boldsymbol{A}_{:,i})\|^2\\&\le\left\|\sum_{j=1}^m\boldsymbol{b}_j\boldsymbol{b}_j^\mathsf{H}\right\|\cdot K\max_{1\le i\le K}\max_{1\le j\le m}|\boldsymbol{A}_{i,j}|^2\\&\le 2m\cdot K\cdot 20\log m\end{aligned}$$

with probability at least $1-O(m^{-100})$. ☐

Next, the following lemma corresponds to Lemma 39 under the Fourier design. Its proof is deferred to Appendix D.3.

Lemma 32. *Suppose that T is the tangent space of $\boldsymbol{h}\boldsymbol{x}^\mathsf{H}$ with $\|\boldsymbol{h}\|_2=\|\boldsymbol{x}\|_2=1$ and $m\ge CK\log^2 m$ for some sufficiently large constant $C>0$. Then there exists some sufficiently large constant $C'>0$ such that*

$$\|\mathcal{P}_T\mathcal{A}^*\mathcal{A}\mathcal{P}_T-m\mathcal{P}_T\|\le C'\sqrt{mK\log m}$$

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

D.2 Proof of Theorem 3

In this section, we proceed to prove Theorem 3 by connecting the convex minimizer with nonconvex iterates, in the same vein as in the Fourier design case (cf. Appendix B). To begin with, a lemma stating the results of Algorithm 2 under the Gaussian design is listed below.

Lemma 33. *Take $\lambda = C_\lambda \sigma\sqrt{mK\log m}$ for some sufficiently large constant $C_\lambda > 0$. Suppose that Assumption 2 holds. Assume the number of measurements obeys $m \geq CK\log^6 m$ for some sufficiently large constant $C > 0$ and the noise satisfies $\sigma\sqrt{K\log^5/m} \leq c$ for some sufficiently small constant $c > 0$. Let stepsize η be c_η/m for some sufficiently small constant $c_\eta > 0$. Then, with probability at least $1 - O\left(m^{-100} + me^{-K}\right)$, the iterates $\{\boldsymbol{h}^t, \boldsymbol{x}^t\}_{0<t\leq t_0}$ of Algorithm 2 satisfy*

$$\mathsf{dist}\left(\boldsymbol{z}^t, \boldsymbol{z}^\star\right) \leq \rho\mathsf{dist}\left(\boldsymbol{z}^{t-1}, \boldsymbol{z}^\star\right) + C_{11}\eta\left(\lambda + \sigma\sqrt{mK\log m}\right), \tag{D.1a}$$

$$\mathsf{dist}\left(\boldsymbol{z}^{t,(l)}, \widetilde{\boldsymbol{z}}^t\right) \leq C_{12}\frac{\sigma\sqrt{K\log^2 m}}{m}, \tag{D.1b}$$

$$\max_{1\leq l\leq m}\left\|\widetilde{\boldsymbol{z}}^{t,(l)} - \widetilde{\boldsymbol{z}}^t\right\|_2 \lesssim C_{12}\frac{\sigma\sqrt{K\log^2 m}}{m}, \tag{D.1c}$$

$$\max_{1\leq l\leq m}\left|\boldsymbol{a}_l^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^t - \boldsymbol{x}^\star\right)\right| \leq C_{13}\frac{\sigma\sqrt{mK\log^2 m}}{m}, \tag{D.1d}$$

$$\max_{1\leq l\leq m}\left|\boldsymbol{b}_l^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^t - \boldsymbol{h}^\star\right)\right| \leq C_{13}\frac{\sigma\sqrt{mK\log^2 m}}{m} \tag{D.1e}$$

for any $0 < t \leq t_0$, where $\rho = 1 - c_\rho c_\eta$ for some small constant $c_\rho > 0$, and we take $t_0 = m^{20}$. Here, C_{11}, C_{12} and C_{13} are positive constants. Additionally, one has

$$\min_{0\leq t\leq t_0}\left\|\nabla f\left(\boldsymbol{h}^t, \boldsymbol{x}^t\right)\right\|_2 \leq \frac{\lambda}{m^{10}}. \tag{D.1f}$$

(D.1a)-(D.1e) can be seen as direct consequences from our analysis in Appendix C, while (D.1f) can be derived by following the proof in Appendix B.3.2. Hence, we do not repeat the proof here for brevity.

Similar to Conditions 1 and 2, we single out two critical conditions on the operators under Assumption 2. The first condition below requires the regularization parameter λ to be large enough, so as to dominate a certain form of noise and the deviation of $\mathcal{T}\left(\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right)$ from its mean $m\left(\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right)$.

Condition 3. The regularization parameter λ satisfies

1. $\left\|\mathcal{T}\left(\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right) - m\left(\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right)\right\| < \lambda/8$.
2. $\left\|\mathcal{A}^*\left(\boldsymbol{\xi}\right)\right\| \leq c\lambda$ for some small constant $c > 0$.

The second condition is concerned with the injectivity property of $\mathcal{A}$.

Condition 4. Let T be the tangent space of $\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}$. Then for all $\boldsymbol{Z} \in T$, one has

$$\left\|\mathcal{A}\left(\boldsymbol{Z}\right)\right\|_2^2 \geq \frac{m}{16}\left\|\boldsymbol{Z}\right\|_{\mathrm{F}}^2.$$

Armed with these two conditions, the following lemma reveals how an approximate nonconvex optimizer can serve as a proxy of the convex minimizer. The proof of this lemma can be developed in the same manner as in Appendix A.8; the details are omitted here for brevity.

Lemma 34. *Suppose that $(\boldsymbol{h}, \boldsymbol{x})$ obeys*

$$\left\|\nabla f\left(\boldsymbol{h}, \boldsymbol{x}\right)\right\|_2 \leq C\frac{\lambda}{m^{10}} \tag{D.2a}$$

for some constants $C > 0$. Then under Conditions 3 and 4, any minimizer $\boldsymbol{Z}_{\mathsf{cvx}}$ of the convex problem (3) satisfies

$$\left\|\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} - \boldsymbol{Z}_{\mathsf{cvx}}\right\|_{\mathrm{F}} \lesssim \left\|\nabla f\left(\boldsymbol{h}, \boldsymbol{x}\right)\right\|_2.$$

Consequently, the conclusions in Theorem 3 can be easily derived from Lemma 34 by similar calculations as proof of Theorem 1 in Appendix B.1, and thus omitted here for brevity.

It remains to demonstrate that Conditions 3 and 4 hold with high probability under the sample size and noise level conditions (13). We start with the first point in Condition 3. Its proof can be directly adaptated from the proof in Appendix B.5, and thus omitted here for simplicity.

Lemma 35. *Suppose that the sample complexity satisfies $m \geq CK \log^4 m$ for some sufficiently large constant $C > 0$. Take $\lambda = C_\lambda \sigma \sqrt{mK \log m}$ for some large enough constant $C_\lambda > 0$. Then with probability at least $1 - O\left(m^{-10} + me^{-CK}\right)$, we have*

$$\left\| \mathcal{T}\left(\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} - \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}}\right) - m\left(\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} - \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}}\right)\right\| < \lambda/8$$

simultaneously for any $(\boldsymbol{h}, \boldsymbol{x})$ obeying (B.5a) and (B.5b).

The next lemma corresponds to the second point in Condition 3.

Lemma 36. *Suppose that Assumption 2 holds and $m \geq CK \log^5 m$ for some sufficiently large constant $C > 0$. Then one has*

$$\left\| \mathcal{A}^*\left(\boldsymbol{\xi}\right)\right\| \lesssim \sigma \sqrt{mK \log m} \tag{D.3}$$

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

Proof. See Appendix F.1. ∎

Turning attention to Condition 4, we have the following lemma.

Lemma 37. *Suppose that the sample complexity satisfies $m \geq CK \log m$ for some sufficiently large constant $C > 0$. Then with probability at least $1 - O\left(m^{-10}\right)$,*

$$\left\|\mathcal{A}\left(\boldsymbol{Z}\right)\right\|_2^2 \geq \frac{m}{16}\left\|\boldsymbol{Z}\right\|_{\mathrm{F}}^2, \quad \forall \boldsymbol{Z} \in T$$

holds simultaneously for all T for which the associated point $(\boldsymbol{h}, \boldsymbol{x})$ obeys (B.5a) and (B.5b). Here, T denotes the tangent space of $\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}$.

The proof is a direct adaptation from Appendix B.7 and thus omitted for brevity.

D.3 Proof of Lemma 32

The framework and notation adopted here are similar to Ahmed et al. [2013, Section 5.2]. To facilitate the proof, we introduce an operator for $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathbb{C}^K$ as follows:

$$\boldsymbol{x}_1 \boldsymbol{y}_1^{\mathsf{H}} \otimes \boldsymbol{x}_2 \boldsymbol{y}_2^{\mathsf{H}} \coloneqq \left\{\overline{y_{1i}} y_{1k} \boldsymbol{x}_1 \boldsymbol{x}_2^{\mathsf{H}}\right\}_{i,k} \in \mathbb{C}^{K^2 \times K^2}.$$

Denote by $\boldsymbol{v}_j = \langle \boldsymbol{h}, \boldsymbol{b}_j \rangle \boldsymbol{a}_j$ and $\boldsymbol{u}_j = \langle \boldsymbol{x}, \boldsymbol{a}_j \rangle \left(\boldsymbol{I}_K - \boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\right) \boldsymbol{b}_j$. Then we can rewrite the operator $\mathcal{P}_T \mathcal{A}^* \mathcal{A} \mathcal{P}_T : \mathbb{C}^{K \times K} \to \mathbb{C}^{K \times K}$ as the following matrix

$$\mathcal{Q} \coloneqq \sum_{j=1}^m \left(\boldsymbol{h}\boldsymbol{v}_j^{\mathsf{H}} \otimes \boldsymbol{h}\boldsymbol{v}_j^{\mathsf{H}} + \boldsymbol{h}\boldsymbol{v}_j^{\mathsf{H}} \otimes \boldsymbol{u}_j \boldsymbol{x}^{\mathsf{H}} + \boldsymbol{u}_j \boldsymbol{x}^{\mathsf{H}} \otimes \boldsymbol{h}\boldsymbol{v}_j^{\mathsf{H}} + \boldsymbol{u}_j \boldsymbol{x}^{\mathsf{H}} \otimes \boldsymbol{u}_j \boldsymbol{x}^{\mathsf{H}}\right) \in \mathbb{C}^{K^2 \times K^2},$$

which satisfies

$$\mathrm{vec}\left(\mathcal{P}_T \mathcal{A}^* \mathcal{A} \mathcal{P}_T(\boldsymbol{X})\right) = \mathcal{Q}\mathrm{vec}\left(\boldsymbol{X}\right)$$

for any $\boldsymbol{X} \in \mathbb{C}^{K \times K}$. This implies that

$$\begin{aligned} &\left\|\mathcal{P}_T \mathcal{A}^* \mathcal{A} \mathcal{P}_T - m\mathcal{P}_T\right\| \\ &\quad = \left\|\mathcal{P}_T \mathcal{A}^* \mathcal{A} \mathcal{P}_T - \mathbb{E}\left[\mathcal{P}_T \mathcal{A}^* \mathcal{A} \mathcal{P}_T\right]\right\| = \left\|\mathcal{Q} - \mathbb{E}\left[\mathcal{Q}\right]\right\| \end{aligned}$$

$$\leq \underbrace{\left\|\sum_{j=1}^{m}\left(\boldsymbol{h}\boldsymbol{v}_j^{\mathsf{H}}\otimes \boldsymbol{h}\boldsymbol{v}_j^{\mathsf{H}} - \mathbb{E}\left[\boldsymbol{h}\boldsymbol{v}_j^{\mathsf{H}}\otimes \boldsymbol{h}\boldsymbol{v}_j^{\mathsf{H}}\right]\right)\right\|}_{\beta_1} + \underbrace{\left\|\sum_{j=1}^{m}\left(\boldsymbol{h}\boldsymbol{v}_j^{\mathsf{H}}\otimes \boldsymbol{u}_j\boldsymbol{x}^{\mathsf{H}} - \mathbb{E}\left[\boldsymbol{h}\boldsymbol{v}_j^{\mathsf{H}}\otimes \boldsymbol{u}_j\boldsymbol{x}^{\mathsf{H}}\right]\right)\right\|}_{\beta_2}$$

$$+ \underbrace{\left\|\sum_{j=1}^{m}\left(\boldsymbol{u}_j\boldsymbol{x}^{\mathsf{H}}\otimes \boldsymbol{h}\boldsymbol{v}_j^{\mathsf{H}} - \boldsymbol{u}_j\boldsymbol{x}^{\mathsf{H}}\otimes \boldsymbol{h}\boldsymbol{v}_j^{\mathsf{H}}\right)\right\|}_{\beta_3} + + \underbrace{\left\|\sum_{j=1}^{m}\left(\boldsymbol{u}_j\boldsymbol{x}^{\mathsf{H}}\otimes \boldsymbol{u}_j\boldsymbol{x}^{\mathsf{H}} - \mathbb{E}\left[\boldsymbol{u}_j\boldsymbol{x}^{\mathsf{H}}\otimes \boldsymbol{u}_j\boldsymbol{x}^{\mathsf{H}}\right]\right)\right\|}_{\beta_4} \tag{D.4}$$

In the sequel, we consider the four terms on the right-hand side of (D.4) separately.

Controlling β_1. Regarding the first term β_1, we denote

$$\boldsymbol{Z}_j := \boldsymbol{h}\boldsymbol{v}_j^{\mathsf{H}}\otimes \boldsymbol{h}\boldsymbol{v}_j^{\mathsf{H}} - \mathbb{E}\left[\boldsymbol{h}\boldsymbol{v}_j^{\mathsf{H}}\otimes \boldsymbol{h}\boldsymbol{v}_j^{\mathsf{H}}\right].$$

Then one has

$$\begin{aligned}
\left\|\left\|\left\|\boldsymbol{Z}_j\right\|\right\|\right\|_{\psi_1} &= \left\|\left\|\left\|\left\{\left(\left|\langle \boldsymbol{h}, \boldsymbol{b}_j\rangle\right|^2 \overline{a_{ji}}a_{jk} - \delta_{ik}\right)\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\right\}_{i,k}\right\|\right\|\right\|_{\psi_1} \\
&\leq \left\|\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\right\| \cdot \left\|\left\|\left\|\left|\langle \boldsymbol{h}, \boldsymbol{b}_j\rangle\right|^2 \boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}} - \boldsymbol{I}\right\|\right\|\right\|_{\psi_1} \\
&\overset{\text{(i)}}{\leq} \left\|\max\left\{\left|\langle \boldsymbol{h}, \boldsymbol{b}_j\rangle\right|^2 \cdot \left\|\boldsymbol{a}_j\right\|_2^2, 1\right\}\right\|_{\psi_1} \\
&\leq \left|\langle \boldsymbol{h}, \boldsymbol{b}_j\rangle\right|^2 \cdot \left\|\left\|\boldsymbol{a}_j\right\|_2\right\|_{\psi_2}^2 + 1 \\
&\overset{\text{(ii)}}{\leq} CK\log m,
\end{aligned}$$

where (i) is due to the fact that $\|\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\| = \|\boldsymbol{h}\|_2^2 = 1$; (ii) uses (F.1) and $\left\|\|\boldsymbol{b}_j\|_2\right\|_{\psi_2} \lesssim \sqrt{K}$ (cf. Vershynin [2018, Theorem 3.1.1]). To compute the variance term $\mathbb{E}[\boldsymbol{Z}_j^{\mathsf{H}}\boldsymbol{Z}_j]$ and $\mathbb{E}[\boldsymbol{Z}_j\boldsymbol{Z}_j^{\mathsf{H}}]$, we express the operation of $\boldsymbol{Z}_j$ on a matrix $\boldsymbol{X}$ as

$$\boldsymbol{Z}_j\left(\boldsymbol{X}\right) = \left|\langle \boldsymbol{h}, \boldsymbol{b}_j\rangle\right|^2 \boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\boldsymbol{X}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}} - \|\boldsymbol{h}\|_2^2\,\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\boldsymbol{X},$$

and hence

$$\boldsymbol{Z}_j^{\mathsf{H}}\boldsymbol{Z}_j\left(\boldsymbol{X}\right) = \left|\langle \boldsymbol{h}, \boldsymbol{b}_j\rangle\right|^4 \|\boldsymbol{h}\|_2^2\,\|\boldsymbol{a}_j\|_2^2\,\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\boldsymbol{X}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}} - 2\left|\langle \boldsymbol{h}, \boldsymbol{b}_j\rangle\right|^2 \|\boldsymbol{h}\|_2^4\,\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\boldsymbol{X}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}} + \|\boldsymbol{h}\|_2^6\,\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\boldsymbol{X}.$$

Then one has

$$\mathbb{E}\left[\boldsymbol{Z}_j^{\mathsf{H}}\boldsymbol{Z}_j\left(\boldsymbol{X}\right)\right] = 3\left(K+2\right)\|\boldsymbol{h}\|_2^4\,\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\boldsymbol{X} - 2\|\boldsymbol{h}\|_2^6\,\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\boldsymbol{X} + \|\boldsymbol{h}\|_2^6\,\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\boldsymbol{X} = \left(3K+5\right)\|\boldsymbol{h}\|_2^4\,\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\boldsymbol{X}.$$

Similarly, one can derive that

$$\mathbb{E}\left[\boldsymbol{Z}_j\boldsymbol{Z}_j^{\mathsf{H}}\left(\boldsymbol{X}\right)\right] = \mathbb{E}\left[\boldsymbol{Z}_j^{\mathsf{H}}\boldsymbol{Z}_j\left(\boldsymbol{X}\right)\right] = \left(3K+5\right)\|\boldsymbol{h}\|_2^4\,\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\boldsymbol{X},$$

thus indicating that

$$\sigma_{\boldsymbol{Z}} := \max\left\{\left\|\sum_{j=1}^{m}\mathbb{E}\left[\boldsymbol{Z}_j^{\mathsf{H}}\boldsymbol{Z}_j\right]\right\|^{1/2}, \left\|\sum_{j=1}^{m}\mathbb{E}\left[\boldsymbol{Z}_j\boldsymbol{Z}_j^{\mathsf{H}}\right]\right\|^{1/2}\right\} \leq \sqrt{\left(3K+5\right)m\,\|\boldsymbol{h}\|_2^6}.$$

By the matrix Bernstein inequality Koltchinskii et al. [2011, Proposition 2], one has

$$\left\|\sum_{j=1}^{m}\boldsymbol{Z}_j\right\| \lesssim \sigma_{\boldsymbol{Z}}\sqrt{\log m} + B_{\boldsymbol{Z}}\log\left(\frac{B_{\boldsymbol{Z}}\sqrt{m}}{\sigma_{\boldsymbol{Z}}}\right)\log m \lesssim \sqrt{mK\log m} \tag{D.5}$$

with high probability.

Controlling β_2. When it comes to the second term β_2, we first set

$$\boldsymbol{H}_j := \boldsymbol{h}\boldsymbol{v}_j^{\mathsf{H}} \otimes \boldsymbol{u}_j\boldsymbol{x}^{\mathsf{H}} - \mathbb{E}\left[\boldsymbol{h}\boldsymbol{v}_j^{\mathsf{H}} \otimes \boldsymbol{u}_j\boldsymbol{x}^{\mathsf{H}}\right],$$

which satisfies

$$\begin{aligned}
\|\boldsymbol{H}_j\| &= \left\|\left\{\overline{\langle \boldsymbol{h}, \boldsymbol{b}_j\rangle\langle \boldsymbol{x}, \boldsymbol{a}_j\rangle a_{ji}} x_k \boldsymbol{h}\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{I}_K - \boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\right)\right\}_{i,k}\right\| \\
&\leq \left\|\langle \boldsymbol{h}, \boldsymbol{b}_j\rangle \langle \boldsymbol{x}, \boldsymbol{a}_j\rangle \boldsymbol{a}_j\boldsymbol{x}^{\mathsf{H}}\right\| \cdot \left\|\boldsymbol{h}\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{I}_K - \boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\right)\right\| \\
&\leq |\langle \boldsymbol{h}, \boldsymbol{b}_j\rangle| \cdot \|\langle \boldsymbol{x}, \boldsymbol{a}_j\rangle \boldsymbol{a}_j\| \cdot \|\boldsymbol{x}\|_2 \cdot \|\boldsymbol{h}\|_2 \cdot \|\boldsymbol{b}_j\|_2 \cdot \left\|\boldsymbol{I}_K - \boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\right\| \\
&\leq |\langle \boldsymbol{h}, \boldsymbol{b}_j\rangle| \cdot |\langle \boldsymbol{x}, \boldsymbol{a}_j\rangle| \cdot \|\boldsymbol{a}_j\|_2 \cdot \|\boldsymbol{x}\|_2 \cdot \|\boldsymbol{h}\|_2 \cdot \|\boldsymbol{b}_j\|_2 \cdot \left\|\boldsymbol{I}_K - \boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\right\|
\end{aligned}$$

By employing $\|\boldsymbol{h}\|_2 = \|\boldsymbol{x}\|_2 = 1$, (F.1) and $\left\|\|\boldsymbol{a}_j\|_2\right\|_{\psi_2} = \left\|\|\boldsymbol{b}_j\|_2\right\|_{\psi_2} \lesssim \sqrt{K}$ (cf. Vershynin [2018, Theorem 3.1.1]), we obtain

$$\left\|\|\boldsymbol{H}_j\|\right\|_{\psi_1} \leq CK\log m.$$

Next, let us consider the operation of $\boldsymbol{H}_j$ and $\boldsymbol{H}_j^{\mathsf{H}}$ on $\boldsymbol{X}$, which obeys

$$\begin{aligned}
\boldsymbol{H}_j(\boldsymbol{X}) &= \overline{\langle \boldsymbol{h}, \boldsymbol{b}_j\rangle\langle \boldsymbol{x}, \boldsymbol{a}_j\rangle}\boldsymbol{h}\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{I}_K - \boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\right)\boldsymbol{X}\boldsymbol{x}\boldsymbol{a}_j^{\mathsf{H}}, \\
\boldsymbol{H}_j^{\mathsf{H}}(\boldsymbol{X}) &= \langle \boldsymbol{h}, \boldsymbol{b}_j\rangle \langle \boldsymbol{x}, \boldsymbol{a}_j\rangle \left(\boldsymbol{I}_K - \boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\right)\boldsymbol{b}_j\boldsymbol{h}^{\mathsf{H}}\boldsymbol{X}\boldsymbol{a}_j\boldsymbol{x}^{\mathsf{H}}.
\end{aligned}$$

Consequently, one can deduce that

$$\boldsymbol{H}_j\boldsymbol{H}_j^{\mathsf{H}}(\boldsymbol{X}) = |\langle \boldsymbol{h}, \boldsymbol{b}_j\rangle|^2 |\langle \boldsymbol{x}, \boldsymbol{a}_j\rangle|^2 \|\boldsymbol{x}\|_2^2 \boldsymbol{h}\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{I}_K - \boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\right)\boldsymbol{b}_j\boldsymbol{h}^{\mathsf{H}}\boldsymbol{X}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}},$$

and

$$\boldsymbol{H}_j^{\mathsf{H}}\boldsymbol{H}_j(\boldsymbol{X}) = |\langle \boldsymbol{h}, \boldsymbol{b}_j\rangle|^2 |\langle \boldsymbol{x}, \boldsymbol{a}_j\rangle|^2 \|\boldsymbol{a}_j\|_2^2 \|\boldsymbol{h}\|_2^2 \left(\boldsymbol{I}_K - \boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\right)\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{I}_K - \boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\right)\boldsymbol{X}\boldsymbol{x}\boldsymbol{x}^{\mathsf{H}}.$$

It follows that their expectations are

$$\mathbb{E}\left[\boldsymbol{H}_j\boldsymbol{H}_j^{\mathsf{H}}(\boldsymbol{X})\right] = \left[(K+2)\|\boldsymbol{h}\|_2^2 - 3\|\boldsymbol{h}\|_2^4\right]\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\boldsymbol{X}\left(2\boldsymbol{x}\boldsymbol{x}^{\mathsf{H}} + \|\boldsymbol{x}\|_2^2\boldsymbol{I}_K\right),$$

and

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{H}_j^{\mathsf{H}}\boldsymbol{H}_j(\boldsymbol{X})\right] &= (K+2)\|\boldsymbol{h}\|_2^2\|\boldsymbol{x}\|_2^2\left(\boldsymbol{I}_K - \boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\right)\left(2\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}} + \|\boldsymbol{h}\|_2^2\boldsymbol{I}_K\right)\left(\boldsymbol{I}_K - \boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\right)\boldsymbol{X}\boldsymbol{x}\boldsymbol{x}^{\mathsf{H}} \\
&= (K+2)\left(\boldsymbol{I}_K - \boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\right)\boldsymbol{X}\boldsymbol{x}\boldsymbol{x}^{\mathsf{H}}.
\end{aligned}$$

Hence, we have

$$\sigma_{\boldsymbol{Z}} := \max\left\{\left\|\sum_{j=1}^m \mathbb{E}\left[\boldsymbol{H}_j^{\mathsf{H}}\boldsymbol{H}_j\right]\right\|^{1/2}, \left\|\sum_{j=1}^m \mathbb{E}\left[\boldsymbol{H}_j\boldsymbol{H}_j^{\mathsf{H}}\right]\right\|^{1/2}\right\} \leq \sqrt{3mK}.$$

By the matrix Bernstein inequality Koltchinskii et al. [2011, Proposition 2], one has

$$\left\|\sum_{j=1}^m \boldsymbol{H}_j\right\| \lesssim \sigma_{\boldsymbol{Z}}\sqrt{\log m} + B_{\boldsymbol{Z}}\log\left(\frac{B_{\boldsymbol{Z}}\sqrt{m}}{\sigma_{\boldsymbol{Z}}}\right)\log m \lesssim \sqrt{mK\log m}. \tag{D.6}$$

Controlling β_3. When being written in matrix form, one has $\boldsymbol{u}_j\boldsymbol{x}^{\mathsf{H}} \otimes \boldsymbol{h}\boldsymbol{v}_j^{\mathsf{H}} - \mathbb{E}[\boldsymbol{u}_j\boldsymbol{x}^{\mathsf{H}} \otimes \boldsymbol{h}\boldsymbol{v}_j^{\mathsf{H}}]$ is the conjugate transpose of $\boldsymbol{h}\boldsymbol{v}_j^{\mathsf{H}} \otimes \boldsymbol{u}_j\boldsymbol{x}^{\mathsf{H}} - \mathbb{E}[\boldsymbol{h}\boldsymbol{v}_j^{\mathsf{H}} \otimes \boldsymbol{u}_j\boldsymbol{x}^{\mathsf{H}}]$, so that their norms are the same and (D.6) holds for $\boldsymbol{u}_j\boldsymbol{x}^{\mathsf{H}} \otimes \boldsymbol{h}\boldsymbol{v}_j^{\mathsf{H}} - \mathbb{E}[\boldsymbol{u}_j\boldsymbol{x}^{\mathsf{H}} \otimes \boldsymbol{h}\boldsymbol{v}_j^{\mathsf{H}}]$ as well.

Controlling β_4. For the last term β_4, we denote

$$\boldsymbol{W}_j := \boldsymbol{u}_j\boldsymbol{x}^{\mathsf{H}} \otimes \boldsymbol{u}_j\boldsymbol{x}^{\mathsf{H}} - \mathbb{E}\left[\boldsymbol{u}_j\boldsymbol{x}^{\mathsf{H}} \otimes \boldsymbol{u}_j\boldsymbol{x}^{\mathsf{H}}\right],$$

which satisfies

$$\begin{aligned}
\|\boldsymbol{W}_j\| &= \left\|\left\{\overline{x_i}x_k \left|\langle \boldsymbol{x}, \boldsymbol{a}_j\rangle\right|^2 \left(\boldsymbol{I}_K - \boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\right)\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{I}_K - \boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\right) - \overline{x_i}x_k \|\boldsymbol{x}\|_2^2\left(\boldsymbol{I}_K - \boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\right)\right\}_{i,k}\right\| \\
&\overset{\text{(i)}}{\leq} \left\|\left\{\overline{x_i}x_k \left|\langle \boldsymbol{x}, \boldsymbol{a}_j\rangle\right|^2 \left(\boldsymbol{I}_K - \boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\right)\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{I}_K - \boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\right)\right\}_{i,k}\right\| + \left\|\left\{\overline{x_i}x_k \|\boldsymbol{x}\|_2^2\left(\boldsymbol{I}_K - \boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\right)\right\}_{i,k}\right\| \\
&\leq \left\|\boldsymbol{I}_K - \boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\right\|^2 \left\|\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\right\| \left|\langle \boldsymbol{x}, \boldsymbol{a}_j\rangle\right|^2 \|\boldsymbol{x}\|_2^2 + \left\|\boldsymbol{I}_K - \boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\right\| \left\|\boldsymbol{x}\boldsymbol{x}^{\mathsf{H}}\right\| \|\boldsymbol{x}\|_2^2 \\
&\overset{\text{(ii)}}{\leq} \|\boldsymbol{b}_j\|_2^2 \left|\langle \boldsymbol{x}, \boldsymbol{a}_j\rangle\right|^2 + 1.
\end{aligned}$$

Here, (i) is due to the triangle inequality, and (ii) applies $\|\boldsymbol{h}\|_2 = \|\boldsymbol{x}\|_2 = 1$ and the fact that $\|\boldsymbol{I}_K - \boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\| \leq 1$. It then follows that

$$\left\|\|\boldsymbol{W}_j\|\right\|_{\psi_1} \leq \max_{1\leq j\leq m} \left|\langle \boldsymbol{x}, \boldsymbol{a}_j\rangle\right|^2 \cdot \left\|\|\boldsymbol{b}_j\|_2\right\|_{\psi_2}^2 + 1 \qquad\qquad \leq CK\log m,$$

where the second inequality uses (F.1) and $\left\|\|\boldsymbol{b}_j\|_2\right\|_{\psi_2} \lesssim \sqrt{K}$ (cf. Vershynin [2018, Theorem 3.1.1]). To calculate the variance term, one observes that

$$\boldsymbol{W}_j\left(\boldsymbol{X}\right) = \boldsymbol{W}_j^{\mathsf{H}}\left(\boldsymbol{X}\right) = \left|\langle \boldsymbol{h}, \boldsymbol{b}_j\rangle\right|^2 \boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\boldsymbol{X}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}} - \|\boldsymbol{h}\|_2^2\,\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\boldsymbol{X},$$

which gives

$$\boldsymbol{W}_j^{\mathsf{H}}\boldsymbol{W}_j\left(\boldsymbol{X}\right) = \left|\langle \boldsymbol{h}, \boldsymbol{b}_j\rangle\right|^4 \|\boldsymbol{h}\|_2^2 \|\boldsymbol{a}_j\|_2^2\,\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\boldsymbol{X}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}} - 2\left|\langle \boldsymbol{h}, \boldsymbol{b}_j\rangle\right|^2 \|\boldsymbol{h}\|_2^4\,\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\boldsymbol{X}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}} + \|\boldsymbol{h}\|_2^6\,\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\boldsymbol{X}.$$

It is then seen that

$$\mathbb{E}\left[\boldsymbol{W}_j^{\mathsf{H}}\boldsymbol{W}_j\left(\boldsymbol{X}\right)\right] = 3\left(K+2\right)\|\boldsymbol{h}\|_2^4\,\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\boldsymbol{X} - 2\|\boldsymbol{h}\|_2^6\,\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\boldsymbol{X} + \|\boldsymbol{h}\|_2^6\,\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\boldsymbol{X} = \left(3K+5\right)\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\boldsymbol{X}$$

and

$$\mathbb{E}\left[\boldsymbol{W}_j^{\mathsf{H}}\boldsymbol{W}_j\left(\boldsymbol{X}\right)\right] = \mathbb{E}\left[\boldsymbol{W}_j\boldsymbol{W}_j^{\mathsf{H}}\left(\boldsymbol{X}\right)\right] = \left(3K+5\right)\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\boldsymbol{X}.$$

Therefore, one has

$$\sigma_{\boldsymbol{Z}} := \max\left\{\left\|\sum_{j=1}^m \mathbb{E}\left[\boldsymbol{W}_j^{\mathsf{H}}\boldsymbol{W}_j\right]\right\|^{1/2}, \left\|\sum_{j=1}^m \mathbb{E}\left[\boldsymbol{W}_j\boldsymbol{W}_j^{\mathsf{H}}\right]\right\|^{1/2}\right\} \leq \sqrt{\left(3K+5\right)m}.$$

By the matrix Bernstein inequality Koltchinskii et al. [2011, Proposition 2], one has

$$\left\|\sum_{j=1}^m \boldsymbol{W}_j\right\| \lesssim \sigma_{\boldsymbol{Z}}\sqrt{\log m} + B_{\boldsymbol{Z}}\log\left(\frac{B_{\boldsymbol{Z}}\sqrt{m}}{\sigma_{\boldsymbol{Z}}}\right)\log m \lesssim \sqrt{mK\log m}. \tag{D.7}$$

Putting all this together. Plugging (D.5), (D.6) and (D.7) into (D.4) yields that with probability at least $1 - O(m^{-10})$,

$$\|\mathcal{P}_T\mathcal{A}^*\mathcal{A}\mathcal{P}_T - m\mathcal{P}_T\| \leq C\sqrt{mK\log m}$$

holds for some large enough constant $C > 0$.

E Proof of Theorem 5

The proof of this lower bound is rather standard, and hence we only provide a proof sketch here. First of all, it suffices to consider the case where $\boldsymbol{h}, \boldsymbol{x} \in \mathbb{R}^K$. We assume that $\boldsymbol{h}^\star \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_K)$ and suppose that there is an oracle informing us of $\boldsymbol{h}^\star$, which reduces the problem to estimating $\boldsymbol{x}^\star$ from linear measurements

$$\boldsymbol{y} = \widetilde{\boldsymbol{A}}\boldsymbol{x}^\star + \boldsymbol{\xi},$$

where $\widetilde{\boldsymbol{A}} := [\widetilde{\boldsymbol{a}}_1, \widetilde{\boldsymbol{a}}_2, \cdots, \widetilde{\boldsymbol{a}}_m]^{\mathsf{H}}$ with $\widetilde{\boldsymbol{a}}_j = \overline{\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star}\boldsymbol{a}_j$. Denoting by $\widetilde{\boldsymbol{A}}_{\mathsf{re}}$ and $\widetilde{\boldsymbol{A}}_{\mathsf{im}}$ the real and the imaginary part of $\widetilde{\boldsymbol{A}}$, respectively, the standard minimax risk results for linear regression (e.g. Candes and Plan [2011, Lemma 3.11]) gives

$$\begin{aligned}\inf_{\widehat{\boldsymbol{x}}}\sup_{\boldsymbol{x}^\star\in\mathbb{C}^K}\mathbb{E}\left[\|\widehat{\boldsymbol{x}}-\boldsymbol{x}^\star\|_2^2\,|\boldsymbol{A}\right] &= \frac{1}{2}\sigma^2\left(\mathsf{tr}\left[\left(\widetilde{\boldsymbol{A}}_{\mathsf{re}}^\top\widetilde{\boldsymbol{A}}_{\mathsf{re}}\right)^{-1}\right] + \mathsf{tr}\left[\left(\widetilde{\boldsymbol{A}}_{\mathsf{im}}^\top\widetilde{\boldsymbol{A}}_{\mathsf{im}}\right)^{-1}\right]\right)\\ &\geq K\sigma^2/\max\left\{\left\|\widetilde{\boldsymbol{A}}_{\mathsf{re}}\right\|^2, \left\|\widetilde{\boldsymbol{A}}_{\mathsf{im}}\right\|^2\right\},\end{aligned} \tag{E.1}$$

where the infimum is over all estimator $\widehat{\boldsymbol{x}}$. It is known from standard Gaussian concentration results that, with high probability,

$$\max\left\{\left\|\widetilde{\boldsymbol{A}}_{\mathsf{re}}\right\|, \left\|\widetilde{\boldsymbol{A}}_{\mathsf{im}}\right\|\right\} \leq \left\{\max_{1\leq j\leq m}\left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\right|\right\}\|\boldsymbol{A}\| \lesssim \sqrt{\frac{K}{m}\log m}\cdot\sqrt{m} \asymp \sqrt{K\log m},$$

which together with (E.1) gives

$$\inf_{\widehat{\boldsymbol{x}}}\sup_{\boldsymbol{x}^\star\in\mathbb{C}^K}\mathbb{E}\left[\|\widehat{\boldsymbol{x}}-\boldsymbol{x}^\star\|_2^2\,|\boldsymbol{A}\right] \gtrsim \sigma^2/\log m.$$

In turn, this oracle lower bound implies that, with high probability,

$$\begin{aligned}\inf_{\widehat{\boldsymbol{Z}}}\sup_{\boldsymbol{Z}^\star\in\mathcal{M}^\star}\mathbb{E}\left[\left\|\widehat{\boldsymbol{Z}}-\boldsymbol{Z}^\star\right\|_{\mathrm{F}}^2\mid\boldsymbol{A}\right] &\gtrsim \inf_{\widehat{\boldsymbol{x}}}\sup_{\boldsymbol{x}^\star\in\mathbb{C}^K}\mathbb{E}\left[\left\|\boldsymbol{h}^\star\widehat{\boldsymbol{x}}^{\mathsf{H}}-\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}}^2\mid\boldsymbol{A}\right] \asymp \inf_{\widehat{\boldsymbol{x}}}\sup_{\boldsymbol{x}^\star\in\mathbb{C}^K}\mathbb{E}\left[\|\widehat{\boldsymbol{x}}-\boldsymbol{x}^\star\|_2^2\,\|\boldsymbol{h}^\star\|_2^2\mid\boldsymbol{A}\right]\\ &\gtrsim \sigma^2K/\log m.\end{aligned}$$

Similarly, for the second case, we assume that $\boldsymbol{h}^\star$ is a unit vector and there is an oracle informing us of $\boldsymbol{h}^\star$. Then we again relates the problem to estimating $\boldsymbol{x}^\star$ from linear measurements

$$\boldsymbol{y} = \check{\boldsymbol{A}}\boldsymbol{x}^\star + \boldsymbol{\xi},$$

where $\check{\boldsymbol{A}} := [\check{\boldsymbol{a}}_1, \check{\boldsymbol{a}}_2, \cdots, \check{\boldsymbol{a}}_m]^{\mathsf{H}}$ with $\check{\boldsymbol{a}}_j = \overline{\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star}\boldsymbol{a}_j$. Denoting by $\check{\boldsymbol{A}}_{\mathsf{re}}$ and $\check{\boldsymbol{A}}_{\mathsf{im}}$ the real and the imaginary part of $\check{\boldsymbol{A}}$, respectively. Similar to (E.1), one has

$$\begin{aligned}\inf_{\widehat{\boldsymbol{x}}}\sup_{\boldsymbol{x}^\star\in\mathbb{C}^K}\mathbb{E}\left[\|\widehat{\boldsymbol{x}}-\boldsymbol{x}^\star\|_2^2\,|\boldsymbol{A},\boldsymbol{B}\right] &= \frac{1}{2}\sigma^2\left(\mathsf{tr}\left[\left(\check{\boldsymbol{A}}_{\mathsf{re}}^\top\check{\boldsymbol{A}}_{\mathsf{re}}\right)^{-1}\right] + \mathsf{tr}\left[\left(\check{\boldsymbol{A}}_{\mathsf{im}}^\top\check{\boldsymbol{A}}_{\mathsf{im}}\right)^{-1}\right]\right)\\ &\geq K\sigma^2/\max\left\{\left\|\check{\boldsymbol{A}}_{\mathsf{re}}\right\|^2, \left\|\check{\boldsymbol{A}}_{\mathsf{im}}\right\|^2\right\},\end{aligned} \tag{E.2}$$

by the standard minimax risk results for linear regression (e.g. Candes and Plan [2011, Lemma 3.11]). From standard Gaussian concentration, we have, with high probability,

$$\max\left\{\left\|\check{\boldsymbol{A}}_{\mathsf{re}}\right\|, \left\|\check{\boldsymbol{A}}_{\mathsf{im}}\right\|\right\} \leq \left\{\max_{1\leq j\leq m}\left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\right|\right\}\|\boldsymbol{A}\| \lesssim \sqrt{\log m}\cdot\sqrt{m} \asymp \sqrt{m\log m},$$

which taken collectively with (E.2) gives

$$\inf_{\widehat{\boldsymbol{x}}}\sup_{\boldsymbol{x}^\star\in\mathbb{C}^K}\mathbb{E}\left[\|\widehat{\boldsymbol{x}}-\boldsymbol{x}^\star\|_2^2\,|\boldsymbol{A},\boldsymbol{B}\right] \gtrsim \frac{\sigma^2K}{m\log m}.$$

Hence, this oracle lower bound implies that,

$$\begin{aligned}\inf_{\widehat{\boldsymbol{Z}}}\sup_{\boldsymbol{Z}^\star\in\mathcal{M}^\star}\mathbb{E}\left[\left\|\widehat{\boldsymbol{Z}}-\boldsymbol{Z}^\star\right\|_{\mathrm{F}}^2\mid\boldsymbol{A},\boldsymbol{B}\right]&\gtrsim\inf_{\widehat{\boldsymbol{x}}}\sup_{\boldsymbol{x}^\star\in\mathbb{C}^K}\mathbb{E}\left[\left\|\boldsymbol{h}^\star\widehat{\boldsymbol{x}}^{\mathsf{H}}-\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}}^2\mid\boldsymbol{A},\boldsymbol{B}\right]\\&\asymp\inf_{\widehat{\boldsymbol{x}}}\sup_{\boldsymbol{x}^\star\in\mathbb{C}^K}\mathbb{E}\left[\left\|\widehat{\boldsymbol{x}}-\boldsymbol{x}^\star\right\|_2^2\left\|\boldsymbol{h}^\star\right\|_2^2\mid\boldsymbol{A},\boldsymbol{B}\right]\\&\gtrsim\frac{\sigma^2K}{m\log m},\end{aligned}$$

with high probability.

F Auxiliary lemmas

In this section, we collect several auxiliary lemmas that are useful for the proofs of our main theorems.

Lemma 38. *Consider any fixed vector* $\boldsymbol{x}$ *independent of* $\{\boldsymbol{a}_l\}_{1\le l\le m}$*. Then with probability at least* $1-O\left(m^{-100}\right)$*, we have*

$$\max_{1\le l\le m}\left|\boldsymbol{a}_l^{\mathsf{H}}\boldsymbol{x}\right|\le 20\sqrt{\log m}\left\|\boldsymbol{x}\right\|_2. \tag{F.1}$$

Additionally, there exists some constant $C>0$ *such that with probability at least* $1-O\left(me^{-CK}\right)$*, we have*

$$\max_{1\le l\le m}\left\|\boldsymbol{a}_l\right\|_2\le 10\sqrt{K}. \tag{F.2}$$

Proof. The first result follows from standard Gaussian concentration inequalities as well as the union bound. The second claim results from Vershynin [2018, Theorem 3.1.1]. □

Lemma 39. *Fix an arbitrarily small constant* $\epsilon>0$*. Suppose that Assumption 1 holds and* $m\ge C\mu^2K\log^2 m/\epsilon^2$ *for some sufficiently large constant* $C>0$*. Then one has*

$$\left\|\mathcal{P}_T\mathcal{A}^*\mathcal{A}\mathcal{P}_T-\mathcal{P}_T\right\|\le\epsilon,$$

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

Proof. This has been established in Ahmed et al. [2013, Section 5.2]. □

Lemma 40. *Under Assumption 2, one has*

$$\left\|\frac{1}{m}\sum_{j=1}^m y_j\boldsymbol{b}_j\boldsymbol{a}_j^{\mathsf{H}}-\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right\|\lesssim\frac{\sqrt{mK\log^2 m}}{m}+\frac{\sigma\sqrt{mK\log m}}{m},$$

holds with probability over $1-O(m^{-10})$*, as long as* $m>CK\log^5 m$ *for some large enough constant* $C>0$*.*

Proof. See Appendix F.2. □

F.1 Proof of Lemma 36

By the definition of $\mathcal{A}^*$, we have

$$\mathcal{A}^*(\boldsymbol{\xi})=\sum_{j=1}^m\underbrace{\xi_j\boldsymbol{b}_j\boldsymbol{a}_j^{\mathsf{H}}\mathbb{1}_{\{|\xi_j|\le C\sigma\log m\}}}_{=:\boldsymbol{X}_j}+\sum_{j=1}^m\xi_j\boldsymbol{b}_j\boldsymbol{a}_j^{\mathsf{H}}\mathbb{1}_{\{|\xi_j|>C\sigma\log m\}}.$$

Since

$$\mathbb{P}\left(\min_{1\le j\le m}|\xi_j|>C\sigma\log m\right)\le\sum_{j=1}^m\mathbb{P}\left(|\xi_j|>C\sigma\log m\right)$$

$$\leq O\left(m^{-100}\right),$$

for sufficiently large constant $C > 0$, we have with probability exceeding $1 - O(m^{-10})$, that

$$\|\mathcal{A}^*(\boldsymbol{\xi})\| = \left\|\sum_{j=1}^{m} \boldsymbol{X}_j\right\|. \tag{F.3}$$

To bound $\|\sum_{j=1}^{m} \boldsymbol{X}_j\|$, we proceed by applying the matrix Bernstein inequality Koltchinskii et al. [2011, Proposition 2]. One has

$$\begin{aligned}
B_{\boldsymbol{Z}} &:= \left\|\left\|\xi_j \boldsymbol{b}_j \boldsymbol{a}_j^{\mathsf{H}} \mathbb{1}_{\{|\xi_j| \leq C\sigma \log m\}}\right\|\right\|_{\psi_1} \\
&= \left\|\left|\xi_j \mathbb{1}_{\{|\xi_j| \leq C\sigma \log m\}}\right| \|\boldsymbol{b}_j\|_2 \|\boldsymbol{a}_j\|_2\right\|_{\psi_1} \\
&\overset{\text{(i)}}{\leq} C\sigma \log m \left\|\|\boldsymbol{b}_j\|_2\right\|_{\psi_2} \left\|\|\boldsymbol{a}_j\|_2\right\|_{\psi_2} \\
&\overset{\text{(ii)}}{\lesssim} C\sigma K \log m,
\end{aligned}$$

where (i) uses Vershynin [2018, Lemma 2.7.7] and (ii) is due to the facts that $\|\|\boldsymbol{a}_j\|_2\|_{\psi_2} \lesssim \sqrt{K}$ and $\|\|\boldsymbol{b}_j\|_2\|_{\psi_2} \lesssim \sqrt{K}$ (cf. Vershynin [2018, Theorem 3.1.1]). Next, we turn to control the variance term. One has

$$\begin{aligned}
\left\|\sum_{j=1}^{m} \mathbb{E}\left[\boldsymbol{X}_j \boldsymbol{X}_j^{\mathsf{H}}\right]\right\| &= \left\|\sum_{j=1}^{m} \mathbb{E}\left[|\xi_j|^2 \boldsymbol{b}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{a}_j \boldsymbol{b}_j^{\mathsf{H}} \mathbb{1}_{\{|\xi_j| \leq C\sigma \log m\}}\right]\right\| \\
&= \left\|\sum_{j=1}^{m} \mathbb{E}\left[|\xi_j|^2 \mathbb{1}_{\{|\xi_j| \leq C\sigma \log m\}}\right] \mathbb{E}\left[\boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}}\right] \mathbb{E}\left[\boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{a}_j\right]\right\| \\
&\leq \sigma^2 m K.
\end{aligned}$$

Since $\{\boldsymbol{a}_j\}_{j=1}^m$ have the same distribution as $\{\boldsymbol{b}_j\}_{j=1}^m$, $\|\sum_{j=1}^{m} \mathbb{E}[\boldsymbol{X}_j^{\mathsf{H}} \boldsymbol{X}_j]\|$ can be controlled in the same way as above. Then, we have

$$\sigma_{\boldsymbol{Z}} := \max\left\{\left\|\sum_{j=1}^{m} \mathbb{E}\left[\boldsymbol{X}_j \boldsymbol{X}_j^{\mathsf{H}}\right]\right\|^{1/2}, \left\|\sum_{j=1}^{m} \mathbb{E}\left[\boldsymbol{X}_j^{\mathsf{H}} \boldsymbol{X}_j\right]\right\|^{1/2}\right\} \leq \sigma\sqrt{mK}.$$

Now we are ready to invoke Koltchinskii et al. [2011, Proposition 2] to derive that with probability over $1 - O(m^{-20})$, there holds

$$\left\|\sum_{j=1}^{m} \boldsymbol{X}_j\right\| \lesssim \sigma_{\boldsymbol{Z}} \sqrt{\log m} + B_{\boldsymbol{Z}} \log\left(\frac{B_{\boldsymbol{Z}}\sqrt{m}}{\sigma_{\boldsymbol{Z}}}\right) \log m \lesssim \sigma\sqrt{mK \log m}, \tag{F.4}$$

where the last inequality holds as long as $m \gg K \log^5 m$. Taking (F.4) collectively with (F.3), one has

$$\|\mathcal{A}^*(\boldsymbol{\xi})\| = \left\|\sum_{j=1}^{m} \boldsymbol{X}_j\right\| \lesssim \sigma\sqrt{mK \log m},$$

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

F.2 Proof of Lemma 40

Denote by $\boldsymbol{M} = \frac{1}{m}\sum_{j=1}^{m} y_j \boldsymbol{b}_j \boldsymbol{a}_j^{\mathsf{H}}$. Then we have

$$\|\boldsymbol{M} - \mathbb{E}[\boldsymbol{M}]\| = \left\|\frac{1}{m}\sum_{j=1}^{m} y_j \boldsymbol{b}_j \boldsymbol{a}_j^{\mathsf{H}} - \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}}\right\|$$

$$\leq \frac{1}{m}\left\|\sum_{j=1}^{m}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}} - m\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\right\| + \frac{1}{m}\left\|\sum_{j=1}^{m}\xi_j\boldsymbol{b}_j\boldsymbol{a}_j^{\mathsf{H}}\right\|. \tag{F.5}$$

The second term can be bounded by Lemma 36. We are left to control the first term.

In view of (F.2), one has

$$\sum_{j=1}^{m}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}} - m\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}} = \sum_{j=1}^{m}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\,\mathbb{1}_{\left\{\left|\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}^{\star}\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^{\star}\right|\leq\left(20\sqrt{\log m}\right)^2\right\}} - m\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}, \tag{F.6}$$

holds with probability over $1 - O(m^{-100})$.

Concentration. For any fixed unit vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, define

$$Z_j := \boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{\left|\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}^{\star}\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^{\star}\right|\leq\left(20\sqrt{\log m}\right)^2\right\}}.$$

Then we invoke the Bernstein inequality Vershynin [2018, Theorem 2.8.2] to control $\|\sum_{j=1}^{m}(Z_j - \mathbb{E}[Z_j])\|$. We have

$$\left\|Z_j - \mathbb{E}\left[Z_j\right]\right\|_{\psi_1} \leq C\left\|Z_j\right\|_{\psi_1} \leq 400C\log m\left\|\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\right\|_{\psi_2}\left\|\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\right\|_{\psi_2} \lesssim \log m.$$

Here, we have used $\|X - \mathbb{E}[X]\|_{\psi_1} \leq C\left\|X\right\|_{\psi_1}$ (cf. Vershynin [2018, Section 2.7]). Then the Bernstein inequality Vershynin [2018, Theorem 2.8.2] allows us to derive that

$$\mathbb{P}\left(\left|\sum_{j=1}^{m}\left(Z_j - \mathbb{E}\left[Z_j\right]\right)\right| \geq t\right) \leq 2\exp\left(-c\min\left(\frac{t^2}{m\log^2 m}, \frac{t}{\log m}\right)\right).$$

Letting $t = C_t\sqrt{mK}\log m$ for some large enough constant $C_t > 0$, we obtain that

$$\left|\sum_{j=1}^{m}\left(X_j - \mathbb{E}\left[X_j\right]\right)\right| \leq C_t\sqrt{mK}\log m, \tag{F.7}$$

holds with probability exceeding $1 - 2\exp(-cC_t^2 K)$.

Union bound. Next, we define $\mathcal{N}_0$ an ϵ-net of the unit sphere $\mathcal{S}^{K-1}$. In view of Vershynin [2018, Corollary 4.2.13], we have

$$|\mathcal{N}_0| \leq \left(1 + \frac{2}{\epsilon}\right)^{2K}.$$

Taking this collectively with the union bound yields that (F.7) holds uniformly for any $\boldsymbol{x} \in \mathcal{N}_{\boldsymbol{x}}$ and $\boldsymbol{u}$, $\boldsymbol{v} \in \mathcal{N}_0$ with probability over

$$1 - \left(1 + \frac{2}{\epsilon}\right)^{4K} \cdot 2\exp\left(-cC_t^2 K\right) \geq 1 - 2\exp\left(-CK\log m\right).$$

Approximation. Then, for any $\boldsymbol{u}$, $\boldsymbol{v} \in \mathcal{S}^{K-1}$, one can choose $\boldsymbol{u}_0 \in \mathcal{N}_0$ and $\boldsymbol{v}_0 \in \mathcal{N}_0$ satisfying $\max\{\|\boldsymbol{u} - \boldsymbol{u}_0\|_2, \|\boldsymbol{v} - \boldsymbol{v}_0\|_2\} \leq \epsilon_2$. Let

$$g\left(\boldsymbol{u}, \boldsymbol{v}\right) := \sum_{j=1}^{m}\left[\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{\left|\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}^{\star}\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^{\star}\right|\leq\left(20\sqrt{\log m}\right)^2\right\}} - m\boldsymbol{u}^{\mathsf{H}}\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{v}\right].$$

Set $\epsilon = 1/4$. By triangle inequality, one has

$$\left|g\left(\boldsymbol{u}, \boldsymbol{v}\right) - g\left(\boldsymbol{u}_0, \boldsymbol{v}_0\right)\right| \leq \left|g\left(\boldsymbol{u}, \boldsymbol{v}\right) - g\left(\boldsymbol{u}_0, \boldsymbol{v}\right)\right| + \left|g\left(\boldsymbol{u}_0, \boldsymbol{v}\right) - g\left(\boldsymbol{u}_0, \boldsymbol{v}_0\right)\right|$$

$$\leq 2\epsilon \left\| \sum_{j=1}^{m} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{v}\, \mathbb{1}_{\left\{\left|\boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star\right| \leq \left(20\sqrt{\log m}\right)^2\right\}} - m\boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}} \right\|.$$

Considering $g(\boldsymbol{u}_0, \boldsymbol{v}_0)$, let

$$Z_j := \boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{v}\, \mathbb{1}_{\left\{\left|\boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star\right| \leq \left(20\sqrt{\log m}\right)^2\right\}}.$$

One has

$$\begin{aligned}
&\left|g\left(\boldsymbol{u}_0, \boldsymbol{v}_0\right)\right| \\
&\leq \left|\sum_{j=1}^{m} \left(Z_j - \mathbb{E}\left[Z_j\right]\right)\right| + \left|\sum_{j=1}^{m} \left(\mathbb{E}\left[Z_j\right] - m\boldsymbol{u}^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}} \boldsymbol{v}\right)\right| \\
&\leq C_t \sqrt{mK} \log m + \left|\sum_{j=1}^{m} \mathbb{E}\left[\boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{v}\, \mathbb{1}_{\left\{\left|\boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star\right| \leq \left(20\sqrt{\log m}\right)^2\right\}}\right]\right| \\
&\leq C_t \sqrt{mK} \log m + \sum_{j=1}^{m} \left|\mathbb{E}\left[\boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{v}\, \mathbb{1}_{\left\{\left|\boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star\right| \leq \left(20\sqrt{\log m}\right)^2\right\}}\right]\right| \\
&\leq 2C_t \sqrt{mK} \log m,
\end{aligned}$$

where we use (F.7) and

$$\begin{aligned}
&\left|\mathbb{E}\left[\boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{v}\, \mathbb{1}_{\left\{\left|\boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star\right| \leq \left(20\sqrt{\log m}\right)^2\right\}}\right]\right| \\
&\leq \sqrt{\mathbb{E}\left[\left(\boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{v}\right)^2\right] \mathbb{P}\left(\left|\boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star\right| \leq \left(20\sqrt{\log m}\right)^2\right)} \\
&\leq O\left(m^{-100}\right).
\end{aligned}$$

Hence we have

$$\left|g\left(\boldsymbol{u}_0, \boldsymbol{v}_0\right)\right| \leq 2C_t \sqrt{mK} \log m.$$

Putting all this together. It then follows that

$$\begin{aligned}
\left|g\left(\boldsymbol{u}, \boldsymbol{v}\right)\right| &\leq \left|g\left(\boldsymbol{u}_0, \boldsymbol{v}_0\right)\right| + \left|g\left(\boldsymbol{u}, \boldsymbol{v}\right) - g\left(\boldsymbol{u}_0, \boldsymbol{v}_0\right)\right| \\
&\leq 2C_t \sqrt{mK} \log m + 2\epsilon \left\| \sum_{j=1}^{m} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{v}\, \mathbb{1}_{\left\{\left|\boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star\right| \leq \left(20\sqrt{\log m}\right)^2\right\}} - m\boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}} \right\|.
\end{aligned}$$

Taking maximum over $\boldsymbol{u}$ and $\boldsymbol{v}$ on the left side yields that

$$\begin{aligned}
\max_{\boldsymbol{u}, \boldsymbol{v} \in \mathcal{S}^{K-1}} \left|g\left(\boldsymbol{u}, \boldsymbol{v}\right)\right| &= \left\| \sum_{j=1}^{m} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}}\, \mathbb{1}_{\left\{\left|\boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star\right| \leq \left(20\sqrt{\log m}\right)^2\right\}} - m\boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}} \right\| \\
&\leq 2C_t \sqrt{mK} \log m + 2\epsilon \left\| \sum_{j=1}^{m} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}}\, \mathbb{1}_{\left\{\left|\boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star\right| \leq \left(20\sqrt{\log m}\right)^2\right\}} - m\boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}} \right\|.
\end{aligned}$$

Rearranging terms and recalling $\epsilon = 1/4$ give rise to

$$\left\| \sum_{j=1}^{m} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{v}\, \mathbb{1}_{\left\{\left|\boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star\right| \leq \left(20\sqrt{\log m}\right)^2\right\}} - m\boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}} \right\| \leq 4C_t \sqrt{mK} \log m. \tag{F.8}$$

Taking (F.6) with (F.8) collectively yields that

$$\left\| \sum_{j=1}^{m} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^{\star} \boldsymbol{x}^{\star\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} - m \boldsymbol{h}^{\star} \boldsymbol{x}^{\star\mathsf{H}} \right\| \leq 4 C_t \sqrt{mK} \log m, \tag{F.9}$$

holds with probability at least $1 - O(\exp(-CK \log m) + m^{-100})$. Plugging (F.9) and (D.3) into (F.5) gives the desired conclusion.

References

A. Ahmed, B. Recht, and J. Romberg. Blind deconvolution using convex programming. *IEEE Transactions on Information Theory*, 60(3):1711–1732, 2013.

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

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

Y. Chen, Y. Chi, J. Fan, C. Ma, and Y. Yan. Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *SIAM Journal on Optimization*, 30(4):3098–3121, 2020.

F. M. Dopico. A note on sin θ theorems for singular subspace variations. *BIT Numerical Mathematics*, 40 (2):395–403, 2000.

V. Koltchinskii, K. Lounici, A. B. Tsybakov, et al. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

X. Li, S. Ling, T. Strohmer, and K. Wei. Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *Applied and computational harmonic analysis*, 47(3):893–934, 2019.

C. Ma, K. Wang, Y. Chi, and Y. Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In *International Conference on Machine Learning*, pages 3345–3354. PMLR, 2018.

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.