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

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

August 3, 2020

Abstract

We investigate the effectiveness of convex relaxation and nonconvex optimization in solving bilinear systems of equations (a.k.a. blind deconvolution under a subspace model). Despite the wide applicability, the theoretical understanding about these two paradigms remains largely inadequate in the presence of noise. The current paper makes two contributions by demonstrating that: (1) convex relaxation achieves minimax-optimal statistical accuracy vis-à-vis random noise, and (2) a two-stage nonconvex algorithm attains minimax-optimal accuracy within a logarithmic number of iterations. Both results improve upon the state-of-the-art results by some factors that scale polynomially in the problem dimension.

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

Contents

Author names are sorted alphabetically.

*Department of Electrical 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`.

1 Introduction

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

$$y_j = \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}} \boldsymbol{a}_j + \xi_j, \qquad 1 \leq j \leq m. \tag{1.1}$$

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.

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 communications engineering [LWDF11, WP98, WBSJ15, TXK94, CW98, CE16]. Particularly worth emphasizing is the application of blind deconvolution [ARR13, KH96, LS15, MWCC17], which involves recovering two unknown signals from their circular convolution. As has been made apparent in the seminal work [ARR13], 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 [ARR13] for details. A variety of approaches have since been put forward for blind deconvolution, most notable of which are convex relaxation and nonconvex optimization [ARR13, LS17, LLSW19, MWCC17, HH18, LS19]. 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 Optimization algorithms and prior theory

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

$$\underset{\boldsymbol{h},\boldsymbol{x}\in\mathbb{C}^K}{\text{minimize}} \quad \sum_{j=1}^{m} \left| \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h} \boldsymbol{x}^{\mathsf{H}} \boldsymbol{a}_j - y_j \right|^2, \tag{1.2}$$

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) [ARR13, LS15, LS17]. More concretely, we consider the following convex program:[1]

$$\underset{\boldsymbol{Z}\in\mathbb{C}^{K\times K}}{\text{minimize}} \quad g\left(\boldsymbol{Z}\right) = \sum_{j=1}^{m} \left| \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{Z} \boldsymbol{a}_j - y_j \right|^2 + 2\lambda \left\| \boldsymbol{Z} \right\|_*, \tag{1.3}$$

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 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 (1.2) have been studied in the literature (e.g. [LLSW19, MWCC17, CDDD19, CCD+19, HH18]), the present paper focuses attention on the following ridge-regularized least-squares problem:

$$\underset{\boldsymbol{h},\boldsymbol{x}\in\mathbb{C}^K}{\text{minimize}} \quad f\left(\boldsymbol{h},\boldsymbol{x}\right) = \sum_{j=1}^{m} \left| \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h} \boldsymbol{x}^{\mathsf{H}} \boldsymbol{a}_j - y_j \right|^2 + \lambda \left\| \boldsymbol{h} \right\|_2^2 + \lambda \left\| \boldsymbol{x} \right\|_2^2, \tag{1.4}$$

with $\lambda > 0$ the regularization parameter. Owing to the nonconvexity of (1.4), 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 $(\boldsymbol{h}^0, \boldsymbol{x}^0)$ by means of a spectral method, followed by Wirtinger gradient descent (GD) that iteratively refines the estimates (to be made precise in (1.6a)). 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 (balanced) ℓ_2 norm (see (1.6b)); this balancing step helps stabilize the algorithm, while facilitating analysis. The whole algorithm is summarized in Algorithm 1.

Prior theoretical guarantees. The aforementioned two algorithms have found intriguing 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). Prior theory asserts that 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) [ARR13, LS17], which, however, exceeds the minimax lower bound by at least a factor of $\sqrt{m}$. In comparison, nonconvex algorithms are provably capable of achieving nearly minimax optimal statistical accuracy, with an iteration complexity on the order of K (up to some log factor) [LLSW19, HH18]. See Table 1 for a summary of existing results.

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

[1] 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.

Algorithm 1 Gradient descent for blind deconvolution

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

$$\boldsymbol{M} := \sum_{j=1}^{m} y_j \boldsymbol{b}_j \boldsymbol{a}_j^{\mathsf{H}}. \tag{1.5}$$

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**

$$\begin{bmatrix} \boldsymbol{h}^{t+1/2} \\ \boldsymbol{x}^{t+1/2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{h}^{t} \\ \boldsymbol{x}^{t} \end{bmatrix} - \eta \begin{bmatrix} \nabla_{\boldsymbol{h}} f(\boldsymbol{h}^t, \boldsymbol{x}^t) \\ \nabla_{\boldsymbol{x}} f(\boldsymbol{h}^t, \boldsymbol{x}^t) \end{bmatrix}, \tag{1.6a}$$

$$\begin{bmatrix} \boldsymbol{h}^{t+1} \\ \boldsymbol{x}^{t+1} \end{bmatrix} = \begin{bmatrix} \sqrt{\frac{\|\boldsymbol{x}^{t+1/2}\|_2}{\|\boldsymbol{h}^{t+1/2}\|_2}}\,\boldsymbol{h}^{t+1/2} \\ \sqrt{\frac{\|\boldsymbol{h}^{t+1/2}\|_2}{\|\boldsymbol{x}^{t+1/2}\|_2}}\,\boldsymbol{x}^{t+1/2} \end{bmatrix}, \tag{1.6b}$$

where $\nabla_{\boldsymbol{h}} f(\cdot)$ and $\nabla_{\boldsymbol{x}} f(\cdot)$ represent the Wirtinger gradient (see [LLSW19, Section 3.3] and Appendix A.2.1) of $f(\cdot)$ w.r.t. $\boldsymbol{h}$ and $\boldsymbol{x}$, respectively.

1. *Is convex relaxation inherently suboptimal when coping with random noise?*
2. *Is it possible to further accelerate the nonconvex algorithm without compromising statistical accuracy?*

The present paper is devoted to answering these questions.

1.2 Main results

This subsection presents our theoretical guarantees for the above two algorithms vis-à-vis random noise.

Model and assumptions. In order to formalize our findings, let us make precise the following assumptions that are commonly assumed in the blind deconvolution literature.

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

- *The entries of* $\boldsymbol{A}$ *are independently drawn from standard complex Gaussian distributions, namely,* $\boldsymbol{a}_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\boldsymbol{0}, \frac{1}{2}\boldsymbol{I}_K\right) + i\mathcal{N}\left(\boldsymbol{0}, \frac{1}{2}\boldsymbol{I}_K\right)$ *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* $\|\xi_i\|_{\psi_2} \leq \sigma$ $(1 \leq i \leq m)$*. See [Ver10, Definition 5.7] for the definition of* $\|\cdot\|_{\psi_2}$*.*

In addition, as pointed out by prior work [ARR13, LLSW19, MWCC17], 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*

$$\left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\right| \leq \frac{\mu}{\sqrt{m}}\|\boldsymbol{h}^\star\|_2, \qquad 1 \leq j \leq m. \tag{1.7}$$

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. $\boldsymbol{h}^\star \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_K)$), the incoherence parameter μ is, with high probability, at most $O(\sqrt{\log m})$.

Table 1: Comparison of our theoretical guarantees 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 estimation error (noisy case)	iteration complexity
[ARR13]	$\mu^2 K$	convex relaxation	$\sigma\sqrt{Km}$	—
[LS17]	$\mu^2 K$	convex relaxation	$\sigma\sqrt{Km}$	—
This paper	$\mu^2 K$	convex relaxation	$\sigma\sqrt{K}$	—
[LLSW19]	$\mu^2 K$	nonconvex regularized GD	$\sigma\sqrt{K}$	mK^2
[HH18]	$\mu^2 K$	Riemannian steepest descent	$\sigma\sqrt{K}$	mK^2
[MWCC17]	$\mu^2 K$	nonconvex vanilla GD	—	mK (noiseless)
This paper	$\mu^2 K$	nonconvex GD (with balancing operations)	$\sigma\sqrt{K}$	mK

Main theory. We are now positioned to state our main theory formally, followed by discussing the implications of our theory. Towards this end, we begin with the statistical guarantees for the convex formulation.

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

$$m \geq C\mu^2 K\log^9 m \qquad \textit{and} \qquad \sigma\sqrt{K\log^5 m} \leq c\left\|\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}} \tag{1.8}$$

for some sufficiently large (resp. small) constant $C>0$ (resp. $c>0$). Then under Assumption 1 and the incoherence condition (1.7), one has with probability exceeding $1-O\left(m^{-3}+me^{-K}\right)$ that

$$\left\|\boldsymbol{Z}_{\mathsf{cvx}}-\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\right\| \leq \left\|\boldsymbol{Z}_{\mathsf{cvx}}-\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}} \lesssim \sigma\sqrt{K\log m}. \tag{1.9}$$

In addition, the bounds in (1.9) continue to hold if $\boldsymbol{Z}_{\mathsf{cvx}}$ is replaced by $\boldsymbol{Z}_{\mathsf{cvx},1} := \arg\min_{\boldsymbol{Z}:\mathsf{rank}(\boldsymbol{Z})\leq 1}\|\boldsymbol{Z}-\boldsymbol{Z}_{\mathsf{cvx}}\|_{\mathrm{F}}$ (i.e. the best rank-1 approximation of $\boldsymbol{Z}_{\mathsf{cvx}}$).

Remark 1. Here and throughout, we shall use $f_1(m,K)\lesssim f_2(m,K)$ or $f_1(m,K)=O(f_2(m,K))$ to indicate that there exists some constant $C_1>0$ such that $f_1(m,K)\leq Cf_2(m,K)$ 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.

Next, we turn to theoretical guarantees for the nonconvex algorithm described in Algorithm 1. For notational convenience, we define

$$\boldsymbol{z}^t := \begin{bmatrix} \boldsymbol{h}^t \\ \boldsymbol{x}^t \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{z}^{\star} := \begin{bmatrix} \boldsymbol{h}^{\star} \\ \boldsymbol{x}^{\star} \end{bmatrix} \tag{1.10}$$

throughout this paper. 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:

$$\mathsf{dist}\left(\boldsymbol{z},\boldsymbol{z}^{\star}\right) := \min_{\alpha\in\mathbb{C}}\sqrt{\left\|\frac{1}{\overline{\alpha}}\boldsymbol{h}-\boldsymbol{h}^{\star}\right\|_2^2+\left\|\alpha\boldsymbol{x}-\boldsymbol{x}^{\star}\right\|_2^2}. \tag{1.11}$$

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

Theorem 2 (Nonconvex optimization)**.** *Set $\lambda = C_\lambda \sigma\sqrt{K\log m}$ for some large enough constant $C_\lambda > 0$. Assume that $\|\boldsymbol{h}^\star\|_2 = \|\boldsymbol{x}^\star\|_2$ without loss of generality. Take $\eta = c_\eta$ for some sufficiently small constant $c_\eta > 0$. Suppose that Assumption 1, the incoherence condition (1.7) and the condition (1.8) 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 Algorithm 1 obey*

$$\mathsf{dist}\left(\boldsymbol{z}^0, \boldsymbol{z}^\star\right) \lesssim \sqrt{\frac{\mu^2 K\log m}{m}}\|\boldsymbol{z}^\star\|_2 + \frac{\sigma\sqrt{K\log m}}{\|\boldsymbol{z}^\star\|_2} \tag{1.12a}$$

$$\mathsf{dist}\left(\boldsymbol{z}^t, \boldsymbol{z}^\star\right) \le \rho^t \mathsf{dist}\left(\boldsymbol{z}^0, \boldsymbol{z}^\star\right) + \frac{C_1\left(\lambda + \sigma\sqrt{K\log m}\right)}{c_\rho\|\boldsymbol{z}^\star\|_2} \tag{1.12b}$$

$$\left\|\boldsymbol{h}^t\left(\boldsymbol{x}^t\right)^{\mathsf{H}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}} \le 2\rho^t \mathsf{dist}\left(\boldsymbol{z}^0, \boldsymbol{z}^\star\right)\|\boldsymbol{z}^\star\|_2 + \frac{2C_1\left(\lambda + \sigma\sqrt{K\log m}\right)}{c_\rho} \tag{1.12c}$$

simultaneously for all $0 \le t \le t_0 \le m^{20}$. 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 $c_\rho > 0$.

1.3 Insights

The above theorems — particularly the statistical guarantees for convex relaxation in Theorem 1 — strengthen our understanding about the performance of these 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 reveal that both convex and nonconvex optimization estimate $\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}$ to within an Euclidean error at most $\sigma\sqrt{K}$ (up to some log factor), provided that the regularization parameter is taken to be $\lambda \asymp \sigma\sqrt{K\log m}$. 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 theorem, whose proof is postponed to Appendix C.

  Theorem 3. *Suppose that the noise components obey $\xi_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2/2) + i\mathcal{N}(0, \sigma^2/2)$. Define*

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

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

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

  where the infimum is taken over all estimator $\widehat{\boldsymbol{Z}}$.

  Encouragingly, this minimax lower bound matches the statistical error bounds in Theorems 1-2 up to some logarithmic factor, thus confirming the near minimaxity of both convex relaxation and nonconvex optimization for noisy blind deconvolution.

- *Fast convergence of nonconvex algorithms.* From the computational perspective, Theorem 2 guarantees linear 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 [LLSW19, HH18], our theory reveals considerably faster convergence and hence improved computational cost, without compromising statistical efficiency. See Table 1 for details.

The careful reader might immediately remark that the validity of the above results requires the assumptions (1.8) on both the sample size and the noise level. Fortunately, a closer inspection of these conditions reveals the broad applicability of these conditions.

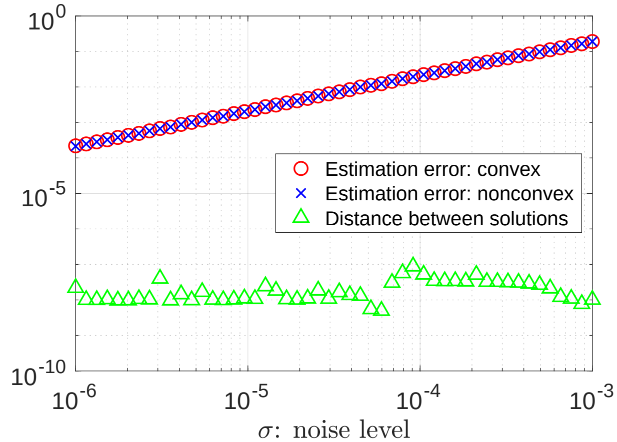

Figure 1: Relative estimation errors of both $\boldsymbol{Z}_{\mathsf{cvx}}$ and $\boldsymbol{Z}_{\mathsf{ncvx}}$ and the relative distance between them vs. the noise level σ. The results are averaged over 20 independent trials.

- *Sample complexity.* The sample size requirement in our theory (as stated in Condition (1.8)) scales as

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

which matches the information-theoretical lower limit even in the absence of noise (modulo some logarithmic factor) [KK17].

- *Signal-to-noise ratio (SNR).* The noise level required for our theory (see Condition (1.8)) to work 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

$$\mathsf{SNR} := \frac{\frac{1}{m}\sum_{k=1}^{m}\mathbb{E}\left[\left|\boldsymbol{b}_k^{\mathsf{H}}\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}\right|^2\right]}{\sigma^2} = \frac{\|\boldsymbol{h}^\star\|_2^2\|\boldsymbol{x}^\star\|_2^2}{m\sigma^2}, \tag{1.13}$$

then our noise requirement can be equivalently phrased as

$$\mathsf{SNR} \gtrsim \frac{K\log^5 m}{m},$$

where the right-hand side 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.

1.4 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 $\mathcal{N}\left(\boldsymbol{0}, \frac{1}{2K}\boldsymbol{I}_K\right) + i\mathcal{N}\left(\boldsymbol{0}, \frac{1}{2K}\boldsymbol{I}_K\right)$ (so that they have approximately unit ℓ_2 norm). The stepsize η is set to be 0.05, whereas the regularization parameter is taken to be $\lambda = 5\sigma\sqrt{K\log m}$. The convex problem is solved by means of the proximal gradient method [PB14].

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}; 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.

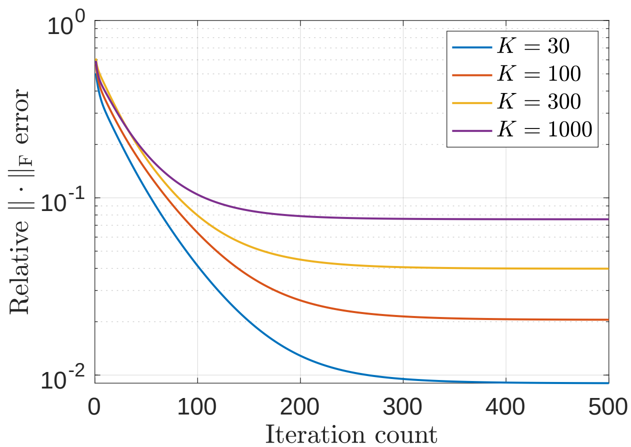

Figure 2: Relative Euclidean error $\left\|\boldsymbol{h}^t\boldsymbol{x}^{t\mathsf{H}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}}$ vs. iteration count.

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 increases as the dimension K increases, which is consistent with Theorem 2.

1.5 Notation

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 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. The notation $f(n) \gg g(n)$ (resp. $f(n) \ll g(n)$) means that there exists a sufficiently large (resp. small) constant $c_1 > 0$ (resp. $c_2 > 0$) such that $f(n) > c_1 g(n)$ (resp. $f(n) \le c_2 g(n)$). In our proof, C serves as a constant whose value might change from line to line.

2 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 (1.3). 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.

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

Algorithm 2 Auxiliary Gradient Descent for Blind Deconvolution

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**

$$\left[\begin{array}{c}\boldsymbol{h}^{t+1/2}\\ \boldsymbol{x}^{t+1/2}\end{array}\right]=\left[\begin{array}{c}\boldsymbol{h}^{t}\\ \boldsymbol{x}^{t}\end{array}\right]-\eta\left[\begin{array}{c}\nabla_{\boldsymbol{h}}f\left(\boldsymbol{h}^t,\boldsymbol{x}^t\right)\\ \nabla_{\boldsymbol{x}}f\left(\boldsymbol{h}^t,\boldsymbol{x}^t\right)\end{array}\right], \tag{2.1a}$$

$$\left[\begin{array}{c}\boldsymbol{h}^{t+1}\\ \boldsymbol{x}^{t+1}\end{array}\right]=\left[\begin{array}{c}\sqrt{\frac{\left\|\boldsymbol{x}^{t+1/2}\right\|_2}{\left\|\boldsymbol{h}^{t+1/2}\right\|_2}}\,\boldsymbol{h}^{t+1/2}\\ \sqrt{\frac{\left\|\boldsymbol{h}^{t+1/2}\right\|_2}{\left\|\boldsymbol{x}^{t+1/2}\right\|_2}}\,\boldsymbol{x}^{t+1/2}\end{array}\right], \tag{2.1b}$$

where $\nabla_{\boldsymbol{h}}f(\cdot)$ and $\nabla_{\boldsymbol{x}}f(\cdot)$ represent the Wirtinger gradient (see [LLSW19, Section 3.3] and Appendix A.2.1) of $f(\cdot)$ w.r.t. $\boldsymbol{h}$ and $\boldsymbol{x}$, respectively.

rescaled iterate

$$\widetilde{\boldsymbol{z}}^t=\left(\widetilde{\boldsymbol{h}}^t,\widetilde{\boldsymbol{x}}^t\right):=\left(\tfrac{1}{\overline{\alpha^t}}\boldsymbol{h}^t,\alpha^t\boldsymbol{x}^t\right). \tag{2.2}$$

with alignment parameter defined by

$$\alpha^t:=\arg\min_{\alpha\in\mathbb{C}}\left\{\left\|\tfrac{1}{\overline{\alpha}}\boldsymbol{h}^t-\boldsymbol{h}^\star\right\|_2^2+\left\|\alpha\boldsymbol{x}^t-\boldsymbol{x}^\star\right\|_2^2\right\}. \tag{2.3}$$

Lemma 1. *Take $\lambda=C_\lambda\sigma\sqrt{K\log m}$ for some large enough constant $C_\lambda>0$. Assume the number of measurements obeys $m\ge C\mu^2K\log^9 m$ 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*

$$\mathsf{dist}\left(\boldsymbol{z}^t,\boldsymbol{z}^\star\right)\le\rho\mathsf{dist}\left(\boldsymbol{z}^{t-1},\boldsymbol{z}^\star\right)+C_5\eta\left(\lambda+\sigma\sqrt{K\log m}\right) \tag{2.4a}$$

$$\max_{1\le j\le m}\left|\boldsymbol{a}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^t-\boldsymbol{x}^\star\right)\right|\le C_7\sqrt{\log m}\left(\lambda+\sigma\sqrt{K\log m}\right) \tag{2.4b}$$

$$\max_{1\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right|\le C_8\left(\frac{\mu}{\sqrt{m}}\log m+\sigma\right) \tag{2.4c}$$

$$\max_{1\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^t-\boldsymbol{h}^\star\right)\right|\le C_9\sigma \tag{2.4d}$$

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

$$\min_{0\le t\le t_0}\left\|\nabla f\left(\boldsymbol{h}^t,\boldsymbol{x}^t\right)\right\|_2\le\frac{\lambda}{m^{10}}. \tag{2.4e}$$

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 (2.4d) 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.2.

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. This is enabled by the following crucial observation.

Lemma 2. *Assume $\sigma\sqrt{K\log m}\le c/\log^2 m$ for some sufficiently small constant $c>0$ and $K\log^4 m/m\ll 1$. Suppose that $(\boldsymbol{h},\boldsymbol{x})$ obeys*

$$\left\|\nabla f\left(\boldsymbol{h},\boldsymbol{x}\right)\right\|_2\lesssim\frac{\lambda}{m^{10}}, \tag{2.5a}$$

$$\left\|\boldsymbol{h}\right\|_2=\left\|\boldsymbol{x}\right\|_2, \quad \left\|\boldsymbol{h}-\boldsymbol{h}^\star\right\|_2\lesssim\lambda+\sigma\sqrt{K\log m}, \quad \left\|\boldsymbol{x}-\boldsymbol{x}^\star\right\|_2\lesssim\lambda+\sigma\sqrt{K\log m}, \tag{2.5b}$$

$$\max_{1\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{h}-\boldsymbol{h}^\star\right)\right|\lesssim\sigma \quad \text{and} \quad \max_{1\le j\le m}\left|\boldsymbol{a}_j^{\mathsf{H}}\left(\boldsymbol{x}-\boldsymbol{x}^\star\right)\right|\lesssim\sqrt{\log m}\left(\lambda+\sigma\sqrt{K\log m}\right). \tag{2.5c}$$

Then, any minimizer $\boldsymbol{Z}_{\mathsf{cvx}}$ of the convex problem (1.3) satisfies

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

Proof. See Appendix B.3. □

In words, if we can find a point $(\boldsymbol{h},\boldsymbol{x})$ that has vanishingly small gradient (cf. (2.5a)) and that enjoys additional properties stated in (2.5b) and (2.5c), then the matrix $\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}$ is guaranteed to be exceedingly close to the solution of the convex program. Encouragingly, Lemma 1 hints at the existence of a point along the trajectory of Algorithm (2) satisfying these conditions (2.5); 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.

Proof of Theorem 1. Armed with the result in Lemma 2 and the properties about the nonconvex trajectory, we are ready to establish Theorem 1 as follows. Let $\bar{t}\coloneqq\arg\min_{0\le t\le t_0}\left\|\nabla f\left(\boldsymbol{h}^t,\boldsymbol{x}^t\right)\right\|_{\mathrm{F}}$, 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 1, we see that $(\boldsymbol{h}_{\mathsf{ncvx}},\boldsymbol{x}_{\mathsf{ncvx}})$ satisfies — with high probability — the small gradient property (2.4e) as well as all conditions required to invoke Lemma 2. As a consequence, invoke Lemma 2 to obtain

$$\left\|\boldsymbol{Z}_{\mathsf{cvx}}-\boldsymbol{h}_{\mathsf{ncvx}}\boldsymbol{x}_{\mathsf{ncvx}}^{\mathsf{H}}\right\|_{\mathrm{F}}\lesssim\left\|\nabla f\left(\boldsymbol{h}_{\mathsf{ncvx}},\boldsymbol{x}_{\mathsf{ncvx}}\right)\right\|_{\mathrm{F}}\lesssim\frac{\lambda}{m^{10}}. \tag{2.6}$$

Further, it is seen that

$$\begin{aligned}
\left\|\boldsymbol{h}_{\mathsf{ncvx}}\left(\boldsymbol{x}_{\mathsf{ncvx}}\right)^{\mathsf{H}}-\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}} &\le\left\|\boldsymbol{h}_{\mathsf{ncvx}}\left(\boldsymbol{x}_{\mathsf{ncvx}}\right)^{\mathsf{H}}-\boldsymbol{h}^\star\left(\boldsymbol{x}_{\mathsf{ncvx}}\right)^{\mathsf{H}}\right\|_{\mathrm{F}}+\left\|\boldsymbol{h}^\star\left(\boldsymbol{x}_{\mathsf{ncvx}}\right)^{\mathsf{H}}-\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}}\\
&\le\left\|\boldsymbol{h}_{\mathsf{ncvx}}-\boldsymbol{h}^\star\right\|_2\left\|\boldsymbol{x}_{\mathsf{ncvx}}\right\|_2+\left\|\boldsymbol{h}^\star\right\|_2\left\|\boldsymbol{x}_{\mathsf{ncvx}}-\boldsymbol{x}^\star\right\|_2\\
&\lesssim 2\left\|\boldsymbol{z}^\star\right\|_2\left(\lambda+\sigma\sqrt{K\log m}\right)\\
&\lesssim\lambda+\sigma\sqrt{K\log m},
\end{aligned} \tag{2.7}$$

where the penultimate line follows from (2.5b) and the inequality that for some constant $C>0$,

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

Taking (2.6) and (2.7) 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.

3 Prior art

As mentioned previously, recent years have witnessed much progress towards understanding convex and nonconvex optimization for solving bilinear systems of equations. Regarding the convex programming approach, [ARR13] 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 a number of nonconvex problems [CSV13, WdM15, CC14, TBSR13, Chi16, CGH14, GW94, SBE14, OJF+15, BR15, KS19, TR14]. Focusing on convex relaxing in the lifted domain, [ARR13] 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 work has also been devoted to understanding convex relaxation under possibly different assumptions. [AAHJ18, AAHJ19] proposed effective convex algorithms 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 bilinear regression [B+19] and the blind demixing problem [LS17, JKS17, SJK17], which is more general than blind deconvolution.

Another line of work has focused on the development of fast nonconvex algorithms [LLSW19, LTR18, MWCC17, HH18, LS19, CDDD19, CCD+19], which was largely motivated by recent advances in efficient nonconvex optimization for tackling statistical estimation problems [CLS15, CC17, CCD+19, KMO09, JNS13, ZCL16, CW15, SL16, ZL16, WGE17, CLPC19, WZG+17, QZEW17, DR19, MXM19] (see [CLC19b] for an overview). [LLSW19] 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 [HH18] 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. In addition, [CDDD19, CCD+19] accounted for outliers in the model and prove theoretical recovery guarantees for subgradient and prox-linear methods. Further, [LS19, DS18, DYS18] extended the nonconvex paradigm to accommodate the blind demixing problem, which subsumes blind deconvolution as a special case.

Going beyond algorithm designs, the past work [CM13, BR15, LLB16, LLB15, KK17] 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 work [WC16, LLJB16, ZLK+17, ZKW19, ZQW20, LB19, QLZ19, KLZW19] 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.

At the technical level, the pivotal idea of our paper lies in bridging convex and nonconvex estimators, which is motivated by prior work [CCF+19, CFMY19, CFMY20] 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 [EK18, CCFM19, CFMW19, DC20, CPC20, DS18, XMR19, CLC+19a, CGZ20, ZB18].

4 Discussion

This paper investigates the effectiveness of both convex relaxation and nonconvex optimization in solving bilinear systems of equations in the presence of random noise. We demonstrate 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 establish — for the first time — optimal statistical guarantees of convex relaxation when applied to blind deconvolution. Both of these results contribute towards demystifying the efficacy of optimization-based methods in solving this fundamental nonconvex problem.

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 [LS17, LS19, DS18]. 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.

Acknowledgements

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 YIP award W911NF-20-1-0097, by the ARO grant W911NF-18-1-0303, by the NSF grants CCF-1907661, IIS-1900140 and DMS-2014279, and by the Princeton SEAS Innovation Award. J. Fan is supported by the NSF grant DMS-1712591, by the ONR grant N00014-19-1-2120, and by the NIH grant 2R01-GM072611-13. B. Wang is supported in part by the Gordon Y. S. Wu Fellowship in Engineering.

References

[AAHJ18] Alireza Aghasi, Ali Ahmed, Paul Hand, and Babhru Joshi. A convex program for bilinear inversion of sparse vectors. In *Advances in Neural Information Processing Systems*, pages 8548–8558, 2018.

[AAHJ19] Alireza Aghasi, Ali Ahmed, Paul Hand, and Babhru Joshi. Branchhull: Convex bilinear inversion from the entrywise product of signals with known signs. *Applied and Computational Harmonic Analysis*, 2019.

[ARR13] Ali Ahmed, Benjamin Recht, and Justin Romberg. Blind deconvolution using convex programming. *IEEE Transactions on Information Theory*, 60(3):1711–1732, 2013.

[B+19] Sohail Bahmani et al. Estimation from nonlinear observations via convex programming with application to bilinear regression. *Electronic Journal of Statistics*, 13(1):1978–2011, 2019.

[BR15] Sohail Bahmani and Justin Romberg. Lifting for blind deconvolution in random mask imaging: Identifiability and convex relaxation. *SIAM Journal on Imaging Sciences*, 8(4):2203–2238, 2015.

[CC14] Yuxin Chen and Yuejie Chi. Robust spectral compressed sensing via structured matrix completion. *IEEE Transactions on Information Theory*, 10(60):6576–6601, 2014.

[CC17] Yuxin Chen and Emmanuel 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.

[CCD+19] Vasileios Charisopoulos, Yudong Chen, Damek Davis, Mateo Díaz, Lijun Ding, and Dmitriy Drusvyatskiy. Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence. *arXiv preprint arXiv:1904.10020*, 2019.

[CCF+19] Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, and Yuling Yan. Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *arXiv preprint arXiv:1902.07698, accepted to SIAM Journal on Optimization*, 2019.

[CCFM19] Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176(1-2):5–37, 2019.

[CDDD19] Vasileios Charisopoulos, Damek Davis, Mateo Díaz, and Dmitriy Drusvyatskiy. Composite optimization for robust blind deconvolution. *arXiv preprint arXiv:1901.01624*, 2019.

[CE16] Patrizio Campisi and Karen Egiazarian. *Blind image deconvolution: theory and applications*. CRC press, 2016.

[CFMW19] Yuxin Chen, Jianqing Fan, Cong Ma, and Kaizheng Wang. Spectral method and regularized mle are both optimal for top-k ranking. *Annals of statistics*, 47(4):2204, 2019.

[CFMY19] Yuxin Chen, Jianqing Fan, Cong Ma, and Yuling Yan. Inference and uncertainty quantification for noisy matrix completion. *Proceedings of the National Academy of Sciences*, 116(46):22931–22937, 2019.

[CFMY20] Yuxin Chen, Jianqing Fan, Cong Ma, and Yuling Yan. Bridging convex and nonconvex optimization in robust pca: Noise, outliers, and missing data. *arXiv preprint arXiv:2001.05484*, 2020.

[CGH14] Yuxin Chen, Leonidas Guibas, and Qixing Huang. Near-optimal joint object matching via convex relaxation. In *International Conference on International Conference on Machine Learning*, 2014.

[CGZ20] Pinhan Chen, Chao Gao, and Anderson Y Zhang. Partial recovery for top-k ranking: Optimality of mle and sub-optimality of spectral method. *arXiv preprint arXiv:2006.16485*, 2020.

[Chi16] Yuejie Chi. Guaranteed blind sparse spikes deconvolution via lifting and convex optimization. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):782–794, 2016.

[CLC+19a] Changxiao Cai, Gen Li, Yuejie Chi, H Vincent Poor, and Yuxin Chen. Subspace estimation from unbalanced and incomplete data matrices: $\ell_{2,\infty}$ statistical guarantees. *arXiv preprint arXiv:1910.04267, accepted to Annals of Statistics*, 2019.

[CLC19b] Yuejie Chi, Yue M Lu, and Yuxin Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.

[CLPC19] Changxiao Cai, Gen Li, H Vincent Poor, and Yuxin Chen. Nonconvex low-rank symmetric tensor completion from noisy data. pages 1863–1874, 2019.

[CLS15] Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[CM13] Sunav Choudhary and Urbashi Mitra. On identifiability in bilinear inverse problems. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4325–4329. IEEE, 2013.

[CP11] Emmanuel J Candes and Yaniv 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.

[CPC20] Changxiao Cai, H Vincent Poor, and Yuxin Chen. Uncertainty quantification for nonconvex tensor completion: Confidence intervals, heteroscedasticity and optimality. *arXiv preprint arXiv:2006.08580*, 2020.

[CSV13] Emmanuel J Candes, Thomas Strohmer, and Vladislav Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

[CW98] Tony F Chan and Chiu-Kwong Wong. Total variation blind deconvolution. *IEEE transactions on Image Processing*, 7(3):370–375, 1998.

[CW15] Yudong Chen and Martin J Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

[DC20] Lijun Ding and Yudong Chen. Leave-one-out approach for matrix completion: Primal and dual analysis. *IEEE Transactions on Information Theory*, 2020.

[DR19] John C Duchi and Feng 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.

[DS18] Jialin Dong and Yuanming Shi. Nonconvex demixing from bilinear measurements. *IEEE Transactions on Signal Processing*, 66(19):5152–5166, 2018.

[DYS18] Jialin Dong, Kai Yang, and Yuanming Shi. Blind demixing for low-latency communication. *IEEE Transactions on Wireless Communications*, 18(2):897–911, 2018.

[EK18] Noureddine 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.

[GW94] Michel X Goemans and David 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.

[HH18] Wen Huang and Paul Hand. Blind deconvolution by a steepest descent algorithm on a quotient manifold. *SIAM Journal on Imaging Sciences*, 11(4):2757–2785, 2018.

[JKS17] Peter Jung, Felix Krahmer, and Dominik Stöger. Blind demixing and deconvolution at near-optimal rate. *IEEE Transactions on Information Theory*, 64(2):704–727, 2017.

[JNS13] Prateek Jain, Praneeth Netrapalli, and Sujay 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.

[KH96] Deepa Kundur and Dimitrios Hatzinakos. Blind image deconvolution. *IEEE signal processing magazine*, 13(3):43–64, 1996.

[KK17] Michael Kech and Felix 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.

[KLT+11] Vladimir Koltchinskii, Karim Lounici, Alexandre 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.

[KLZW19] Han-Wen Kuo, Yenson Lau, Yuqian Zhang, and John Wright. Geometry and symmetry in short-and-sparse deconvolution. In *International Conference on Machine Learning*, pages 3570–3580, 2019.

[KMO09] Raghunandan Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from noisy entries. In *Advances in neural information processing systems*, pages 952–960, 2009.

[KS19] Felix Krahmer and Dominik Stöger. On the convex geometry of blind deconvolution and matrix completion. *arXiv preprint arXiv:1902.11156*, 2019.

[LB19] Yanjun Li and Yoram Bresler. Multichannel sparse blind deconvolution on the sphere. *IEEE Transactions on Information Theory*, 65(11):7415–7436, 2019.

[LLB15] Yanjun Li, Kiryung Lee, and Yoram Bresler. A unified framework for identifiability analysis in bilinear inverse problems with applications to subspace and sparsity models. *arXiv preprint arXiv:1501.06120*, 2015.

[LLB16] Yanjun Li, Kiryung Lee, and Yoram Bresler. Identifiability in blind deconvolution with subspace or sparsity constraints. *IEEE Transactions on information Theory*, 62(7):4266–4275, 2016.

[LLJB16] Kiryung Lee, Yanjun Li, Marius Junge, and Yoram Bresler. Blind recovery of sparse signals from subsampled convolution. *IEEE Transactions on Information Theory*, 63(2):802–821, 2016.

[LLSW19] Xiaodong Li, Shuyang Ling, Thomas Strohmer, and Ke Wei. Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *Applied and computational harmonic analysis*, 47(3):893–934, 2019.

[LS15] Shuyang Ling and Thomas Strohmer. Self-calibration and biconvex compressive sensing. *Inverse Problems*, 31(11):115002, 2015.

[LS17] Shuyang Ling and Thomas Strohmer. Blind deconvolution meets blind demixing: Algorithms and performance bounds. *IEEE Transactions on Information Theory*, 63(7):4497–4520, 2017.

[LS19] Shuyang Ling and Thomas 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.

[LTR18] Kiryung Lee, Ning Tian, and Justin Romberg. Fast and guaranteed blind multichannel deconvolution under a bilinear system model. *IEEE Transactions on Information Theory*, 64(7):4792–4818, 2018.

[LWDF11] Anat Levin, Yair Weiss, Fredo Durand, and William T Freeman. Understanding blind deconvolution algorithms. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2354–2367, 2011.

[MWCC17] Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion and blind deconvolution. *arXiv preprint arXiv:1711.10467*, 2017.

[MXM19] Junjie Ma, Ji Xu, and Arian 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.

[OJF+15] Samet Oymak, Amin Jalali, Maryam Fazel, Yonina C Eldar, and Babak Hassibi. Simultaneously structured models with application to sparse and low-rank matrices. *IEEE Transactions on Information Theory*, 61(5):2886–2908, 2015.

[PB14] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239, 2014.

[QLZ19] Qing Qu, Xiao Li, and Zhihui Zhu. A nonconvex approach for exact and efficient multichannel sparse blind deconvolution. In *Advances in Neural Information Processing Systems*, pages 4015–4026, 2019.

[QZEW17] Qing Qu, Yuqian Zhang, Yonina Eldar, and John Wright. Convolutional phase retrieval. In *Advances in Neural Information Processing Systems*, pages 6086–6096, 2017.

[SBE14] Yoav Shechtman, Amir Beck, and Yonina C Eldar. Gespar: Efficient phase retrieval of sparse signals. *IEEE transactions on signal processing*, 62(4):928–938, 2014.

[SJK17] Dominik Stöger, Peter Jung, and Felix Krahmer. Blind demixing and deconvolution with noisy data: Near-optimal rate. In *WSA 2017; 21th International ITG Workshop on Smart Antennas*, pages 1–5. VDE, 2017.

[SL16] Ruoyu Sun and Zhi-Quan Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016.

[TBSR13] Gongguo Tang, Badri Narayan Bhaskar, Parikshit Shah, and Benjamin Recht. Compressed sensing off the grid. *IEEE transactions on information theory*, 59(11):7465–7490, 2013.

[TR14] Gongguo Tang and Benjamin Recht. Convex blind deconvolution with random masks. In *Computational Optical Sensing and Imaging*, pages CW4C–1. Optical Society of America, 2014.

[TXK94] Lang Tong, Guanghan Xu, and Thomas Kailath. Blind identification and equalization based on second-order statistics: A time domain approach. *IEEE Transactions on information Theory*, 40(2):340–349, 1994.

[Ver10] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

[WBSJ15] Gerhard Wunder, Holger Boche, Thomas Strohmer, and Peter Jung. Sparse signal processing concepts for efficient 5g system design. *IEEE Access*, 3:195–208, 2015.

[WC16] Liming Wang and Yuejie Chi. Blind deconvolution from multiple sparse inputs. *IEEE Signal Processing Letters*, 23(10):1384–1388, 2016.

[WdM15] Irene Waldspurger, Alexandre d'Aspremont, and Stephane Mallat. Phase recovery, maxcut and complex semidefinite programming. *Mathematical Programming*, 149(1-2):47–81, 2015.

[WGE17] Gang Wang, Georgios B Giannakis, and Yonina C Eldar. Solving systems of random quadratic equations via truncated amplitude flow. *IEEE Transactions on Information Theory*, 64(2):773–794, 2017.

[WP98] Xiaodong Wang and H Vincent Poor. Blind equalization and multiuser detection in dispersive cdma channels. *IEEE Transactions on Communications*, 46(1):91–103, 1998.

[WZG+17] Gang Wang, Liang Zhang, Georgios B Giannakis, Mehmet Akçakaya, and Jie Chen. Sparse phase retrieval via truncated amplitude flow. *IEEE Transactions on Signal Processing*, 66(2):479–491, 2017.

[XMR19] Ji Xu, Arian Maleki, and Kamiar Rahnama Rad. Consistent risk estimation in high-dimensional linear regression. *arXiv preprint arXiv:1902.01753*, 2019.

[ZB18] Yiqiao Zhong and Nicolas Boumal. Near-optimal bounds for phase synchronization. *SIAM Journal on Optimization*, 28(2):989–1016, 2018.

[ZCL16] Huishuai Zhang, Yuejie Chi, and Yingbin Liang. Provable non-convex phase retrieval with outliers: Median truncatedwirtinger flow. In *International conference on machine learning*, pages 1022–1031, 2016.

[ZKW19] Yuqian Zhang, Han-Wen Kuo, and John Wright. Structured local optima in sparse blind deconvolution. *IEEE Transactions on Information Theory*, 66(1):419–452, 2019.

[ZL16] Qinqing Zheng and John Lafferty. Convergence analysis for rectangular matrix completion using burer-monteiro factorization and gradient descent. *arXiv preprint arXiv:1605.07051*, 2016.

[ZLK+17] Yuqian Zhang, Yenson Lau, Han-wen Kuo, Sky Cheung, Abhay Pasupathy, and John 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.

[ZQW20] Yuqian Zhang, Qing Qu, and John Wright. From symmetry to geometry: Tractable nonconvex problems. *arXiv preprint arXiv:2007.06753*, 2020.

A Analysis of the nonconvex gradient method

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

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

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

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

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

Additionally, we shall also define

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

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

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:

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

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

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

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 $0 < t \leq t_0$, while the hypotheses (A.6b) and (A.6c) are made for all $0 \leq t \leq t_0$. Clearly, if the hypotheses (A.6a) can be established, then simple recursion yields

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

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:

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

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

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

$$\left|\left|\alpha^0\right| - 1\right| \leq 1/4. \tag{A.6h}$$

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 [CLS15, Section 6] and [MWCC17, 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 (1.4):

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

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

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

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 $f(\boldsymbol{h}, \boldsymbol{x})$ and $f(\boldsymbol{z})$ interchangeably for any $\boldsymbol{z} = \begin{bmatrix} \boldsymbol{h} \\ \boldsymbol{x} \end{bmatrix}$, whenever it is clear from the context.

For notational convenience, we define throughout the following operators: for any $\boldsymbol{z} = [z_j]_{1 \le j \le m}$ and any $\boldsymbol{Z} \in \mathbb{C}^{K \times K}$,

$$\begin{aligned} \mathcal{A}(\boldsymbol{Z}) &:= \left\{\boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{Z} \boldsymbol{a}_j\right\}_{j=1}^{m}, \qquad \mathcal{A}^*(\boldsymbol{z}) = \sum_{j=1}^{m} z_j \boldsymbol{b}_j \boldsymbol{a}_j^{\mathsf{H}}, \\ \mathcal{T}(\boldsymbol{Z}) &:= \mathcal{A}^* \mathcal{A}(\boldsymbol{Z}) = \sum_{j=1}^{m} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{Z} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \qquad \text{and} \\ \mathcal{T}^{\mathsf{debias}}(\boldsymbol{Z}) &:= \mathcal{T}(\boldsymbol{Z}) - \boldsymbol{Z} = (\mathcal{A}^* \mathcal{A} - \mathcal{I})(\boldsymbol{Z}) = \sum_{j=1}^{m} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{Z} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} - \boldsymbol{Z}. \end{aligned} \tag{A.8}$$

Below are two useful properties of the operator $\mathcal{A}$ and the design vectors $\{\boldsymbol{b}_j\}_{j=1}^{m}$.

Lemma 3. *For $\mathcal{A}$ defined in (A.8), with probability at least $1 - m^{-\gamma}$,*

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

Proof. See [LLSW19, Lemma 5.12]. □

Lemma 4. *For any $m \ge 3$ and any $1 \le l \le m$, we have*

$$\sum_{j=1}^{m} \left|\boldsymbol{b}_l^{\mathsf{H}} \boldsymbol{b}_j\right| \le 4 \log m.$$

Proof. See [MWCC17, 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 [MWCC17]. 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 3. 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 3 The lth leave-one-out sequence for nonconvex blind deconvolution

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

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

$$\boldsymbol{M}^{(l)} := \sum_{j: j\neq l} y_j \boldsymbol{b}_j \boldsymbol{a}_j^{\mathsf{H}}, \tag{A.9}$$

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**

$$\begin{aligned}
\begin{bmatrix} \boldsymbol{h}^{t+1/2,(l)} \\ \boldsymbol{x}^{t+1/2,(l)} \end{bmatrix} &= \begin{bmatrix} \boldsymbol{h}^{t,(l)} \\ \boldsymbol{x}^{t,(l)} \end{bmatrix} - \eta \begin{bmatrix} \nabla_{\boldsymbol{h}} f^{(l)}\left(\boldsymbol{h}^t, \boldsymbol{x}^t\right) \\ \nabla_{\boldsymbol{x}} f^{(l)}\left(\boldsymbol{h}^t, \boldsymbol{x}^t\right) \end{bmatrix}, \\
\begin{bmatrix} \boldsymbol{h}^{t+1,(l)} \\ \boldsymbol{x}^{t+1,(l)} \end{bmatrix} &= \begin{bmatrix} \sqrt{\frac{\left\|\boldsymbol{x}^{t+1/2,(l)}\right\|_2}{\left\|\boldsymbol{h}^{t+1/2,(l)}\right\|_2}}\, \boldsymbol{h}^{t+1/2,(l)} \\ \sqrt{\frac{\left\|\boldsymbol{h}^{t+1/2,(l)}\right\|_2}{\left\|\boldsymbol{x}^{t+1/2,(l)}\right\|_2}}\, \boldsymbol{x}^{t+1/2,(l)} \end{bmatrix}.
\end{aligned} \tag{A.10a}$$

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

$$\alpha^{t,(l)} := \arg\min_{\alpha\in\mathbb{C}} \left\{ \left\| \tfrac{1}{\overline{\alpha}} \boldsymbol{h}^{t,(l)} - \boldsymbol{h}^\star \right\|_2^2 + \left\| \alpha \boldsymbol{x}^{t,(l)} - \boldsymbol{x}^\star \right\|_2^2 \right\}, \tag{A.11a}$$

$$\alpha^{t+1/2,(l)} := \arg\min_{\alpha\in\mathbb{C}} \left\{ \left\| \tfrac{1}{\overline{\alpha}} \boldsymbol{h}^{t+1/2,(l)} - \boldsymbol{h}^\star \right\|_2^2 + \left\| \alpha \boldsymbol{x}^{t+1/2,(l)} - \boldsymbol{x}^\star \right\|_2^2 \right\}, \tag{A.11b}$$

along with the properly rescaled iterates

$$\widetilde{\boldsymbol{z}}^{t,(l)} = \begin{bmatrix} \widetilde{\boldsymbol{h}}^{t,(l)} \\ \widetilde{\boldsymbol{x}}^{t,(l)} \end{bmatrix} := \begin{bmatrix} \frac{1}{\overline{\alpha^{t,(l)}}} \boldsymbol{h}^{t,(l)} \\ \alpha^{t,(l)} \boldsymbol{x}^{t,(l)} \end{bmatrix}, \tag{A.12a}$$

$$\widetilde{\boldsymbol{z}}^{t+1/2,(l)} = \begin{bmatrix} \widetilde{\boldsymbol{h}}^{t+1/2,(l)} \\ \widetilde{\boldsymbol{x}}^{t+1/2,(l)} \end{bmatrix} := \begin{bmatrix} \frac{1}{\overline{\alpha^{t+1/2,(l)}}} \boldsymbol{h}^{t+1/2,(l)} \\ \alpha^{t+1/2,(l)} \boldsymbol{x}^{t+1/2,(l)} \end{bmatrix}. \tag{A.12b}$$

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

$$\alpha_{\text{mutual}}^{t,(l)} := \arg\min_{\alpha\in\mathbb{C}} \left\{ \left\| \tfrac{1}{\overline{\alpha}} \boldsymbol{h}^{t,(l)} - \tfrac{1}{\overline{\alpha^t}} \boldsymbol{h}^t \right\|_2^2 + \left\| \alpha \boldsymbol{x}^{t,(l)} - \alpha^t \boldsymbol{x}^t \right\|_2^2 \right\}, \tag{A.13a}$$

$$\alpha_{\text{mutual}}^{t+1/2,(l)} := \arg\min_{\alpha\in\mathbb{C}} \left\{ \left\| \frac{1}{\overline{\alpha}} \boldsymbol{h}^{t+1/2,(l)} - \frac{1}{\overline{\alpha^{t+1/2}}} \boldsymbol{h}^{t+1/2} \right\|_2^2 + \left\| \alpha \boldsymbol{x}^{t+1/2,(l)} - \alpha^{t+1/2} \boldsymbol{x}^{t+1/2} \right\|_2^2 \right\}. \tag{A.13b}$$

Hereafter, we shall also denote

$$\widehat{\boldsymbol{z}}^{t,(l)} := \begin{bmatrix} \widehat{\boldsymbol{h}}^{t,(l)} \\ \widehat{\boldsymbol{x}}^{t,(l)} \end{bmatrix} = \begin{bmatrix} \frac{1}{\overline{\alpha_{\text{mutual}}^{t,(l)}}} \boldsymbol{h}^{t,(l)} \\ \alpha_{\text{mutual}}^{t,(l)} \boldsymbol{x}^{t,(l)} \end{bmatrix}, \tag{A.14a}$$

$$\widehat{\boldsymbol{z}}^{t+1/2,(l)} := \begin{bmatrix} \widehat{\boldsymbol{h}}^{t+1/2,(l)} \\ \widehat{\boldsymbol{x}}^{t+1/2,(l)} \end{bmatrix} = \begin{bmatrix} \frac{1}{\overline{\alpha_{\text{mutual}}^{t+1/2,(l)}}} \boldsymbol{h}^{t+1/2,(l)} \\ \alpha_{\text{mutual}}^{t+1/2,(l)} \boldsymbol{x}^{t+1/2,(l)} \end{bmatrix}. \tag{A.14b}$$

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

$$\mathsf{dist}\left(\boldsymbol{z}^{t,(l)}, \widetilde{\boldsymbol{z}}^{t}\right) \le C_2 \left(\frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^9 m}{m}} + \frac{\sigma}{\log^2 m} \right) \tag{A.15a}$$

$$\left\| \widetilde{\boldsymbol{z}}^{t,(l)} - \widetilde{\boldsymbol{z}}^{t} \right\|_2 \lesssim C_2 \left(\frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^9 m}{m}} + \frac{\sigma}{\log^2 m} \right) \tag{A.15b}$$

$$\mathsf{dist}\left(\boldsymbol{z}^{0,(l)}, \boldsymbol{z}^{\star}\right) \lesssim \sqrt{\frac{\mu^2 K \log m}{m}} + \sigma \sqrt{K \log m} \tag{A.15c}$$

$$\mathsf{dist}\left(\boldsymbol{z}^{0,(l)}, \widetilde{\boldsymbol{z}}^{0}\right) \lesssim \frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^5 m}{m}} + \frac{\sigma}{\log^2 m} \tag{A.15d}$$

for some constant $C_2 \gg C_4^2$. Furthermore, there are several immediate consequences of the hypotheses (A.6) and (A.15) 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.14a) and (A.2a), respectively.

Lemma 5. *Instate the notation and assumptions in Theorem 2. For $t \ge 0$, suppose that the hypotheses (A.6) and (A.15) hold in the first t iterations. Then there exist some constants $C_1, C > 0$ such that for any $1 \le l \le m$,*

$$\mathsf{dist}\left(\boldsymbol{z}^{t}, \boldsymbol{z}^{\star}\right) \le C_1 \left(\sqrt{\frac{\mu^2 K \log m}{m}} + \lambda + \sigma \sqrt{K \log m} \right), \tag{A.16a}$$

$$\left\| \boldsymbol{h}^{t} \left(\boldsymbol{x}^{t}\right)^{\mathsf{H}} - \boldsymbol{h}^{\star} \boldsymbol{x}^{\star \mathsf{H}} \right\| \le C \left(\sqrt{\frac{\mu^2 K \log m}{m}} + \lambda + \sigma \sqrt{K \log m} \right), \tag{A.16b}$$

$$\left\| \widetilde{\boldsymbol{z}}^{t,(l)} - \boldsymbol{z}^{\star} \right\|_2 \le 2C_1 \left(\sqrt{\frac{\mu^2 K \log m}{m}} + \lambda + \sigma \sqrt{K \log m} \right), \tag{A.16c}$$

$$\frac{1}{2} \le \left\| \widetilde{\boldsymbol{x}}^{t} \right\|_2 \le \frac{3}{2}, \qquad \frac{1}{2} \le \left\| \widetilde{\boldsymbol{h}}^{t} \right\|_2 \le \frac{3}{2}, \tag{A.16d}$$

$$\frac{1}{2} \le \left\| \widetilde{\boldsymbol{x}}^{t,(l)} \right\|_2 \le \frac{3}{2}, \qquad \frac{1}{2} \le \left\| \widetilde{\boldsymbol{h}}^{t,(l)} \right\|_2 \le \frac{3}{2}, \tag{A.16e}$$

$$\frac{1}{2} \le \left\| \widehat{\boldsymbol{x}}^{t,(l)} \right\|_2 \le \frac{3}{2}, \qquad \frac{1}{2} \le \left\| \widehat{\boldsymbol{h}}^{t,(l)} \right\|_2 \le \frac{3}{2}. \tag{A.16f}$$

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

$$\left\| \widehat{\boldsymbol{z}}^{t-1/2} - \boldsymbol{z}^{\star} \right\|_2 \le C \left(\sqrt{\frac{\mu^2 K \log m}{m}} + \lambda + \sigma \sqrt{K \log m} \right). \tag{A.16g}$$

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 [MWCC17, 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

$$f_{\text{reg-free}}(\boldsymbol{z}) = f_{\text{reg-free}}(\boldsymbol{h}, \boldsymbol{x}) := \sum_{j=1}^{m} \left| \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h} \boldsymbol{x}^{\mathsf{H}} \boldsymbol{a}_j - y_j \right|^2. \tag{A.17}$$

Our result is this:

Lemma 6. *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*

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

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*

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

- *$\boldsymbol{z}_1 := (\boldsymbol{h}_1, \boldsymbol{x}_1)$ 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*

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

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

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

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 6.

Lemma 7. *Set $\lambda = C_\lambda \sigma\sqrt{K\log m}$ for some large constant $C_\lambda > 0$. The stepsize parameter $\eta > 0$ in Algorithm 3 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*

$$\mathsf{dist}\left(\boldsymbol{z}^{t+1}, \boldsymbol{z}^\star\right) \le \left\|\widehat{\boldsymbol{z}}^{t+1/2} - \boldsymbol{z}^\star\right\|_2 \le \rho\mathsf{dist}\left(\boldsymbol{z}^t, \boldsymbol{z}^\star\right) + C_1\eta\left(\lambda + \sigma\sqrt{K\log m}\right) \tag{A.18}$$

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.15) hold in the first $t-1$ iterations. Then there exists some constant $C > 0$ such that for any $1 \le l \le m$, one has*

$$\left|\left|\alpha^t\right| - 1\right| \lesssim \mathsf{dist}\left(\widetilde{\boldsymbol{z}}^t, \boldsymbol{z}^\star\right) \lesssim \sqrt{\frac{\mu^2 K\log m}{m}} + \lambda + \sigma\sqrt{K\log m}, \tag{A.19a}$$

$$\left|\frac{\alpha^{t-1/2}}{\alpha^{t-1}} - 1\right| \lesssim \eta\left(\sqrt{\frac{\mu^2 K\log m}{m}} + \lambda + \sigma\sqrt{K\log m}\right), \tag{A.19b}$$

$$\left|\left|\alpha_{\text{mutual}}^{t,(l)}\right| - 1\right| \lesssim \left\|\widetilde{\boldsymbol{z}}^{t,(l)} - \boldsymbol{z}^\star\right\|_2 \lesssim \sqrt{\frac{\mu^2 K\log m}{m}} + \lambda + \sigma\sqrt{K\log m}, \tag{A.19c}$$

$$\frac{1}{2} \le \left\|\boldsymbol{x}^t\right\|_2 \le \frac{3}{2}, \qquad \frac{1}{2} \le \left\|\boldsymbol{h}^t\right\|_2 \le \frac{3}{2}, \tag{A.19d}$$

$$\frac{1}{2} \le \left\|\boldsymbol{x}^{t,(l)}\right\|_2 \le \frac{3}{2}, \qquad \frac{1}{2} \le \left\|\boldsymbol{h}^{t,(l)}\right\|_2 \le \frac{3}{2} \tag{A.19e}$$

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.15a).

Lemma 8. *Suppose the sample complexity obeys $m \ge C\mu^2 K\log^9 m$ 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*

$$\max_{1\le l\le m} \mathsf{dist}\left(\boldsymbol{z}^{t+1,(l)}, \widetilde{\boldsymbol{z}}^{t+1}\right) \le C_2\left(\frac{\mu}{\sqrt{m}}\sqrt{\frac{\mu^2 K\log^9 m}{m}} + \frac{\sigma}{\log^2 m}\right) \tag{A.20a}$$

$$\text{and} \qquad \max_{1\le l\le m} \left\|\widetilde{\boldsymbol{z}}^{t+1,(l)} - \widetilde{\boldsymbol{z}}^{t+1}\right\|_2 \lesssim C_2\left(\frac{\mu}{\sqrt{m}}\sqrt{\frac{\mu^2 K\log^9 m}{m}} + \frac{\sigma}{\log^2 m}\right), \tag{A.20b}$$

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

$$\max_{1\le l\le m}\left|\boldsymbol{a}_l^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^{t+1,(l)}-\boldsymbol{x}^{\star}\right)\right|\le 20\sqrt{\log m}\max_{1\le l\le m}\left\|\widetilde{\boldsymbol{x}}^{t+1,(l)}-\boldsymbol{x}^{\star}\right\|_2 \tag{A.21}$$

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

$$\begin{aligned}\left|\boldsymbol{a}_l^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^{t+1}-\boldsymbol{x}^{\star}\right)\right| &\le \left|\boldsymbol{a}_l^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^{t+1}-\widetilde{\boldsymbol{x}}^{t+1,(l)}\right)\right|+\left|\boldsymbol{a}_l^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^{t+1,(l)}-\boldsymbol{x}^{\star}\right)\right|\\
&\le \left\|\boldsymbol{a}_l\right\|_2\left\|\widetilde{\boldsymbol{x}}^{t+1}-\widetilde{\boldsymbol{x}}^{t+1,(l)}\right\|_2+\left|\boldsymbol{a}_l^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^{t+1,(l)}-\boldsymbol{x}^{\star}\right)\right|\\
&\le 10\sqrt{K}C_2\left(\frac{\mu}{\sqrt{m}}\sqrt{\frac{\mu^2K\log^9 m}{m}}+\frac{\sigma}{\log^2 m}\right)\\
&\quad+20\sqrt{\log m}\cdot 2C_1\left(\sqrt{\frac{\mu^2K\log m}{m}}+\lambda+\sigma\sqrt{K\log m}\right)\\
&\le C_3\left(\sqrt{\frac{\mu^2K\log^2 m}{m}}+\lambda+\sigma\sqrt{K}\log m\right),\end{aligned} \tag{A.22}$$

where $C_3\gg C_1$, the penultimate inequality follows from (D.2), (A.20b), (A.21) and (A.16c). 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 9. *Suppose the sample complexity obeys $m\ge C\mu^2K\log^9 m$ 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*

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

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

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 10. *Suppose the sample size obeys $m\ge C\mu^2K\log^4 m$ 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 $\left||\alpha^0| - 1\right| \le 1/4$.

In view of the definition of $\mathsf{dist}(\cdot,\cdot)$, we can invoke Lemma 10 to reach

$$\begin{aligned}\mathsf{dist}\big(\boldsymbol{z}^0, \boldsymbol{z}^\star\big) &= \min_{\alpha\in\mathbb{C}} \sqrt{\left\|\tfrac{1}{\overline{\alpha}}\boldsymbol{h}^0 - \boldsymbol{h}^\star\right\|_2^2 + \left\|\alpha \boldsymbol{x}^0 - \boldsymbol{x}^\star\right\|_2^2} \le \min_{\alpha\in\mathbb{C}} \left\{\left\|\tfrac{1}{\overline{\alpha}}\boldsymbol{h}^0 - \boldsymbol{h}^\star\right\|_2 + \left\|\alpha \boldsymbol{x}^0 - \boldsymbol{x}^\star\right\|_2\right\} \\ &\le \min_{\alpha\in\mathbb{C}, |\alpha|=1} \left\{\left\|\alpha \boldsymbol{h}^0 - \boldsymbol{h}^\star\right\|_2 + \left\|\alpha \boldsymbol{x}^0 - \boldsymbol{x}^\star\right\|_2\right\} \le C_1\left(\sqrt{\frac{\mu^2 K \log m}{m}} + \sigma\sqrt{K\log m}\right). \qquad \text{(A.23)}\end{aligned}$$

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

$$\mathsf{dist}\big(\boldsymbol{z}^{0,(l)}, \boldsymbol{z}^\star\big) \le C_1\left(\sqrt{\frac{\mu^2 K \log m}{m}} + \sigma\sqrt{K\log m}\right), \qquad 1 \le l \le m, \qquad \text{(A.24)}$$

and $\left||\alpha^{0,(l)}| - 1\right| \le 1/4$, as asserted in the hypothesis (A.15c).

The following lemma justifies (A.15d) as well as (A.6c) for the base case.

Lemma 11. *Suppose the sample size 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 $c > 0$. 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*

$$\max_{1\le l\le m} \mathsf{dist}\big(\boldsymbol{z}^{0,(l)}, \widetilde{\boldsymbol{z}}^0\big) \lesssim \frac{\mu}{\sqrt{m}}\sqrt{\frac{\mu^2 K \log^5 m}{m}} + \frac{\sigma}{\log^2 m}, \qquad \text{(A.25a)}$$

$$\max_{1\le l\le m} \left|\boldsymbol{b}_l^{\mathsf{H}} \widetilde{\boldsymbol{h}}^0\right| \lesssim \frac{\mu \log^2 m}{\sqrt{m}} + \sigma, \qquad \text{(A.25b)}$$

$$\max_{1\le j\le \tau} \left|\left(\boldsymbol{b}_j - \boldsymbol{b}_1\right)^{\mathsf{H}} \widetilde{\boldsymbol{h}}^0\right| \lesssim \frac{\mu}{\sqrt{m}}\frac{1}{\log m} + \frac{\sigma}{\log m}. \qquad \text{(A.25c)}$$

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 12. *Suppose the sample size obeys $m \ge C\mu^2 K \log^6 m$ for some sufficiently large constant $C > 0$ and $\sigma\sqrt{K\log^5 m} \le c$ 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*

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

The proof of these three lemmas can be easily obtained via straightforward modifications to [MWCC17, 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.23) and (A.6d). Regarding (1.12c), it follows that

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

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

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

This concludes the proof.

A.4 Proof of Lemma 5

1. Condition (A.16a) follows directly from the ℓ_2 contraction (A.6a) and the bound (A.6e) for the base case.

2. (A.16b) is direct consequence of (A.16a) and triangle inequality. We have

$$\begin{aligned}
\left\| \boldsymbol{h}^t \boldsymbol{x}^{t\mathsf{H}} - \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}} \right\|_{\mathrm{F}} &= \left\| \widetilde{\boldsymbol{h}}^t \widetilde{\boldsymbol{x}}^{t\mathsf{H}} - \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}} \right\|_{\mathrm{F}} \\
&\leq \left\| \widetilde{\boldsymbol{h}}^t \widetilde{\boldsymbol{x}}^{t\mathsf{H}} - \widetilde{\boldsymbol{h}}^t \boldsymbol{x}^{\star\mathsf{H}} \right\|_{\mathrm{F}} + \left\| \widetilde{\boldsymbol{h}}^t \boldsymbol{x}^{\star\mathsf{H}} - \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}} \right\|_{\mathrm{F}} \\
&\leq \left\| \widetilde{\boldsymbol{h}}^t \right\|_2 \left\| \widetilde{\boldsymbol{x}}^t - \boldsymbol{x}^\star \right\|_2 + \left\| \widetilde{\boldsymbol{h}}^t - \boldsymbol{h}^\star \right\|_2 \left\| \boldsymbol{x}^\star \right\|_2 \\
&\leq \left(1 + \mathsf{dist}\left(\boldsymbol{z}^t, \boldsymbol{z}^\star\right)\right) \mathsf{dist}\left(\boldsymbol{z}^t, \boldsymbol{z}^\star\right) + \mathsf{dist}\left(\boldsymbol{z}^t, \boldsymbol{z}^\star\right) \\
&\leq C \left(\sqrt{\frac{\mu^2 K \log m}{m}} + \lambda + \sigma \sqrt{K \log m} \right),
\end{aligned}$$

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.16c), it follows from the triangle inequality that

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

for $t > 0$. Here, the penultimate inequality follows from the distance bounds (A.15b) and (A.16a), 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.15c).

4. Condition (A.16d) immediately results from (A.16a), 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.16e) and (A.16f), 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.16f), one has

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

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.15a). 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.16e), apply (A.15b) and (A.16d) 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.15d), (A.16d) and triangle inequality.

6. When it comes to Condition (A.16g), it is seen from (A.6a) and the choice $\rho = 1 - c_\rho \eta$ that

$$\begin{aligned}\left\|\widehat{\boldsymbol{z}}^{t-1/2} - \boldsymbol{z}^\star\right\|_2 &\leq \rho^t \mathsf{dist}\left(\boldsymbol{z}^0, \boldsymbol{z}^\star\right) + \frac{C_1}{1-\rho}\eta\left(\lambda + \sigma\sqrt{K\log m}\right)\\ &= \rho^t \mathsf{dist}\left(\boldsymbol{z}^0, \boldsymbol{z}^\star\right) + \frac{C_1}{c_\rho}\left(\lambda + \sigma\sqrt{K\log m}\right).\end{aligned}$$

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

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

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

A.5 Proof of Corollary 1

1. To establish (A.19a), we recall that the balancing operation (1.6b) 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

$$\begin{aligned}0 = \left|\alpha^t\right|^2\left\|\widetilde{\boldsymbol{h}}^t\right\|_2^2 - \frac{1}{\left|\alpha^t\right|^2}\left\|\widetilde{\boldsymbol{x}}^t\right\|_2^2 &\leq \left|\alpha^t\right|^2\left(1 + \left\|\widetilde{\boldsymbol{h}}^t - \boldsymbol{h}^\star\right\|_2\right)^2 - \frac{\left(1 - \|\widetilde{\boldsymbol{x}}^t - \boldsymbol{x}^\star\|_2\right)^2}{\left|\alpha^t\right|^2};\\ 0 = \left|\alpha^t\right|^2\left\|\widetilde{\boldsymbol{h}}^t\right\|_2^2 - \frac{1}{\left|\alpha^t\right|^2}\left\|\widetilde{\boldsymbol{x}}^t\right\|_2^2 &\geq \left|\alpha^t\right|^2\left(1 - \left\|\widetilde{\boldsymbol{h}}^t - \boldsymbol{h}^\star\right\|_2\right)^2 - \frac{\left(1 + \|\widetilde{\boldsymbol{x}}^t - \boldsymbol{x}^\star\|_2\right)^2}{\left|\alpha^t\right|^2}.\end{aligned}$$

Rearranging terms, we are left with

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

Combining this with (A.16a), we arrive at

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

2. Regarding (A.19a), 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 [MWCC17, Lemma 54]. Towards this, observe that

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

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

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.16a) and (A.18). Hence, the condition of [MWCC17, Lemma 54] is satisfied. Note that the statement of [MWCC17, 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 $\alpha_2=1$. [MWCC17, Lemma 54] tells us that

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

Additionally, the gradient update rule (1.6a) 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_{\mathsf{reg\text{-}free}}\big(\widetilde{\boldsymbol{z}}^{t-1}\big)-\eta\lambda\widetilde{\boldsymbol{h}}^{t-1}\\ -\eta\left|\alpha^{t-1}\right|^2\nabla_{\boldsymbol{x}}f_{\mathsf{reg\text{-}free}}\big(\widetilde{\boldsymbol{z}}^{t-1}\big)-\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_{\mathsf{reg\text{-}free}}\big(\widetilde{\boldsymbol{z}}^{t-1}\big)-\nabla_{\boldsymbol{h}}f_{\mathsf{reg\text{-}free}}\left(\boldsymbol{z}^{\star}\right)\right)-\eta\lambda\widetilde{\boldsymbol{h}}^{t-1}-\frac{\eta}{|\alpha^{t-1}|^2}\nabla_{\boldsymbol{h}}f_{\mathsf{reg\text{-}free}}\left(\boldsymbol{z}^{\star}\right)\\ -\eta\left|\alpha^{t-1}\right|^2\left(\nabla_{\boldsymbol{x}}f_{\mathsf{reg\text{-}free}}\big(\widetilde{\boldsymbol{z}}^{t-1}\big)-\nabla_{\boldsymbol{x}}f_{\mathsf{reg\text{-}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_{\mathsf{reg\text{-}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_{\mathsf{reg\text{-}free}}\big(\widetilde{\boldsymbol{z}}^{t-1}\big)-\nabla_{\boldsymbol{h}}f_{\mathsf{reg\text{-}free}}\left(\boldsymbol{z}^{\star}\right)\right)\\ \eta\left|\alpha^{t-1}\right|^2\left(\nabla_{\boldsymbol{x}}f_{\mathsf{reg\text{-}free}}\big(\widetilde{\boldsymbol{z}}^{t-1}\big)-\nabla_{\boldsymbol{x}}f_{\mathsf{reg\text{-}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_{\mathsf{reg\text{-}free}}\left(\boldsymbol{z}^{\star}\right)\\ \eta\left|\alpha^{t-1}\right|^2\nabla_{\boldsymbol{x}}f_{\mathsf{reg\text{-}free}}\left(\boldsymbol{z}^{\star}\right)\end{bmatrix}\right\|_2\\
&\le 4\eta\left\|\nabla f_{\mathsf{reg-free}}\left(\widetilde{\boldsymbol{z}}^{t-1}\right)-\nabla f_{\mathsf{reg-free}}\left(\boldsymbol{z}^{\star}\right)\right\|_2+\eta\lambda\left\|\widetilde{\boldsymbol{z}}^{t-1}\right\|_2+4\eta\left\|\nabla f_{\mathsf{reg-free}}\left(\boldsymbol{z}^{\star}\right)\right\|_2,
\end{aligned}
$$

where the last inequality utilizes the consequence of (A.19a) 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

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

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

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

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

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

Consequently, it follows that

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

where the last inequality results from (A.16a), (A.16d), and (A.32). Hence, we arrive at

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

3. Similarly, the balancing step (A.10a) implies $\left\|\boldsymbol{h}^{t,(l)}\right\|_2^2=\left\|\boldsymbol{x}^{t,(l)}\right\|_2^2$. From the definitions of $\alpha_{\text{mutual}}^{t,(l)}$ (cf. (A.13a)), $\widehat{\boldsymbol{h}}^{t,(l)}$ and $\widehat{\boldsymbol{x}}^{t,(l)}$ (cf. (A.14a)), we have

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

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

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

which in turn lead to

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

Taking this together with (A.15a) and (A.16a), we reach

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

where the second line follows from the distance bounds (A.15a) and (A.16a), and the last line holds with the proviso that $m\geq\mu^2K\log^8 m$. This establishes the claim (A.19c).

4. Finally, (A.19d) and (A.19e) are direct consequences of (A.19a), (A.19c) 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 6

Define another loss function as follows

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

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

$$\nabla^2 f_{\text{reg-free}}(\boldsymbol{z}) = \nabla^2 f_{\text{clean}}(\boldsymbol{z}) + \left[\begin{array}{cc} \boldsymbol{M} & \boldsymbol{0} \\ 0 & \overline{\boldsymbol{M}} \end{array}\right], \tag{A.26}$$

where

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

By setting

$$\boldsymbol{u} = \left[\begin{array}{c} \boldsymbol{h}_1 - \boldsymbol{h}_2 \\ \boldsymbol{x}_1 - \boldsymbol{x}_2 \\ \overline{\boldsymbol{h}_1 - \boldsymbol{h}_2} \\ \overline{\boldsymbol{x}_1 - \boldsymbol{x}_2} \end{array}\right] =: \left[\begin{array}{c} \boldsymbol{u}_{\boldsymbol{h}} \\ \boldsymbol{u}_{\boldsymbol{x}} \\ \overline{\boldsymbol{u}_{\boldsymbol{h}}} \\ \overline{\boldsymbol{u}_{\boldsymbol{x}}} \end{array}\right]$$

and recalling the definitions of $\boldsymbol{D}$, γ_1, γ_2 in the statement of Lemma 6, we arrive at

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

Consequently, with high probability one has

$$\begin{aligned}
&\left|\boldsymbol{u}^{\mathsf{H}}\left[\boldsymbol{D}\nabla^2 f_{\text{reg-free}}(\boldsymbol{z}) + \nabla^2 f_{\text{reg-free}}(\boldsymbol{z})\boldsymbol{D}\right]\boldsymbol{u} - \boldsymbol{u}^{\mathsf{H}}\left[\boldsymbol{D}\nabla^2 f_{\text{clean}}(\boldsymbol{z}) + \nabla^2 f_{\text{clean}}(\boldsymbol{z})\boldsymbol{D}\right]\boldsymbol{u}\right| \\
&\leq 4(\gamma_1+\gamma_2)\left|\mathrm{Re}\left(\boldsymbol{u}_{\boldsymbol{h}}^{\mathsf{H}}\sum_{j=1}^{m}\xi_j\boldsymbol{b}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{u}_{\boldsymbol{x}}\right)\right| \leq 4(\gamma_1+\gamma_2)\left\|\sum_{j=1}^{m}\xi_j\boldsymbol{b}_j\boldsymbol{a}_j^{\mathsf{H}}\right\|\|\boldsymbol{u}\|_2^2 \\
&\lesssim \sigma\sqrt{K\log m}\,\|\boldsymbol{u}\|_2^2 =: \mathcal{E}_{\text{res}}
\end{aligned} \tag{A.27}$$

for any vector $\boldsymbol{u}$, where the last inequality follows from Lemma 18 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 [MWCC17]. In particular, it has been shown in [MWCC17] that

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

under the assumptions stated in the lemma. These bounds together with (A.27) yield

$$\boldsymbol{u}^{\mathsf{H}}\left[\boldsymbol{D}\nabla^2 f_{\text{reg-free}}(\boldsymbol{z}) + \nabla^2 f_{\text{reg-free}}(\boldsymbol{z})\boldsymbol{D}\right]\boldsymbol{u} \geq (1/4)\cdot\|\boldsymbol{u}\|_2^2 - \mathcal{E}_{\text{res}} \geq (1/8)\cdot\|\boldsymbol{u}\|_2^2, \tag{A.28a}$$

$$\text{and} \quad \left\|\nabla^2 f_{\text{reg-free}}(\boldsymbol{z})\right\| \leq \left\|\nabla^2 f_{\text{clean}}(\boldsymbol{z})\right\| + \sup_{\boldsymbol{u}\neq\boldsymbol{0}}\frac{\mathcal{E}_{\text{res}}}{\|\boldsymbol{u}\|_2^2} \leq 7/2, \tag{A.28b}$$

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

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

which combined with (A.28) and the assumption $\lambda \leq C_\lambda\sigma\sqrt{K\log m} \leq C_\lambda c_1/\log^2 m \ll 1$ yields

$$\boldsymbol{u}^{\mathsf{H}}\left[\boldsymbol{D}\nabla^2 f(\boldsymbol{z}) + \nabla^2 f(\boldsymbol{z})\boldsymbol{D}\right]\boldsymbol{u} = \boldsymbol{u}^{\mathsf{H}}\left[\boldsymbol{D}\nabla^2 f_{\text{reg-free}}(\boldsymbol{z}) + \nabla^2 f_{\text{reg-free}}(\boldsymbol{z})\boldsymbol{D}\right]\boldsymbol{u} + 2\lambda\boldsymbol{u}^{\mathsf{H}}\boldsymbol{D}\boldsymbol{u}$$

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

and

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

A.7 Proof of Lemma 7

Recognizing that

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

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

$$
\textsf{dist}\left(\boldsymbol{z}^{t+1},\boldsymbol{z}^{\star}\right) = \textsf{dist}\left(\boldsymbol{z}^{t+1/2},\boldsymbol{z}^{\star}\right) \leq \left\|\begin{bmatrix} \frac{1}{\overline{\alpha^{t}}}\boldsymbol{h}^{t+1/2}-\boldsymbol{h}^{\star} \\ \alpha^{t}\boldsymbol{x}^{t+1/2}-\boldsymbol{x}^{\star} \end{bmatrix}\right\|_{2} \tag{A.29}
$$

$$
\begin{aligned}
&= \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}} \\
&\qquad + \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}} + \eta\lambda\underbrace{\left\|\begin{bmatrix} \widetilde{\boldsymbol{h}}^{t} \\ \widetilde{\boldsymbol{x}}^{t} \end{bmatrix}\right\|_{2}}_{=:\beta_{3}}.
\end{aligned} \tag{A.30}
$$

Using an argument similar to the proof idea of [MWCC17, Equation (210)], we can obtain

$$
\begin{aligned}
\beta_{1}^{2} &= \left\|\widetilde{\boldsymbol{h}}^{t} - \frac{\eta}{|\alpha^{t}|^{2}}\nabla_{\boldsymbol{h}} f_{\textsf{reg-free}}\left(\widetilde{\boldsymbol{z}}^{t}\right) - \left(\boldsymbol{h}^{\star} - \frac{\eta}{|\alpha^{t}|^{2}}\nabla_{\boldsymbol{h}} f_{\textsf{reg-free}}\left(\boldsymbol{z}^{\star}\right)\right)\right\|_{2}^{2} \\
&\qquad + \left\|\widetilde{\boldsymbol{x}}^{t} - \eta\left|\alpha^{t}\right|^{2}\nabla_{\boldsymbol{x}} f_{\textsf{reg-free}}\left(\widetilde{\boldsymbol{z}}^{t}\right) - \left(\boldsymbol{x}^{\star} - \eta\left|\alpha^{t}\right|^{2}\nabla_{\boldsymbol{x}} f_{\textsf{reg-free}}\left(\boldsymbol{z}^{\star}\right)\right)\right\|_{2}^{2} \\
&\leq \left(1-\frac{\eta}{8}\right)\left\|\widetilde{\boldsymbol{z}}^{t}-\boldsymbol{z}^{\star}\right\|_{2}^{2}.
\end{aligned} \tag{A.31}
$$

Regarding β_2, we first invoke Lemma 19 and the fact $\nabla f_{\textsf{clean}}\left(\boldsymbol{z}^{\star}\right) = \boldsymbol{0}$ to derive

$$
\begin{aligned}
\left\|\nabla f_{\textsf{reg-free}}\left(\boldsymbol{z}^{\star}\right)\right\|_{2} &\leq \left\|\nabla f_{\textsf{clean}}\left(\boldsymbol{z}^{\star}\right)\right\|_{2} + \left\|\mathcal{A}^{*}\left(\boldsymbol{\xi}\right)\right\|\left\|\boldsymbol{h}^{\star}\right\|_{2} + \left\|\mathcal{A}^{*}\left(\boldsymbol{\xi}\right)\right\|\left\|\boldsymbol{x}^{\star}\right\|_{2} \\
&\lesssim \sigma\sqrt{K\log m}.
\end{aligned} \tag{A.32}
$$

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

$$\beta_3^2 = \eta^2\lambda^2 \big\|\widetilde{\boldsymbol{h}}^t\big\|_2^2 + \eta^2\lambda^2 \big\|\widetilde{\boldsymbol{x}}^t\big\|_2^2 \le 8\eta^2\lambda^2,$$

using the fact that $\big\|\widetilde{\boldsymbol{x}}^t\big\|_2 \asymp \big\|\widetilde{\boldsymbol{h}}^t\big\|_2 \asymp 1$ (see Lemma 5).

As a result, as long as $\eta > 0$ is taken to be some constant small enough, combining (A.30) and the above bounds on β_1, β_2 gives

$$\mathsf{dist}\left(\boldsymbol{z}^{t+1}, \boldsymbol{z}^\star\right) \le \left\|\widehat{\boldsymbol{z}}^{t+1/2} - \boldsymbol{z}^\star\right\|_2^2 \le \sqrt{(1-\eta/8)}\,\left\|\widetilde{\boldsymbol{z}}^t - \boldsymbol{z}^\star\right\|_2 + C_1\eta\left(\lambda + \sigma\sqrt{K\log m}\right),$$

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

$$\begin{aligned}\mathsf{dist}\left(\boldsymbol{z}^{t+1}, \boldsymbol{z}^\star\right) &\le \left\|\widehat{\boldsymbol{z}}^{t+1/2} - \boldsymbol{z}^\star\right\|_2 \le (1-\eta/16)\,\left\|\widetilde{\boldsymbol{z}}^t - \boldsymbol{z}^\star\right\|_2 + C_1\eta\left(\lambda + \sigma\sqrt{K\log m}\right)\\ &= (1-\eta/16)\,\mathsf{dist}\left(\boldsymbol{z}^t, \boldsymbol{z}^\star\right) + C_1\eta\left(\lambda + \sigma\sqrt{K\log m}\right).\end{aligned}$$

The advertised claim then follows, provided that C_1 is large enough.

A.8 Proof of Lemma 8

The lemma can be established in a similar manner as [MWCC17, Lemma 17]. We have

$$\begin{aligned}\mathsf{dist}\big(\boldsymbol{z}^{t+1,(l)}, \widetilde{\boldsymbol{z}}^{t+1}\big) &= \mathsf{dist}\big(\boldsymbol{z}^{t+1/2,(l)}, \widetilde{\boldsymbol{z}}^{t+1/2}\big)\\ &\le \max\left\{\left|\frac{\alpha^{t+1/2}}{\alpha^t}\right|, \left|\frac{\alpha^t}{\alpha^{t+1/2}}\right|\right\}\left\|\left[\begin{array}{c}\frac{1}{\alpha_{\text{mutual}}^{t,(l)}}\boldsymbol{h}^{t+1/2,(l)} - \frac{1}{\alpha^t}\boldsymbol{h}^{t+1/2}\\ \alpha_{\text{mutual}}^{t,(l)}\boldsymbol{x}^{t+1,(l)} - \alpha^t\boldsymbol{x}^{t+1/2}\end{array}\right]\right\|_2,\end{aligned} \tag{A.33}$$

where the second line comes from the same calculation as [MWCC17, Eqn. (212)]. Repeating the analysis in [MWCC17, Appendix C.3] and using the gradient update rule, we obtain

$$\begin{aligned}&\left[\begin{array}{c}\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{array}\right]\\ &= \underbrace{\left[\begin{array}{c}\widehat{\boldsymbol{h}}^{t,(l)} - \frac{\eta}{\left|\alpha_{\text{mutual}}^{t,(l)}\right|^2}\nabla_{\boldsymbol{h}} f_{\textsf{reg-free}}\big(\widehat{\boldsymbol{z}}^{t,(l)}\big) - \left(\widetilde{\boldsymbol{h}}^t - \frac{\eta}{\left|\alpha_{\text{mutual}}^{t,(l)}\right|^2}\nabla_{\boldsymbol{h}} f_{\textsf{reg-free}}\big(\widetilde{\boldsymbol{z}}^t\big)\right)\\ \widehat{\boldsymbol{x}}^{t,(l)} - \eta\big|\alpha_{\text{mutual}}^{t,(l)}\big|^2\nabla_{\boldsymbol{x}} f_{\textsf{reg-free}}\big(\widehat{\boldsymbol{z}}^{t,(l)}\big) - \left(\widetilde{\boldsymbol{x}}^t - \eta\big|\alpha_{\text{mutual}}^{t,(l)}\big|^2\nabla_{\boldsymbol{x}} f_{\textsf{reg-free}}\big(\widetilde{\boldsymbol{z}}^t\big)\right)\end{array}\right]}_{=:\boldsymbol{\nu}_1}\\ &\quad + \eta\underbrace{\left[\begin{array}{c}\left(\frac{1}{|\alpha^t|^2} - \frac{1}{\left|\alpha_{\text{mutual}}^{t,(l)}\right|^2}\right)\nabla_{\boldsymbol{h}} f_{\textsf{reg-free}}\big(\widehat{\boldsymbol{z}}^t\big)\\ \left(|\alpha^t|^2 - \left|\alpha_{\text{mutual}}^{t,(l)}\right|^2\right)\nabla_{\boldsymbol{x}} f_{\textsf{reg-free}}\big(\widehat{\boldsymbol{z}}^t\big)\end{array}\right]}_{=:\boldsymbol{\nu}_2} - \eta\underbrace{\left[\begin{array}{c}\frac{1}{\left|\alpha_{\text{mutual}}^{t,(l)}\right|^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)}\\ \left|\alpha_{\text{mutual}}^{t,(l)}\right|^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{a}_l\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\end{array}\right]}_{=:\boldsymbol{\nu}_3}\\ &\quad + \eta\lambda\underbrace{\left[\begin{array}{c}\widehat{\boldsymbol{h}}^{t,(l)} - \widetilde{\boldsymbol{h}}^t\\ \widehat{\boldsymbol{x}}^{t,(l)} - \widetilde{\boldsymbol{x}}^t\end{array}\right]}_{=:\boldsymbol{\nu}_4}.\end{aligned} \tag{A.34}$$

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 [MWCC17, Lemma 17] that

$$\|\boldsymbol{\nu}_1\|_2 \le (1-\eta/16)\,\big\|\widehat{\boldsymbol{z}}^{t,(l)} - \widetilde{\boldsymbol{z}}^t\big\|_2; \qquad \|\boldsymbol{\nu}_2\|_2 \lesssim C_1\frac{1}{\log^2 m}\big\|\widehat{\boldsymbol{z}}^{t,(l)} - \widetilde{\boldsymbol{z}}^t\big\|_2. \tag{A.35}$$

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

$$\begin{aligned}
\left\|\boldsymbol{\nu}_3\right\|_2 &= \sqrt{\frac{1}{\left|\alpha_{\text{mutual}}^{t,(l)}\right|^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 + \left|\alpha_{\text{mutual}}^{t,(l)}\right|^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}{\left|\alpha_{\text{mutual}}^{t,(l)}\right|^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 + \left|\alpha_{\text{mutual}}^{t,(l)}\right|^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}{\left|\alpha_{\text{mutual}}^{t,(l)}\right|^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 + \left|\alpha_{\text{mutual}}^{t,(l)}\right|^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}{\left|\alpha_{\text{mutual}}^{t,(l)}\right|^2}\underbrace{\left\|\xi_l\boldsymbol{b}_l\boldsymbol{a}_l^{\mathsf{H}}\widehat{\boldsymbol{x}}^{t,(l)}\right\|_2}_{=:\nu_{33}} + \left|\alpha_{\text{mutual}}^{t,(l)}\right|^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}$$

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 [MWCC17, Equation (217)] and are thus omitted here for simplicity. The quantity ν_{31} can be upper bounded by

$$\begin{aligned}
\nu_{31} &\leq \left|\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\widehat{\boldsymbol{x}}^{t,(l)\mathsf{H}}\boldsymbol{a}_l - \boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_l\right|\left\|\boldsymbol{b}_l\right\|_2\left|\boldsymbol{a}_l^{\mathsf{H}}\widehat{\boldsymbol{x}}^{t,(l)}\right| \\
&\leq \left|\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\widehat{\boldsymbol{x}}^{t,(l)\mathsf{H}}\boldsymbol{a}_l - \boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_l\right| \cdot \sqrt{\frac{K}{m}} \cdot 20\sqrt{\log m} \cdot \left\|\widehat{\boldsymbol{x}}^{t,(l)}\right\|_2 \\
&\leq 40\sqrt{\frac{K\log m}{m}}\left|\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\widehat{\boldsymbol{x}}^{t,(l)\mathsf{H}}\boldsymbol{a}_l - \boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_l\right|, \qquad \text{(A.36a)}
\end{aligned}$$

where the penultimate inequality follows from the fact that $\left\|\boldsymbol{b}_l\right\|_2 = \sqrt{K/m}$ and (D.1), and the last line makes use of (A.16f). Regarding ν_{32}, one has

$$\begin{aligned}
\nu_{32} &\leq \left|\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\widehat{\boldsymbol{x}}^{t,(l)\mathsf{H}}\boldsymbol{a}_l - \boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_l\right|\left\|\boldsymbol{a}_l\right\|_2\left|\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\right| \\
&\leq \left|\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\widehat{\boldsymbol{x}}^{t,(l)\mathsf{H}}\boldsymbol{a}_l - \boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_l\right| \cdot 10\sqrt{K} \cdot \left(\sqrt{\frac{K}{m}}\left\|\widehat{\boldsymbol{h}}^{t,(l)} - \widetilde{\boldsymbol{h}}^{t}\right\|_2 + \left|\boldsymbol{b}_l^{\mathsf{H}}\widetilde{\boldsymbol{h}}^{t}\right|\right) \\
&\leq \left|\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\widehat{\boldsymbol{x}}^{t,(l)\mathsf{H}}\boldsymbol{a}_l - \boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_l\right| \cdot 10\sqrt{K} \cdot \sqrt{\frac{K}{m}}C_2\left(\frac{\mu}{\sqrt{m}}\sqrt{\frac{\mu^2 K\log^9 m}{m}} + \frac{\sigma}{\log^2 m}\right) \\
&\quad + \left|\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\widehat{\boldsymbol{x}}^{t,(l)\mathsf{H}}\boldsymbol{a}_l - \boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_l\right| \cdot 10\sqrt{K} \cdot C_4\left(\frac{\mu}{\sqrt{m}}\log^2 m + \sigma\right) \\
&\leq 20C_4\left(\frac{\mu\sqrt{K}}{\sqrt{m}}\log^2 m + \sigma\sqrt{K}\right)\left|\boldsymbol{b}_l^{\mathsf{H}}\widehat{\boldsymbol{h}}^{t,(l)}\widehat{\boldsymbol{x}}^{t,(l)\mathsf{H}}\boldsymbol{a}_l - \boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_l\right|, \qquad \text{(A.36b)}
\end{aligned}$$

where the second line follows from (D.2), triangle inequality and the fact that $\left\|\boldsymbol{b}_l\right\|_2 = \sqrt{K/m}$; the penultimate inequality follows from (A.15a) and (A.6c); the last line holds as long as $m \gg \mu^2 K\log^3 m$. Further we have

$$\begin{aligned}
\left|\boldsymbol{b}_l^{\mathsf{H}}\left(\widehat{\boldsymbol{h}}^{t,(l)} - \boldsymbol{h}^{\star}\right)\right| &\leq \left|\boldsymbol{b}_l^{\mathsf{H}}\left(\widehat{\boldsymbol{h}}^{t,(l)} - \widetilde{\boldsymbol{h}}^{t}\right)\right| + \left|\boldsymbol{b}_l^{\mathsf{H}}\widetilde{\boldsymbol{h}}^{t}\right| + \left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^{\star}\right| \\
&\leq \sqrt{\frac{K}{m}}\left\|\widehat{\boldsymbol{h}}^{t,(l)} - \widetilde{\boldsymbol{h}}^{t}\right\|_2 + \left|\boldsymbol{b}_l^{\mathsf{H}}\widetilde{\boldsymbol{h}}^{t}\right| + \left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{h}^{\star}\right|
\end{aligned}$$

$$\le \sqrt{\frac{K}{m}} C_2 \left(\frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^9 m}{m}} + \frac{\sigma}{\log^2 m} \right) + C_4 \left(\frac{\mu}{\sqrt{m}} \log^2 m + \sigma \right) + \frac{\mu}{\sqrt{m}}$$

$$\le 2C_4 \left(\frac{\mu}{\sqrt{m}} \log^2 m + \sigma \right), \tag{A.36c}$$

where the second line follows from the fact that $\|\boldsymbol{b}_l\|_2 = \sqrt{K/m}$; the penultimate inequality follows from (A.15a), (A.6c) and (1.7); the last line holds as long as $m \gg \mu^2 K \log^3 m$. Therefore,

$$\begin{aligned}
&\left| \boldsymbol{b}_l^{\mathsf{H}} \widehat{\boldsymbol{h}}^{t,(l)} \widehat{\boldsymbol{x}}^{t,(l)\mathsf{H}} \boldsymbol{a}_l - \boldsymbol{b}_l^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}} \boldsymbol{a}_l \right| \\
&\le \left| \boldsymbol{b}_l^{\mathsf{H}} \widehat{\boldsymbol{h}}^{t,(l)} \left(\widehat{\boldsymbol{x}}^{t,(l)} - \boldsymbol{x}^\star \right)^{\mathsf{H}} \boldsymbol{a}_l \right| + \left| \boldsymbol{b}_l^{\mathsf{H}} (\widehat{\boldsymbol{h}}^{t,(l)} - \boldsymbol{h}^\star) \boldsymbol{x}^{\star\mathsf{H}} \boldsymbol{a}_l \right| \\
&\le \left(\left| \boldsymbol{b}_l^{\mathsf{H}} (\widehat{\boldsymbol{h}}^{t,(l)} - \boldsymbol{h}^\star) \right| + \left| \boldsymbol{b}_l^{\mathsf{H}} \boldsymbol{h}^\star \right| \right) \cdot 20 \sqrt{\log m} \left(\left\| \widehat{\boldsymbol{x}}^{t,(l)} - \widetilde{\boldsymbol{x}}^t \right\|_2 + \left\| \widetilde{\boldsymbol{x}}^t - \boldsymbol{x}^\star \right\|_2 \right) + \left| \boldsymbol{b}_l^{\mathsf{H}} (\widehat{\boldsymbol{h}}^{t,(l)} - \boldsymbol{h}^\star) \right| \cdot \left| \boldsymbol{x}^{\star\mathsf{H}} \boldsymbol{a}_l \right| \\
&\le 2C_4 \left(\frac{\mu}{\sqrt{m}} \log^2 m + \sigma \right) \cdot 20 \sqrt{\log m} \cdot C_2 \left(\frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^9 m}{m}} + \frac{\sigma}{\log^2 m} \right) \\
&\quad + 2C_4 \left(\frac{\mu}{\sqrt{m}} \log^2 m + \sigma \right) \cdot C_1 \left(\sqrt{\frac{\mu^2 K \log m}{m}} + \lambda + \sigma \sqrt{K \log m} \right) \cdot 20 \sqrt{\log m} \\
&\quad + 2C_4 \left(\frac{\mu}{\sqrt{m}} \log^2 m + \sigma \right) \cdot 20 \sqrt{\log m} \\
&\lesssim C_4 \left(\frac{\mu}{\sqrt{m}} \log^{2.5} m + \sigma \sqrt{\log m} \right),
\end{aligned} \tag{A.36d}$$

where the second inequality follows from triangle inequality and (D.1); the penultimate inequality follows from (A.36c), (A.15a), (A.16a) and (D.1); the last line holds as long as $m \gg \mu^2 K \log m$. Substituting (A.36d) into (A.36a) and (A.36b), we reach

$$\nu_{31} + \nu_{32} \lesssim \left(40 \sqrt{\frac{K \log m}{m}} + 20 C_4 \left(\frac{\mu \sqrt{K}}{\sqrt{m}} \log^2 m + \sigma \sqrt{K} \right) \right) C_4 \left(\frac{\mu}{\sqrt{m}} \log^{2.5} m + \sigma \sqrt{\log m} \right)$$

$$\le (C_4)^2 \frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^9 m}{m}} + C_4 \frac{\sigma}{\log^2 m}, \tag{A.36e}$$

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

$$\left\| \xi_l \boldsymbol{b}_l \boldsymbol{a}_l^{\mathsf{H}} \widehat{\boldsymbol{x}}^{t,(l)} \right\|_2 \le |\xi_l| \left\| \boldsymbol{b}_l \right\|_2 \left| \boldsymbol{a}_l^{\mathsf{H}} \widehat{\boldsymbol{x}}^{t,(l)} \right| \overset{\text{(i)}}{\lesssim} \sigma \sqrt{\frac{K}{m}} \left\| \widehat{\boldsymbol{x}}^{t,(l)} \right\|_2 \log m \overset{\text{(ii)}}{\le} 2\sigma \sqrt{\frac{K}{m}} \log m, \tag{A.36f}$$

$$\begin{aligned}
\left\| \overline{\xi_l} \boldsymbol{a}_l \boldsymbol{b}_l^{\mathsf{H}} \widehat{\boldsymbol{h}}^{t,(l)} \right\|_2 &\le |\xi_l| \left\| \boldsymbol{a}_l \right\|_2 \left| \boldsymbol{b}_l^{\mathsf{H}} \widehat{\boldsymbol{h}}^{t,(l)} \right| \overset{\text{(iii)}}{\lesssim} \sigma \sqrt{K} \left(\left| \boldsymbol{b}_l^{\mathsf{H}} \left(\widehat{\boldsymbol{h}}^{t,(l)} - \boldsymbol{h}^\star \right) \right| + \left| \boldsymbol{b}_l^{\mathsf{H}} \boldsymbol{h}^\star \right| \right) \\
&\overset{\text{(iv)}}{\lesssim} \sigma \sqrt{K} \left(2C_4 \left(\frac{\mu}{\sqrt{m}} \log^2 m + \sigma \right) + \frac{\mu}{\sqrt{m}} \right) \\
&\lesssim C_4 \frac{\sigma}{\log^{2.5} m} + C_4 \sigma \sqrt{\frac{\mu^2 K \log^4 m}{m}},
\end{aligned} \tag{A.36g}$$

where (i) holds by the property of sub-Gaussian variables (cf. [Ver18, Proposition 2.5.2]) and the independence between $\xi_l, \boldsymbol{a}_l$ and $\widehat{\boldsymbol{x}}^{t,(l)}$, (ii) holds by (A.16f), (iii) is due to Lemma (18), the triangle inequality and (1.7), and (iv) follows from (A.36c) and (1.7). Consequently, by (A.36e)-(A.36g) we have

$$\left\| \boldsymbol{\nu}_3 \right\|_2 \lesssim (C_4)^2 \frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^9 m}{m}} + C_4 \frac{\sigma}{\log^2 m}. \tag{A.37}$$

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

$$\|\boldsymbol{\nu}_4\|_2 = \left\| \begin{bmatrix} \widehat{\boldsymbol{h}}^{t,(l)} - \widetilde{\boldsymbol{h}}^t \\ \widehat{\boldsymbol{x}}^{t,(l)} - \widetilde{\boldsymbol{x}}^t \end{bmatrix} \right\|_2 = \left\| \widehat{\boldsymbol{z}}^{t,(l)} - \widetilde{\boldsymbol{z}}^t \right\|_2. \tag{A.38}$$

With the above bounds in place, we can demonstrate that

$$\begin{aligned}
\mathsf{dist}\big(\boldsymbol{z}^{t+1,(l)}, \widetilde{\boldsymbol{z}}^{t+1}\big) &\le \max\left\{ \left| \frac{\alpha^{t+1/2}}{\alpha^t} \right|, \left| \frac{\alpha^t}{\alpha^{t+1/2}} \right| \right\} \left\| \begin{bmatrix} \frac{1}{\alpha_{\text{mutual}}^{t,(l)}} \boldsymbol{h}^{t+1/2,(l)} - \frac{1}{\alpha^t} \boldsymbol{h}^{t+1/2} \\ \alpha_{\text{mutual}}^{t,(l)} \boldsymbol{x}^{t+1/2,(l)} - \alpha^t \boldsymbol{x}^{t+1/2} \end{bmatrix} \right\|_2 \\
&\overset{\text{(i)}}{\le} \frac{1-\eta/32}{1-\eta/16} \left(\|\boldsymbol{\nu}_1\|_2 + \|\boldsymbol{\nu}_2\|_2 + \|\boldsymbol{\nu}_3\|_2 + \|\boldsymbol{\nu}_4\|_2 \right) \\
&\overset{\text{(ii)}}{\le} (1-\eta/32) \left\| \widehat{\boldsymbol{z}}^{t,(l)} - \widetilde{\boldsymbol{z}}^t \right\|_2 + \frac{1-\eta/32}{1-\eta/16} C\eta \times C_1 \frac{1}{\log^2 m} \left\| \widehat{\boldsymbol{z}}^{t,(l)} - \widetilde{\boldsymbol{z}}^t \right\|_2 \\
&\quad + \frac{1-\eta/32}{1-\eta/16} C\eta \left((C_4)^2 \frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^9 m}{m}} + C_4 \frac{\sigma}{\log^2 m} \right) + \frac{1-\eta/32}{1-\eta/16} \eta\lambda \left\| \widehat{\boldsymbol{z}}^{t,(l)} - \widetilde{\boldsymbol{z}}^t \right\|_2 \\
&\le \left(1 - \eta/32 + \frac{1-\eta/32}{1-\eta/16} \eta\lambda + \frac{1-\eta/32}{1-\eta/16} C C_1 \frac{\eta}{\log^2 m} \right) \left\| \widehat{\boldsymbol{z}}^{t,(l)} - \widetilde{\boldsymbol{z}}^t \right\|_2 \\
&\quad + \frac{1-\eta/32}{1-\eta/16} C\eta \left((C_4)^2 \frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^9 m}{m}} + C_4 \frac{\sigma}{\log^2 m} \right) \\
&\le \left(1 - \frac{\eta}{64} \right) \mathsf{dist}\big(\boldsymbol{z}^{t,(l)}, \widetilde{\boldsymbol{z}}^t\big) + \eta C (C_4)^2 \frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^9 m}{m}} + \eta C C_4 \frac{\sigma}{\log^2 m} \\
&\le C_2 \left(\frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^9 m}{m}} + \frac{\sigma}{\log^2 m} \right),
\end{aligned} \tag{A.39}$$

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

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

as shown in Corollary 1, which implies that

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

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

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

In addition, (ii) follows from (A.35), (A.37) and (A.38), whereas the last inequality of (A.39) relies on the hypothesis (A.15a).

Next, we turn to the second inequality claimed in the lemma. In view of (A.16a) in Lemma 5, we have

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

which together with the triangle inequality and (A.39) yields

$$\left\| \widehat{\boldsymbol{z}}^{t+1,(l)} - \boldsymbol{z}^\star \right\|_2 \le \left\| \widehat{\boldsymbol{z}}^{t+1,(l)} - \widetilde{\boldsymbol{z}}^{t+1} \right\|_2 + \left\| \widetilde{\boldsymbol{z}}^{t+1} - \boldsymbol{z}^\star \right\|_2$$

$$\leq C_2 \left(\frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^9 m}{m}} + \frac{\sigma}{\log^2 m} \right) + C_1 \left(\sqrt{\frac{\mu^2 K \log m}{m}} + \lambda + \sigma \sqrt{K \log m} \right)$$
$$\lesssim \sqrt{\frac{\mu^2 K \log m}{m}} + \sigma \sqrt{K \log m} + \lambda. \tag{A.40}$$

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 [MWCC17, 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 [MWCC17, Lemma 55] yields

$$\left\| \widetilde{\boldsymbol{z}}^{t+1,(l)} - \widetilde{\boldsymbol{z}}^{t+1} \right\|_2 \lesssim \left\| \widehat{\boldsymbol{z}}^{t+1,(l)} - \widetilde{\boldsymbol{z}}^{t+1} \right\|_2 \leq C_2 \left(\frac{\mu}{\sqrt{m}} \sqrt{\frac{\mu^2 K \log^9 m}{m}} + \frac{\sigma}{\log^2 m} \right), \tag{A.41}$$

where the last inequality follows from (A.40).

A.9 Proof of Lemma 9

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

$$\left| \frac{\alpha^{t+1/2}}{\alpha^t} - 1 \right| \leq C\eta \left(\sqrt{\frac{\mu^2 K \log m}{m}} + \lambda + \sigma \sqrt{K \log m} \right) =: \delta, \tag{A.42}$$

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

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

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

$$\frac{1}{\overline{\alpha^t}} \boldsymbol{h}^{t+1/2} = \widetilde{\boldsymbol{h}}^t - \frac{\eta}{|\alpha^t|^2} \sum_{j=1}^m \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \left(\widetilde{\boldsymbol{h}}^t \widetilde{\boldsymbol{x}}^{t\mathsf{H}} - \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}} \right) \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \widetilde{\boldsymbol{x}}^t + \frac{\eta}{|\alpha^t|^2} \sum_{j=1}^m \xi_j \boldsymbol{b}_j \boldsymbol{a}_j^{\mathsf{H}} \widetilde{\boldsymbol{x}}^t - \eta\lambda \widetilde{\boldsymbol{h}}^t$$
$$= \underbrace{\left(1 - \eta\lambda - \frac{\eta}{|\alpha^t|^2} \|\boldsymbol{x}^\star\|_2^2 \right) \widetilde{\boldsymbol{h}}^t}_{=:\boldsymbol{\nu}_0} - \frac{\eta}{|\alpha^t|^2} \underbrace{\sum_{j=1}^m \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \widetilde{\boldsymbol{h}}^t \left(\left| \boldsymbol{a}_j^{\mathsf{H}} \widetilde{\boldsymbol{x}}^t \right|^2 - \left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 \right)}_{=:\boldsymbol{\nu}_1}$$
$$- \frac{\eta}{|\alpha^t|^2} \underbrace{\sum_{j=1}^m \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \widetilde{\boldsymbol{h}}^t \left(\left| \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{x}^\star \right|^2 - \|\boldsymbol{x}^\star\|_2^2 \right)}_{=:\boldsymbol{\nu}_2} + \frac{\eta}{|\alpha^t|^2} \underbrace{\sum_{j=1}^m \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \widetilde{\boldsymbol{x}}^t}_{=:\boldsymbol{\nu}_3} + \frac{\eta}{|\alpha^t|^2} \underbrace{\sum_{j=1}^m \xi_j \boldsymbol{b}_j \boldsymbol{a}_j^{\mathsf{H}} \widetilde{\boldsymbol{x}}^t}_{=:\boldsymbol{\nu}_4}. \tag{A.43}$$

The first three terms can be controlled via the same arguments as [MWCC17, 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 [MWCC17, 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 [MWCC17, Appendix C.4] are listed below:

$$\left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{\nu}_1\right| \le 0.1 \max_{1\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right|, \tag{A.44a}$$

$$\left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{\nu}_2\right| \le 0.2 \max_{1\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right| + \max_{1\le j\le\tau}\left|(\boldsymbol{b}_j - \boldsymbol{b}_1)^{\mathsf{H}}\, \widetilde{\boldsymbol{h}}^t\right| \log m, \tag{A.44b}$$

$$\left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{\nu}_3\right| \lesssim \frac{\mu}{\sqrt{m}} + \frac{\mu}{\sqrt{m}}\log^{3/2} m \max_{1\le j\le m}\left|\boldsymbol{a}_j^{\mathsf{H}}(\widetilde{\boldsymbol{x}}^t - \boldsymbol{x}^\star)\right|. \tag{A.44c}$$

When it comes to the last term of (A.43) 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

$$\begin{aligned}
\varsigma_1 &\le \sum_{j=1}^m \left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{b}_j\right| \cdot \max_{1\le j\le m}|\xi_j| \cdot \max_{1\le j\le m}\left|\boldsymbol{a}_j^{\mathsf{H}}(\widetilde{\boldsymbol{x}}^t - \boldsymbol{x}^\star)\right| \\
&\lesssim (4\log m)\cdot \sigma\sqrt{\log m}\cdot \max_{1\le j\le m}\left|\boldsymbol{a}_j^{\mathsf{H}}(\widetilde{\boldsymbol{x}}^t - \boldsymbol{x}^\star)\right| \\
&\asymp \sigma \log^{1.5} m \max_{1\le j\le m}\left|\boldsymbol{a}_j^{\mathsf{H}}(\widetilde{\boldsymbol{x}}^t - \boldsymbol{x}^\star)\right|,
\end{aligned}$$

  where the second inequality follows from [MWCC17, 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 Bernstein's inequality [Ver18, Theorem 2.8.2], we have

$$\mathbb{P}\left\{\left|\sum_{j=1}^m \xi_j \boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}^\star\right| \ge t\right\} \le 2\exp\left(-c\min\left(\frac{\tau^2}{\sigma^2\sum_{j=1}^m\left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{b}_j\right|^2}, \frac{\tau}{\sigma\max_{1\le j\le m}\left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{b}_j\right|}\right)\right) \tag{A.45}$$

  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 $1 - O\left(m^{-100}\right)$,

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

$$\le \sigma \log^{1.5} m \cdot C_3\left(\sqrt{\frac{\mu^2 K\log^2 m}{m}} + \sqrt{\log m}\left(\lambda + \sigma\sqrt{K\log m}\right)\right) + \sigma\sqrt{\frac{K}{m}\log m}$$

$$\lesssim C_3\sigma. \tag{A.47}$$

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

Combining the bounds (A.44) with (A.43) and (A.46), we arrive at

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

as long as $m \gg \mu^2 K\log^9 m$ 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

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

with $\delta \ll 1$ defined in (A.42). This inequality allows us to turn attention to $\frac{1}{\overline{\alpha^{t-1}}}\left(\boldsymbol{b}_j - \boldsymbol{b}_1\right)^{\mathsf{H}}\boldsymbol{h}^{t-1/2}$ 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

$$\begin{aligned}
\left(\boldsymbol{b}_j - \boldsymbol{b}_1\right)^{\mathsf{H}}\frac{1}{\overline{\alpha^{t-1}}}\boldsymbol{h}^t &= \left(1 - \eta\lambda - \frac{\eta}{|\alpha^t|^2}\left\|\widetilde{\boldsymbol{x}}^{t-1}\right\|_2^2\right)\left(\boldsymbol{b}_j - \boldsymbol{b}_1\right)^{\mathsf{H}}\widetilde{\boldsymbol{h}}^{t-1} + \frac{\eta}{|\alpha^t|^2}\underbrace{\left(\boldsymbol{b}_j - \boldsymbol{b}_1\right)\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\widetilde{\boldsymbol{x}}^{t-1}}_{=:\beta_1} \\
&\quad - \frac{\eta}{|\alpha^t|^2}\underbrace{\left(\boldsymbol{b}_j - \boldsymbol{b}_1\right)^{\mathsf{H}}\sum_{l=1}^m \boldsymbol{b}_l\boldsymbol{b}_l^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^{t-1}\widetilde{\boldsymbol{x}}^{t-1\mathsf{H}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right)\left(\boldsymbol{a}_l\boldsymbol{a}_l^{\mathsf{H}} - \boldsymbol{I}_k\right)\widetilde{\boldsymbol{x}}^{t-1}}_{=:\beta_2} \\
&\quad + \frac{\eta}{|\alpha^t|^2}\underbrace{\left(\boldsymbol{b}_j - \boldsymbol{b}_1\right)^{\mathsf{H}}\sum_{l=1}^m \xi_l\boldsymbol{b}_l\boldsymbol{a}_l^{\mathsf{H}}\widetilde{\boldsymbol{x}}^{t-1}}_{=:\beta_3}.
\end{aligned} \tag{A.48}$$

Except β_3, the bounds of the other terms can be obtained by the same arguments as in [MWCC17, Appendix C.4.3]; we thus omit the detailed proof but only list the results below:

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

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

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

- Regarding ω_1, we have

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

  where the second inequality follows from [MWCC17, Lemma 50] and standard sub-Gaussian concentration inequalities.

- For ω_2, similar to (A.45), we can invoke Bernstein's inequality [Ver18, Theorem 2.8.2] to reach

$$\begin{aligned}&\mathbb{P}\left\{\left|\sum_{l=1}^{m}\xi_l\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\boldsymbol{b}_l\boldsymbol{a}_l^{\mathsf{H}}\boldsymbol{x}^{\star}\right|\ge\tau\right\}\\ &\quad\le 2\exp\left(-c\min\left(\frac{\tau^2}{\sigma^2\sum_{l=1}^{m}\left|\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\boldsymbol{b}_l\right|^2},\frac{\tau}{\sigma\max_{1\le j\le m}\left|\left(\boldsymbol{b}_j-\boldsymbol{b}_1\right)^{\mathsf{H}}\boldsymbol{b}_l\right|}\right)\right)\end{aligned}\tag{A.49}$$

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

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

  where the last inequality follows from [MWCC17, Lemma 48, 49]. Taking $\tau = C\sigma\sqrt{K\log^2 m/m}$ in (A.49) for some large enough constant $C > 0$, one arrives at

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

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

$$|\beta_3| \lesssim \frac{\sigma}{\log^{1.5} m}\max_{1\le j\le m}\left|\boldsymbol{a}_j^{\mathsf{H}}(\widetilde{\boldsymbol{x}}^t-\boldsymbol{x}^{\star})\right| + \sigma\sqrt{\frac{K\log m}{m}}$$

$$
\begin{aligned}
&\lesssim C_3 \frac{\sigma}{\log^{1.5} m}\left(\sqrt{\frac{\mu^2 K \log^2 m}{m}}+\sqrt{\log m}\left(\lambda+\sigma\sqrt{K\log m}\right)\right)+\sigma\sqrt{\frac{K\log m}{m}} \\
&\lesssim \frac{\sigma}{\log^3 m}. \qquad\qquad \text{(A.50)}
\end{aligned}
$$

Putting together the above results, we demonstrate that

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

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 5.

B Analysis: connections between convex and nonconvex solutions

In this section, we establish the key lemmas for justifying Theorem 1.

B.1 Preliminary facts

Before proceeding, there are a couple of immediate consequences of Lemma 1 that will prove useful, which we summarize as follows.

Lemma 13. *Instate the notation and assumptions in Theorem 2. For $t \ge 0$, suppose that the hypotheses (B.2) hold in the first t iterations. Then there exist some constants $C_1, C > 0$ such that for any $1 \le l \le m$,*

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

$$\left\|\widetilde{\boldsymbol{z}}^{t,(l)}-\boldsymbol{z}^\star\right\|_2 \le 2\frac{C_5}{c_\rho}\left(\lambda+\sigma\sqrt{K\log m}\right), \tag{B.1b}$$

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

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

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

$$\left\|\boldsymbol{h}^t\right\|_2^2=\left\|\boldsymbol{x}^t\right\|_2^2=\left\|\boldsymbol{h}^t\right\|_2\left\|\boldsymbol{x}^t\right\|_2=\|\widetilde{\boldsymbol{h}}^{t-1/2}\|_2\|\widetilde{\boldsymbol{x}}^{t-1/2}\|_2=\|\widetilde{\boldsymbol{h}}^t\|_2\|\widetilde{\boldsymbol{x}}^t\|_2. \tag{B.1f}$$

In addition, for an integer $t > 0$, suppose that the hypotheses (B.2) hold in the first $t-1$ iterations. Then there exists some constant $C > 0$ such that

$$\left\|\widehat{\boldsymbol{z}}^t-\boldsymbol{z}^\star\right\|_2 \le \frac{C_5}{c_\rho}\left(\lambda+\sigma\sqrt{K\log m}\right), \tag{B.1g}$$

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

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

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

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

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

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

Proof. The proof follows from the same argument as in the proof of Lemma 5 and Corollary 1, and is thus omitted here for brevity. $\square$

B.2 Proof of Lemma 1

After the introduction of the proof idea in Appendix A, we state a more complete version of Lemma 1 here.

Lemma 14. *Take* $\lambda = C_\lambda \sigma \sqrt{K \log m}$ *for some large enough constant* $C_\lambda > 0$. *Assume the number of measurements 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* $c > 0$. *Then, with probability at least* $1 - O\left(m^{-100} + me^{-cK}\right)$ *for some constant* $c > 0$, *the iterates* $\{\boldsymbol{h}^t, \boldsymbol{x}^t\}_{0<t\leq t_0}$ *of Algorithm (2) satisfy*

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

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

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

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

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

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

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

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

The claims (B.2a)-(B.2e) are direct consequences of Lemma 7, Lemma 8, the relation (A.22), and Lemma 9. As a result, the remaining steps lie in proving (B.2f) and (B.2g).

B.2.1 Proof of the claim (B.2f)

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

$$\max_{1\leq j\leq m} \left|\boldsymbol{b}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^t - \boldsymbol{h}^\star\right)\right| \leq C_9 \sigma \tag{B.3}$$

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

It is self-evident that (B.3) holds for the base case (i.e. $t = 0$) when $\boldsymbol{h}^0 = \boldsymbol{h}^\star$. Assume for the moment that (B.3) 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 (2.1), 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.4}
$$

where (i) comes from the gradient update rule (2.1) and (ii) is due to the expression (A.7).

- Applying a similar argument as for [MWCC17, Equation (219)] yields

$$\left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{\nu}_1\right|\le 0.1\max_{1\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\big(\widetilde{\boldsymbol{h}}^t-\boldsymbol{h}^\star\big)\right|.$$

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

$$
\begin{aligned}
\left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{\nu}_2\right|&\le 0.2\max_{1\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\big(\widetilde{\boldsymbol{h}}^t-\boldsymbol{h}^\star\big)\right|+C\log m\max_{0\le l\le m-\tau,1\le j\le\tau}\left|\left(\boldsymbol{b}_{l+j}-\boldsymbol{b}_{l+1}\right)^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^t-\boldsymbol{h}^\star\right)\right|\\
&\le 0.2\max_{1\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\big(\widetilde{\boldsymbol{h}}^t-\boldsymbol{h}^\star\big)\right|+\left(C\log m\right)C_{11}\frac{\sigma}{\log^3 m}.
\end{aligned}
$$

  The first inequality can be derived via a similar argument as in [MWCC17, 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

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

Proof. See Appendix B.2.3. ∎

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

$$\left|\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{\nu}_3\right|\le\left|\sum_{j=1}^m\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\left(\widetilde{\boldsymbol{x}}^t-\boldsymbol{x}^\star\right)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^t-\boldsymbol{x}^\star\right)\right|+\left|\sum_{j=1}^m\boldsymbol{b}_l^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\left(\widetilde{\boldsymbol{x}}^t-\boldsymbol{x}^\star\right)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}^\star\right|$$

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

Here, the penultimate inequality follows from the incoherence condition (B.2d) and Lemma 4, whereas the last inequality follows from the induction hypothesis (B.2d).

- Finally, we turn to the term $\boldsymbol{\nu}_4$. Clearly, it is of the same form as $\boldsymbol{\nu}_4$ in (A.43); therefore, via the same line of analysis, one can deduce the following bound (similar to (A.46))

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

where the last inequality invokes (B.2d).

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

$$
\begin{aligned}
&\left|\frac{\overline{\alpha^{t+1/2}}}{\overline{\alpha^t}}\right|\max_{1\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^{t+1}-\boldsymbol{h}^{\star}\right)\right| \\
&\le \left(1-\eta\lambda-\frac{\overline{\alpha^{t+1/2}}}{\overline{\alpha^t}}\right)\max_{1\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^{\star}\right| + \left(1-\eta\lambda-\frac{\eta}{|\alpha^t|^2}\right)\max_{1\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^t-\boldsymbol{h}^{\star}\right)\right| \\
&\quad + \frac{\eta}{|\alpha^t|^2}\left(0.3\max_{1\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^t-\boldsymbol{h}^{\star}\right)\right| + \log m\times C_{11}\frac{\sigma}{\log^3 m}\right) \\
&\quad + \frac{\eta}{|\alpha^t|^2}CC_7\frac{\mu}{\sqrt{m}}\log^2 m\left(\lambda+\sigma\sqrt{K\log m}\right) + \frac{\eta C}{|\alpha^t|^2}\left(\sigma\log^{1.5} m\left(C_7\sqrt{\log m}\left(\lambda+\sigma\sqrt{K\log m}\right)\right)+\sigma\sqrt{\frac{K}{m}\log m}\right) \\
&\overset{\text{(i)}}{\le} \left(1-\frac{7\eta}{40}\right)\max_{1\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^t-\boldsymbol{h}^{\star}\right)\right| + \left(\eta\lambda+\left|1-\frac{\alpha^{t+1/2}}{\alpha^t}\right|\right)\frac{\mu}{\sqrt{m}} + \frac{4C_{11}\eta\sigma}{\log^2 m} + CC_7\frac{\mu}{\sqrt{m}}\log^2 m\left(\lambda+\sigma\sqrt{K\log m}\right) \\
&\quad + 4\eta C\left[\sigma\log^{1.5} m\left(C_7\sqrt{\log m}\left(\lambda+\sigma\sqrt{K\log m}\right)\right)+\sigma\sqrt{\frac{K}{m}\log m}\right] \\
&\le \left(1-\frac{7\eta}{40}\right)C_9\sigma + c\eta\sigma,
\end{aligned}
$$

for some constant $C>0$ and sufficiently small constant $c>0$. Here (i) uses triangle inequality and (B.1l) 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.1i) we obtain

$$
\begin{aligned}
\max_{1\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^{t+1}-\boldsymbol{h}^{\star}\right)\right| &\le \frac{\left(1-\frac{7\eta}{40}\right)C_9\sigma+c\eta\sigma}{\left|\frac{\overline{\alpha^{t+1/2}}}{\overline{\alpha^t}}\right|} \le \frac{\left(1-\frac{7\eta}{40}\right)C_9\sigma+c\eta\sigma}{1-\left|\frac{\alpha^{t+1/2}}{\alpha^t}-1\right|} \\
&\le \frac{\left(1-\frac{7\eta}{40}\right)C_9\sigma+c\eta\sigma}{1-\eta\frac{CC_5}{c_\rho}\left(\lambda+\sigma\sqrt{K\log m}\right)} \\
&\le C_9\sigma,
\end{aligned}
$$

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

B.2.2 Proof of the claim (B.2g)

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

Claim 3. If the iterates satisfy the induction hypotheses (B.2a)-(B.2e) in the tth iteration, then with probability exceeding $1 - O\left(m^{-100} + e^{-CK} \log m\right)$,

$$f\left(\boldsymbol{h}^{t+1}, \boldsymbol{x}^{t+1}\right) \leq f\left(\boldsymbol{h}^{t}, \boldsymbol{x}^{t}\right) - \frac{\eta}{2} \left\|\nabla f\left(\boldsymbol{h}^{t}, \boldsymbol{x}^{t}\right)\right\|_2^2 . \tag{B.5}$$

Proof. See Appendix B.2.4. □

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

$$\min_{0\leq t<t_0} \left\|\nabla f\left(\boldsymbol{z}^{t}\right)\right\|_2 \leq \left\{\frac{1}{t_0} \sum_{t=0}^{t_0-1} \left\|\nabla f\left(\boldsymbol{z}^{t}\right)\right\|_2^2\right\}^{1/2} \leq \left\{\frac{2}{\eta t_0}\left[f\left(\boldsymbol{z}^{\star}\right) - f\left(\boldsymbol{z}^{t_0}\right)\right]\right\}^{1/2}, \tag{B.6}$$

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. [MWCC17, Appendix D.3.1]), we can write

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

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

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

The last inequality in the above formula invokes Lemma 6, whose assumptions are verified in the proof of Claim 3 (see Appendix (B.2.4)). Further, the relations (B.17) and (B.11) in the proof of Claim 3 lead to

$$f\left(\boldsymbol{z}^{\star}\right) - f\left(\boldsymbol{z}^{t_0}\right) \lesssim \left(\lambda + \sigma\sqrt{K \log m}\right)^2 . \tag{B.7}$$

It then follows from (B.6) and (B.7) 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.2.3 Proof of Claim 2

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

$$\max_{0\leq l\leq m-\tau,\, 1\leq j\leq \tau} \left|\left(\boldsymbol{b}_{l+j} - \boldsymbol{b}_{l+1}\right)^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^{t} - \boldsymbol{h}^{\star}\right)\right| \leq C_{11}\frac{\sigma}{\log^3 m} . \tag{B.8}$$

Apparently, (B.8) holds when $t = 0$ given that $\boldsymbol{h}^0 = \boldsymbol{h}^{\star}$. In what follows, we shall assume that (B.8) holds true at the tth iteration, and examine this condition for the $(t+1)$th iteration.

Similar to the derivation of (B.4), 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

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

  where the second inequality follows from [MWCC17, Lemma 50] and the last inequality utilizes the following consequence of (B.2d) and Lemma 18:

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

- With regards to $\boldsymbol{\nu}_2$, we invoke the induction hypothesis (B.2d) at the tth iteration to obtain

$$
\begin{aligned}
\left|\left(\boldsymbol{b}_{l}-\boldsymbol{b}_{1}\right)^{\mathsf{H}}\boldsymbol{\nu}_{2}\right| &\leq \sum_{j=1}^{m}\left|\left(\boldsymbol{b}_{l}-\boldsymbol{b}_{1}\right)^{\mathsf{H}}\boldsymbol{b}_{j}\right|\max_{1\leq j\leq m}\left|\boldsymbol{b}_{j}^{\mathsf{H}}\boldsymbol{h}^{\star}\right|\max_{1\leq j\leq m}\left|\left(\widetilde{\boldsymbol{x}}^{t}-\boldsymbol{x}^{\star}\right)^{\mathsf{H}}\boldsymbol{a}_{j}\boldsymbol{a}_{j}^{\mathsf{H}}\widetilde{\boldsymbol{x}}^{t}\right|\\
&\leq \frac{c}{\log^{2}m}\frac{\mu}{\sqrt{m}}\left(\max_{1\leq j\leq m}\left|\left(\widetilde{\boldsymbol{x}}^{t}-\boldsymbol{x}^{\star}\right)^{\mathsf{H}}\boldsymbol{a}_{j}\right|^{2}+\max_{1\leq j\leq m}\left|\left(\widetilde{\boldsymbol{x}}^{t}-\boldsymbol{x}^{\star}\right)^{\mathsf{H}}\boldsymbol{a}_{j}\right|\max_{1\leq j\leq m}\left|\boldsymbol{a}_{j}^{\mathsf{H}}\boldsymbol{x}^{\star}\right|\right)\\
&\lesssim C_{8}\frac{\mu}{\log m\sqrt{m}}\left(\lambda+\sigma\sqrt{\log m}\right),
\end{aligned}
$$

  where the second inequality applies [MWCC17, Lemma 50] and (1.7), and the last inequality results from (B.2d) and (D.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.48), we can apply the analysis leading to (A.50) to derive

$$
\begin{aligned}
\left|\left(\boldsymbol{b}_{l}-\boldsymbol{b}_{1}\right)^{\mathsf{H}}\boldsymbol{\nu}_{3}\right| &\lesssim \frac{\sigma}{\log^{1.5}m}\max_{1\leq j\leq m}\left|\boldsymbol{a}_{j}^{\mathsf{H}}(\widetilde{\boldsymbol{x}}^{t}-\boldsymbol{x}^{\star})\right|+\sigma\sqrt{\frac{K\log^{2}m}{m}}\\
&\lesssim \frac{\sigma}{\log^{1.5}m}\left(C_{7}\sqrt{\log m}\left(\lambda+\sigma\sqrt{K\log m}\right)\right)+\sigma\sqrt{\frac{K\log^{2}m}{m}}
\end{aligned}
$$

With the preceding results in hand, we have

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

$$
\begin{aligned}
&\leq \left|1-\eta\lambda-\frac{\overline{\alpha^{t+1/2}}}{\overline{\alpha^{t}}}\right| \max_{0\leq l\leq m-\tau,1\leq j\leq\tau}\left|\left(\boldsymbol{b}_{l+j}-\boldsymbol{b}_{l+1}\right)^{\mathsf{H}}\boldsymbol{h}^{\star}\right| \\
&\quad+\left(1-\eta\lambda-\eta\left\|\boldsymbol{x}^{t}\right\|_{2}^{2}\right)\max_{0\leq l\leq m-\tau,1\leq j\leq\tau}\left|\left(\boldsymbol{b}_{l+j}-\boldsymbol{b}_{l+1}\right)^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^{t}-\boldsymbol{h}^{\star}\right)\right| \\
&\quad+\frac{\eta C C_{9}}{|\alpha^{t}|^{2}}\frac{\mu\log m}{\sqrt{m}}\left(\lambda+\sigma\sqrt{K\log m}\right)+\frac{\eta C C_{8}}{|\alpha^{t}|^{2}}\left(\frac{\mu}{\log m\sqrt{m}}\left(\lambda+\sigma\sqrt{K\log m}\right)\right) \\
&\quad+\frac{\eta C}{|\alpha^{t}|^{2}}\left[\frac{\sigma}{\log^{1.5}m}\left(C_{7}\sqrt{\log m}\left(\lambda+\sigma\sqrt{K\log m}\right)\right)+\sigma\sqrt{\frac{K\log^{2}m}{m}}\right] \\
&\overset{\text{(i)}}{\leq}\left(\eta\lambda+\left|1-\frac{\alpha^{t+1/2}}{\alpha^{t}}\right|\right)\frac{2\mu}{\sqrt{m}}+\left(1-\frac{\eta}{16}\right)C_{11}\frac{\sigma}{\log^{3}m} \\
&\quad+4\eta C C_{9}\frac{\mu\log m}{\sqrt{m}}\left(\lambda+\sigma\sqrt{K\log m}\right)+4\eta C C_{8}\left(\frac{\mu}{\log m\sqrt{m}}\left(\lambda+\sigma\sqrt{K\log m}\right)\right) \\
&\quad+4\eta C\left[\frac{\sigma}{\log^{1.5}m}\left(C_{7}\sqrt{\log m}\left(\lambda+\sigma\sqrt{K\log m}\right)\right)+\sigma\sqrt{\frac{K\log^{2}m}{m}}\right] \\
&\overset{\text{(ii)}}{\leq}\left(1-\frac{\eta}{16}\right)\frac{C_{11}\sigma}{\log^{3}m}+c\frac{\eta\sigma}{\log^{3}m}
\end{aligned}
$$

for some constant $C>0$ and some sufficiently small constant $c>0$. Here, the relation (i) comes from the triangle inequality, (B.1l), as well as the consequence of (B.1c) and (B.1l)

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

the inequality (ii) invokes (B.1i) and holds with the proviso that $m\gg\mu^{2}K\log^{8}m$ and $\sigma\sqrt{K\log^{5}m}\ll1$.

Finally, by (B.1i) we obtain

$$
\begin{aligned}
\max_{0\leq l\leq m-\tau,1\leq j\leq\tau}\left|\left(\boldsymbol{b}_{l+j}-\boldsymbol{b}_{l+1}\right)^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^{t+1}-\boldsymbol{h}^{\star}\right)\right| &\leq\frac{\left(1-\frac{\eta}{16}\right)C_{11}\frac{\sigma}{\log^{3}m}+c\frac{\eta\sigma}{\log^{3}m}}{\left|\frac{\overline{\alpha^{t+1/2}}}{\overline{\alpha^{t}}}\right|} \\
&\leq\frac{\left(1-\frac{\eta}{16}\right)C_{11}\frac{\sigma}{\log^{3}m}+c\frac{\eta\sigma}{\log^{3}m}}{1-\left|\frac{\alpha^{t+1/2}}{\alpha^{t}}-1\right|} \\
&\leq\frac{\left(1-\frac{\eta}{16}\right)C_{11}\frac{\sigma}{\log^{3}m}+c\frac{\eta\sigma}{\log^{3}m}}{1-\eta\frac{CC_{5}}{c_{\rho}}\left(\lambda+\sigma\sqrt{K\log m}\right)} \\
&\leq C_{11}\frac{\sigma}{\log^{3}m},
\end{aligned}
$$

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

B.2.4 Proof of Claim 3

Before proceeding, we note that

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

and

$$
\left[\begin{array}{c}\nabla_{\boldsymbol{h}}f\left(\frac{\boldsymbol{h}}{\overline{\alpha}},\alpha\boldsymbol{x}\right)\\ \nabla_{\boldsymbol{x}}f\left(\frac{\boldsymbol{h}}{\overline{\alpha}},\alpha\boldsymbol{x}\right)\end{array}\right]=\left[\begin{array}{c}\alpha\nabla_{\boldsymbol{h}}f_{\textsf{reg-free}}\left(\boldsymbol{h},\boldsymbol{x}\right)\\ \frac{1}{\overline{\alpha}}\nabla_{\boldsymbol{x}}f_{\textsf{reg-free}}\left(\boldsymbol{h},\boldsymbol{x}\right)\end{array}\right]+\lambda\left[\begin{array}{c}\frac{\boldsymbol{h}}{\overline{\alpha}}\\ \alpha\boldsymbol{x}\end{array}\right]. \tag{B.9}
$$

Another fact of use is that

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

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

$$\begin{aligned}
\left\|\nabla f\left(\overline{\boldsymbol{h}}^t, \overline{\boldsymbol{x}}^t\right)\right\|_2 &= \left\| \begin{bmatrix} \beta^t \nabla_{\boldsymbol{h}} f_{\textsf{reg-free}}\left(\boldsymbol{h}^t, \boldsymbol{x}^t\right) \\ \frac{1}{\overline{\beta^t}} \nabla_{\boldsymbol{x}} f_{\textsf{reg-free}}\left(\boldsymbol{h}^t, \boldsymbol{x}^t\right) \end{bmatrix} + \lambda \begin{bmatrix} \frac{\boldsymbol{h}^t}{\overline{\beta^t}} \\ \beta^t \boldsymbol{x}^t \end{bmatrix} \right\|_2 \\
&= \left\| \begin{bmatrix} \nabla_{\boldsymbol{h}} f_{\textsf{reg-free}}\left(\boldsymbol{h}^t, \boldsymbol{x}^t\right) \\ \nabla_{\boldsymbol{x}} f_{\textsf{reg-free}}\left(\boldsymbol{h}^t, \boldsymbol{x}^t\right) \end{bmatrix} + \lambda \begin{bmatrix} \boldsymbol{h}^t \\ \boldsymbol{x}^t \end{bmatrix} \right\|_2 \\
&= \left\|\nabla f\left(\boldsymbol{h}^t, \boldsymbol{x}^t\right)\right\|_2, \qquad\qquad (B.10)
\end{aligned}$$

where the first inequality is due to (B.9), 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 $f\left(\boldsymbol{h}^{t+1}, \boldsymbol{x}^{t+1}\right)$ is upper bounded by $f\left(\boldsymbol{h}^{t+1/2}, \boldsymbol{x}^{t+1/2}\right)$, that is,

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

where (i) and (ii) come from (B.1f), and (iii) is due to the elementary inequality $2ab \leq a^2 + b^2$. In order to control $f\left(\boldsymbol{h}^{t+1/2}, \boldsymbol{x}^{t+1/2}\right)$, 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_{\textsf{reg-free}}\left(\boldsymbol{z}^t\right) + \lambda \boldsymbol{h}^t\right), \overline{\boldsymbol{x}}^t - \eta \beta^t \left(\nabla_{\boldsymbol{x}} f_{\textsf{reg-free}}\left(\boldsymbol{z}^t\right) + \lambda \boldsymbol{x}^t\right)\right) \\
&\overset{\text{(ii)}}{=} f\left(\overline{\boldsymbol{h}}^t - \eta \nabla_{\boldsymbol{h}} f\left(\overline{\boldsymbol{z}}^t\right), \overline{\boldsymbol{x}}^t - \eta \nabla_{\boldsymbol{x}} f\left(\overline{\boldsymbol{z}}^t\right)\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\left(\widehat{\boldsymbol{z}}\right) \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 (2.1); (ii) utilizes the relation (B.9); (iii) comes from the mean value theorem [MWCC17, Appendix D.3.1]; (iv) follows from Lemma 6 (which we shall verify shortly); (v) holds true for sufficiently small $\eta>0$; and the last equality follows from the identity (B.10). Therefore, it only remains to verify the conditions required to invoke Lemma 6 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 6.

- We first show that $\overline{\boldsymbol{z}}^t$ satisfies the conditions of Lemma 6. Towards this, it is first seen that

$$
\begin{aligned}
\left\|\overline{\boldsymbol{h}}^t-\boldsymbol{h}^\star\right\|_2^2+\left\|\overline{\boldsymbol{x}}^t-\boldsymbol{x}^\star\right\|_2^2 &= \left\|\frac{\boldsymbol{h}^t}{\overline{\alpha^t/|\alpha^t|}}-\boldsymbol{h}^\star\right\|_2^2+\left\|\frac{\alpha^t}{|\alpha^t|}\boldsymbol{x}^t-\boldsymbol{x}^\star\right\|_2^2\\
&\leq\left(\left\|\frac{\boldsymbol{h}^t}{\overline{\alpha^t/|\alpha^t|}}-\frac{\boldsymbol{h}^t}{\overline{\alpha^t}}\right\|_2+\left\|\frac{\boldsymbol{h}^t}{\overline{\alpha^t}}-\boldsymbol{h}^\star\right\|_2\right)^2+\left(\left\|\frac{\alpha^t}{|\alpha^t|}\boldsymbol{x}^t-\alpha^t\boldsymbol{x}^t\right\|_2+\left\|\alpha^t\boldsymbol{x}^t-\boldsymbol{x}^\star\right\|_2\right)^2\\
&=\left(\left||\alpha^t|-1\right|\left\|\widetilde{\boldsymbol{h}}^t\right\|_2+\left\|\widetilde{\boldsymbol{h}}^t-\boldsymbol{h}^\star\right\|_2\right)^2+\left(\left|\frac{|\alpha^t|-1}{|\alpha^t|}\right|\left\|\widetilde{\boldsymbol{x}}^t\right\|_2+\left\|\widetilde{\boldsymbol{x}}^t-\boldsymbol{x}^\star\right\|_2\right)^2\\
&\lesssim\left(\frac{C_5}{c_\rho}\left(\lambda+\sigma\sqrt{K\log m}\right)\right)^2, \qquad\text{(B.11)}
\end{aligned}
$$

where the last inequality comes from (B.1a) and (B.1h). Further,

$$
\begin{aligned}
\max_{1\leq j\leq m}\left|\boldsymbol{a}_j^{\mathsf{H}}\left(\overline{\boldsymbol{x}}^t-\boldsymbol{x}^\star\right)\right| &\leq \max_{1\leq j\leq m}\left|\boldsymbol{a}_j^{\mathsf{H}}\left(\frac{\alpha^t}{|\alpha^t|}\boldsymbol{x}^t-\alpha^t\boldsymbol{x}^t\right)\right|+\max_{1\leq j\leq m}\left|\boldsymbol{a}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^t-\boldsymbol{x}^\star\right)\right|\\
&\leq\left|\frac{|\alpha^t|-1}{|\alpha^t|}\right|\max_{1\leq j\leq m}\left|\boldsymbol{a}_j^{\mathsf{H}}\widetilde{\boldsymbol{x}}^t\right|+\max_{1\leq j\leq m}\left|\boldsymbol{a}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^t-\boldsymbol{x}^\star\right)\right|\\
&\leq\left|\frac{|\alpha^t|-1}{|\alpha^t|}\right|\left(\max_{1\leq j\leq m}\left|\boldsymbol{a}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^t-\boldsymbol{x}^\star\right)\right|+\max_{1\leq j\leq m}\left|\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{x}^\star\right|\right)+\max_{1\leq j\leq m}\left|\boldsymbol{a}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{x}}^t-\boldsymbol{x}^\star\right)\right|\\
&\lesssim\left(\lambda+\sigma\sqrt{K\log m}\right)\sqrt{\log m}, \qquad\text{(B.12)}
\end{aligned}
$$

where the last inequality follows from (B.1h), (B.2d) and Lemma 18. Similarly, one has

$$
\begin{aligned}
\max_{1\leq j\leq m}\left|\boldsymbol{b}_j^{\mathsf{H}}\overline{\boldsymbol{h}}^t\right| &\leq \max_{1\leq j\leq m}\left|\boldsymbol{b}_j^{\mathsf{H}}\left(\frac{\boldsymbol{h}^t}{\overline{\alpha^t/|\alpha^t|}}-\frac{\boldsymbol{h}^t}{\overline{\alpha^t}}\right)\right|+\max_{1\leq j\leq m}\left|\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right|\\
&\leq\left||\alpha^t|-1\right|\max_{1\leq j\leq m}\left|\boldsymbol{b}_j^{\mathsf{H}}\frac{\boldsymbol{h}^t}{\overline{\alpha^t}}\right|+\max_{1\leq j\leq m}\left|\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right|\\
&\leq 2\max_{1\leq j\leq m}\left|\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right| \qquad\text{(B.13)}\\
&\lesssim\frac{\mu}{\sqrt{m}}\log m+\sigma, \qquad\text{(B.14)}
\end{aligned}
$$

where the last inequality comes from (B.2e). Given that $\overline{\boldsymbol{z}}^t$ satisfies the conditions in Lemma 6, we can invoke Lemma 6 to demonstrate that

$$
\left\|\nabla_{\boldsymbol{h}} f\left(\overline{\boldsymbol{z}}^t\right)-\nabla_{\boldsymbol{h}} f\left(\boldsymbol{z}^\star\right)\right\|_2\leq 4\left\|\overline{\boldsymbol{z}}^t-\boldsymbol{z}^\star\right\|_2. \qquad\text{(B.15)}
$$

- 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 6. To begin with,

$$\left\|\overline{\boldsymbol{z}}^t - \eta \nabla f\left(\overline{\boldsymbol{z}}^t\right) - \boldsymbol{z}^\star\right\|_2 \le \left\|\overline{\boldsymbol{z}}^t - \boldsymbol{z}^\star\right\|_2 + \eta\left\|\nabla f\left(\overline{\boldsymbol{z}}^t\right) - \nabla f\left(\boldsymbol{z}^\star\right)\right\|_2 + \eta\left\|\nabla f\left(\boldsymbol{z}^\star\right)\right\|_2. \tag{B.16}$$

We observe that

$$\begin{aligned}\left\|\nabla f\left(\boldsymbol{z}^\star\right)\right\|_2 &\le \left\|\nabla f_{\mathsf{clean}}\left(\boldsymbol{z}^\star\right)\right\|_2 + \left\|\mathcal{A}^*\left(\boldsymbol{\xi}\right)\boldsymbol{h}^\star\right\|_2 + \left\|\mathcal{A}^*\left(\boldsymbol{\xi}\right)\boldsymbol{x}^\star\right\|_2 + \lambda\left\|\boldsymbol{h}^\star\right\|_2 + \lambda\left\|\boldsymbol{z}^\star\right\|_2 \\ &\lesssim \lambda + \sigma\sqrt{K\log m}. \end{aligned}\tag{B.17}$$

Taking (B.17), (B.15), (B.11) and (B.16) 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

$$\begin{aligned}&\max_{1\le j\le m}\left|\boldsymbol{a}_j^{\mathsf{H}}\left(\overline{\boldsymbol{x}}^t - \eta\nabla_{\boldsymbol{x}} f\left(\overline{\boldsymbol{z}}^t\right) - \boldsymbol{x}^\star\right)\right| \\ &\le \max_{1\le j\le m}\left|\boldsymbol{a}_j^{\mathsf{H}}\left(\overline{\boldsymbol{x}}^t - \boldsymbol{x}^\star\right)\right| + \eta\max_{1\le j\le m}\left|\boldsymbol{a}_j^{\mathsf{H}}\nabla_{\boldsymbol{x}} f\left(\overline{\boldsymbol{z}}^t\right)\right| \\ &\le \max_{1\le j\le m}\left|\boldsymbol{a}_j^{\mathsf{H}}\left(\overline{\boldsymbol{x}}^t - \boldsymbol{x}^\star\right)\right| + \eta\left(\max_{1\le j\le m}\left|\boldsymbol{a}_j^{\mathsf{H}}\nabla_{\boldsymbol{x}} f\left(\overline{\boldsymbol{z}}^t - \widetilde{\boldsymbol{z}}^{t,(l)}\right)\right| + \max_{1\le j\le m}\left|\boldsymbol{a}_j^{\mathsf{H}}\nabla_{\boldsymbol{x}} f\left(\widetilde{\boldsymbol{z}}^{t,(l)}\right)\right|\right) \\ &\le C\sqrt{\log m}\left(\lambda + \sigma\sqrt{K\log m}\right) + 4\eta\left(10\sqrt{K}\times 4\max_{1\le j\le m}\left\|\widetilde{\boldsymbol{z}}^t - \widetilde{\boldsymbol{z}}^{t,(l)}\right\|_2 + 20\sqrt{\log m}\max_{1\le j\le m}\left\|\nabla_{\boldsymbol{x}} f\left(\widetilde{\boldsymbol{z}}^{t,(l)}\right)\right\|_2\right),\end{aligned}\tag{B.18}$$

where the last inequality follows from (B.12) for some constant $C > 0$, (B.15) and Lemma 18. Further, it is self-evident that $\widetilde{\boldsymbol{z}}^{t,(l)}$ satisfies the conditions of Lemma 6, so that we have

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

where the second inequality invokes Lemma 6 and (B.17). This together with (B.18) and (B.2) gives

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

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

$$\begin{aligned}&\max_{1\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| \\ &\quad\le \max_{1\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\left(\overline{\boldsymbol{h}}^t - \boldsymbol{h}^\star\right)\right| + \eta\max_{1\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\nabla_{\boldsymbol{h}} f\left(\overline{\boldsymbol{z}}^t\right)\right| \\ &\quad\le \max_{1\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\left(\overline{\boldsymbol{h}}^t - \boldsymbol{h}^\star\right)\right| + \eta\max_{1\le j\le 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\le \max_{1\le j\le 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\le j\le 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\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right| + \left|\alpha^t\right|^{-1}\underbrace{\max_{1\le j\le 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\le \left|\left|\alpha^t\right| - 1\right|\max_{1\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right| + \max_{1\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\left(\widetilde{\boldsymbol{h}}^t - \boldsymbol{h}^\star\right)\right| + \eta\left(2\lambda\max_{1\le j\le m}\left|\boldsymbol{b}_j^{\mathsf{H}}\widetilde{\boldsymbol{h}}^t\right| + 2\tau\right).\end{aligned}\tag{B.19}$$

Here, the last inequality utilizes the fact $\|\boldsymbol{b}_j\|_2 = \sqrt{K/m}$ and (B.1h). 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.43), and the last inequality is a direct consequence of Appendix A.9. Finally, continue the bound (B.19) 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.1h), (B.2e) and (B.2f).

B.3 Proof of Lemma 2

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

$$\boldsymbol{Z} := \boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}, \tag{B.20}$$

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

$$T := \left\{\boldsymbol{X} : \boldsymbol{X} = \boldsymbol{h}\boldsymbol{v}^{\mathsf{H}} + \boldsymbol{u}\boldsymbol{x}^{\mathsf{H}}, \boldsymbol{v}\in\mathbb{C}^K, \boldsymbol{u}\in\mathbb{C}^K\right\}. \tag{B.21}$$

Further, define two associated projection operators as follows

$$\mathcal{P}_T\left(\boldsymbol{X}\right) := \frac{1}{\|\boldsymbol{h}\|_2^2}\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\boldsymbol{X} + \frac{1}{\|\boldsymbol{x}\|_2^2}\boldsymbol{X}\boldsymbol{x}\boldsymbol{x}^{\mathsf{H}} - \frac{1}{\|\boldsymbol{h}\|_2^2\|\boldsymbol{x}\|_2^2}\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\boldsymbol{X}\boldsymbol{x}\boldsymbol{x}^{\mathsf{H}}, \tag{B.22a}$$

$$\mathcal{P}_{T^\perp}\left(\boldsymbol{X}\right) := \left(\boldsymbol{I} - \frac{1}{\|\boldsymbol{h}\|_2^2}\boldsymbol{h}\boldsymbol{h}^{\mathsf{H}}\right)\boldsymbol{X}\left(\boldsymbol{I} - \frac{1}{\|\boldsymbol{x}\|_2^2}\boldsymbol{x}\boldsymbol{x}^{\mathsf{H}}\right). \tag{B.22b}$$

We further introduce three key lemmas below. Basically, Lemma (15) 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.

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

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

holds simultaneously for all T for which the associated point $(\boldsymbol{h},\boldsymbol{x})$ obeys (2.5b) and (2.5c). Here, T denotes the tangent space of $\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}$.

Proof. See Appendix B.3.1. □

Recall the definition of operator $\mathcal{T}$ in (A.8). The second lemma 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 $\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}} - \boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}$.

Lemma 16. *Suppose that the sample complexity satisfies $m \ge C\mu^2 K\log^4 m$ 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 (2.5b) and (2.5c).

Proof. See Appendix B.3.3. $\square$

The next lemma proves useful in connecting the first order optimality conditions of convex and nonconvex formulation.

Lemma 17. *Under the assumptions of Lemma 2, one has*

$$\mathcal{T}\left(\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}-\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\right)-\mathcal{A}^{*}\left(\boldsymbol{\xi}\right)=-\frac{\lambda}{\left\|\boldsymbol{h}\right\|_{2}\left\|\boldsymbol{x}\right\|_{2}}\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}+\boldsymbol{R},$$

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

$$\left\|\mathcal{P}_{T}\left(\boldsymbol{R}\right)\right\|_{\mathrm{F}}\leq2\left\|\nabla f\left(\boldsymbol{h},\boldsymbol{x}\right)\right\|_{2}\qquad\textit{and}\qquad\left\|\mathcal{P}_{T^{\perp}}\left(\boldsymbol{R}\right)\right\|\leq\lambda/2.$$

Proof. See Appendix B.3.4. $\square$

With these supporting lemmas in hand, we are ready to prove Lemma 2. Suppose $\boldsymbol{Z}_{\mathsf{cvx}}$ is the minimizer of (1.3).

1. Let $\boldsymbol{\Delta}\coloneqq\boldsymbol{Z}_{\mathsf{cvx}}-\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}$. 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 $\|\boldsymbol{W}\|\leq1$, 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 $\boldsymbol{p}\coloneqq\boldsymbol{h}/\left\|\boldsymbol{h}\right\|_{2}$ and $\boldsymbol{q}\coloneqq\boldsymbol{x}/\left\|\boldsymbol{x}\right\|_{2}$. 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

$$\begin{aligned}0\leq\left\|\mathcal{A}\left(\boldsymbol{\Delta}\right)\right\|_{2}^{2}&\leq-\left\langle\mathcal{T}\left(\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}-\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\right)-\mathcal{A}^{*}\left(\boldsymbol{\xi}\right),\boldsymbol{\Delta}\right\rangle-2\lambda\left\langle\boldsymbol{p}\boldsymbol{q}^{\mathsf{H}}+\boldsymbol{W},\boldsymbol{\Delta}\right\rangle\\&=-\left\langle\mathcal{T}\left(\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}-\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}\right)-\mathcal{A}^{*}\left(\boldsymbol{\xi}\right),\boldsymbol{\Delta}\right\rangle-2\lambda\left\langle\boldsymbol{p}\boldsymbol{q}^{\mathsf{H}},\boldsymbol{\Delta}\right\rangle-2\lambda\left\|\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{\Delta}\right)\right\|_{*}\\&\overset{\text{(i)}}{=}-\left\langle\boldsymbol{R},\boldsymbol{\Delta}\right\rangle-2\lambda\left\|\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{\Delta}\right)\right\|_{*}\\&=-\left\langle\mathcal{P}_{\boldsymbol{T}}\left(\boldsymbol{R}\right),\boldsymbol{\Delta}\right\rangle-\left\langle\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{R}\right),\boldsymbol{\Delta}\right\rangle-2\lambda\left\|\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{\Delta}\right)\right\|_{*},\end{aligned}\tag{B.23}$$

where $\boldsymbol{R}$ in (i) is defined in Lemma 17. Hence,

$$\begin{aligned}&-\left\|\mathcal{P}_{\boldsymbol{T}}\left(\boldsymbol{R}\right)\right\|_{\mathrm{F}}\left\|\mathcal{P}_{\boldsymbol{T}}\left(\boldsymbol{\Delta}\right)\right\|_{\mathrm{F}}-\left\|\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{R}\right)\right\|\left\|\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{\Delta}\right)\right\|_{*}+2\lambda\left\|\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{\Delta}\right)\right\|_{*}\\&\leq\left\langle\mathcal{P}_{\boldsymbol{T}}\left(\boldsymbol{R}\right),\boldsymbol{\Delta}\right\rangle+\left\langle\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{R}\right),\boldsymbol{\Delta}\right\rangle+2\lambda\left\|\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{\Delta}\right)\right\|_{*}\leq0.\end{aligned}$$

Lemma 17 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 17. We then arrive at

$$\left\|\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{\Delta}\right)\right\|_{\mathrm{F}}\leq\left\|\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{\Delta}\right)\right\|_{*}\leq C\frac{4}{3m^{10}}\left\|\mathcal{P}_{\boldsymbol{T}}\left(\boldsymbol{\Delta}\right)\right\|_{\mathrm{F}}\leq\left\|\mathcal{P}_{\boldsymbol{T}}\left(\boldsymbol{\Delta}\right)\right\|_{\mathrm{F}}.\tag{B.24}$$

2. Next, we return to (B.23) to deduce that

$$
\begin{aligned}
\left\|\mathcal{A}\left(\boldsymbol{\Delta}\right)\right\|_2^2 &\leq -\left\langle \mathcal{P}_{\boldsymbol{T}}\left(\boldsymbol{R}\right), \boldsymbol{\Delta}\right\rangle - \left\langle \mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{R}\right), \boldsymbol{\Delta}\right\rangle - 2\lambda \left\|\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{\Delta}\right)\right\|_* \\
&\leq \left\|\mathcal{P}_{\boldsymbol{T}}\left(\boldsymbol{R}\right)\right\|_{\mathrm{F}} \left\|\mathcal{P}_{\boldsymbol{T}}\left(\boldsymbol{\Delta}\right)\right\|_{\mathrm{F}} + \left\|\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{R}\right)\right\| \left\|\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{\Delta}\right)\right\|_* - 2\lambda \left\|\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{\Delta}\right)\right\|_* &\text{(B.25)} \\
&\overset{\text{(i)}}{\leq} \left\|\mathcal{P}_{\boldsymbol{T}}\left(\boldsymbol{R}\right)\right\|_{\mathrm{F}} \left\|\mathcal{P}_{\boldsymbol{T}}\left(\boldsymbol{\Delta}\right)\right\|_{\mathrm{F}} - \frac{3\lambda}{2} \left\|\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{\Delta}\right)\right\|_* \\
&\leq \left\|\mathcal{P}_{\boldsymbol{T}}\left(\boldsymbol{R}\right)\right\|_{\mathrm{F}} \left\|\mathcal{P}_{\boldsymbol{T}}\left(\boldsymbol{\Delta}\right)\right\|_{\mathrm{F}} &\text{(B.26)} \\
&\overset{\text{(ii)}}{\leq} 2\left\|\nabla f\left(\boldsymbol{h}, \boldsymbol{x}\right)\right\|_2 \left\|\boldsymbol{\Delta}\right\|_{\mathrm{F}}, &\text{(B.27)}
\end{aligned}
$$

where (i) and (ii) come from Lemma 17.

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

$$
\begin{aligned}
\left\|\mathcal{A}\left(\boldsymbol{\Delta}\right)\right\|_2 &= \left\|\mathcal{A}\left(\mathcal{P}_{\boldsymbol{T}}\left(\boldsymbol{\Delta}\right)\right) + \mathcal{A}\left(\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{\Delta}\right)\right)\right\|_2 \\
&\geq \left\|\mathcal{A}\left(\mathcal{P}_{\boldsymbol{T}}\left(\boldsymbol{\Delta}\right)\right)\right\|_2 - \left\|\mathcal{A}\left(\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{\Delta}\right)\right)\right\|_2 \\
&\geq \left\|\mathcal{P}_{\boldsymbol{T}}\left(\boldsymbol{\Delta}\right)\right\|_{\mathrm{F}}/4 - \sqrt{2K\log K + \gamma \log m}\left\|\mathcal{P}_{\boldsymbol{T}^{\perp}}\left(\boldsymbol{\Delta}\right)\right\|_{\mathrm{F}}, &\text{(B.28)}
\end{aligned}
$$

where the last inequality comes from Lemma 15 and Lemma 3. Since (B.24) gives

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

as long as $m \gg K$, (B.28) yields

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

In addition, (B.24) implies

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

Consequently,

$$
\left\|\mathcal{A}\left(\boldsymbol{\Delta}\right)\right\|_2 \geq \frac{1}{8}\left\|\mathcal{P}_{\boldsymbol{T}}\left(\boldsymbol{\Delta}\right)\right\|_{\mathrm{F}} \geq \frac{1}{16}\left\|\boldsymbol{\Delta}\right\|_{\mathrm{F}}. \tag{B.29}
$$

Combining (B.26) and (B.29), we have

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

and therefore

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

B.3.1 Proof of Lemma 15

By the definition of T (cf. (B.21)), 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

$$
(\boldsymbol{h}, \boldsymbol{x}) := \arg\min_{(\widetilde{\boldsymbol{h}}, \widetilde{\boldsymbol{x}})} \left\{\frac{1}{2}\|\widetilde{\boldsymbol{h}}\|_2^2 + \frac{1}{2}\left\|\widetilde{\boldsymbol{x}}\right\|_2^2 \mid \boldsymbol{Z} = \widetilde{\boldsymbol{h}}\boldsymbol{u}^{\mathsf{H}} + \boldsymbol{v}\widetilde{\boldsymbol{x}}^{\mathsf{H}} \text{ for some } \boldsymbol{u} \text{ and } \boldsymbol{v}\right\}. \tag{B.30}
$$

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

$$\|\boldsymbol{Z}\|_{\mathrm{F}}^2 \leq 8\left(\|\boldsymbol{u}\|_2^2+\|\boldsymbol{v}\|_2^2\right), \tag{B.31}$$

and (2) demonstrating that

$$\|\mathcal{A}(\boldsymbol{Z})\|_2^2 \geq \frac{1}{2}\left(\|\boldsymbol{u}\|_2^2+\|\boldsymbol{v}\|_2^2\right). \tag{B.32}$$

The first claim (B.31) can be justified in the same way as [CCF+19, Equation (81)]; we thus omit this part here for brevity.

It then boils down to justifying the second claim (B.32), towards which we first decompose

$$\|\mathcal{A}(\boldsymbol{Z})\|_2^2 = \underbrace{\|\mathcal{A}(\boldsymbol{Z})\|_2^2 - \|\boldsymbol{Z}\|_2^2}_{=:\alpha_1} + \underbrace{\|\boldsymbol{Z}\|_2^2}_{=:\alpha_2}. \tag{B.33}$$

By repeating the same argument as in [CCF+19, Appendix C.3.1, 2(a)], we can lower bound α_2 by

$$\alpha_2 \geq \left\|\boldsymbol{h}^\star\boldsymbol{u}^{\mathsf{H}}\right\|_{\mathrm{F}}^2 + \left\|\boldsymbol{v}\boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}}^2 - \frac{1}{50}\left(\|\boldsymbol{u}\|_2^2+\|\boldsymbol{v}\|_2^2\right).$$

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

$$\begin{aligned}
\boldsymbol{h}\boldsymbol{u}^{\mathsf{H}} + \boldsymbol{v}\boldsymbol{x}^{\mathsf{H}} &= \left(\boldsymbol{h}^\star + \boldsymbol{\Delta}_{\boldsymbol{h}}\right)\boldsymbol{u}^{\mathsf{H}} + \boldsymbol{v}\left(\boldsymbol{x}^\star + \boldsymbol{\Delta}_{\boldsymbol{x}}\right)^{\mathsf{H}} \\
&= \boldsymbol{h}^\star\boldsymbol{u}^{\mathsf{H}} + \boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{u}^{\mathsf{H}} + \boldsymbol{v}\boldsymbol{x}^{\star\mathsf{H}} + \boldsymbol{v}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}.
\end{aligned}$$

This implies that α_1 can be expanded as follows

$$\begin{aligned}
\alpha_1 = &\underbrace{\left\|\mathcal{A}\left(\boldsymbol{h}^\star\boldsymbol{u}^{\mathsf{H}} + \boldsymbol{v}\boldsymbol{x}^{\star\mathsf{H}}\right)\right\|_2^2 - \left\|\boldsymbol{h}^\star\boldsymbol{u}^{\mathsf{H}} + \boldsymbol{v}\boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}}^2}_{=:\gamma_1} + \underbrace{\left\|\mathcal{A}\left(\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{u}^{\mathsf{H}} + \boldsymbol{v}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\right)\right\|_2^2 - \left\|\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{u}^{\mathsf{H}} + \boldsymbol{v}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\right\|_{\mathrm{F}}^2}_{=:\gamma_2} \\
&+ 2\underbrace{\left\langle\mathcal{A}\left(\boldsymbol{h}^\star\boldsymbol{u}^{\mathsf{H}} + \boldsymbol{v}\boldsymbol{x}^{\star\mathsf{H}}\right), \mathcal{A}\left(\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{u}^{\mathsf{H}} + \boldsymbol{v}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\right)\right\rangle - \left\langle\boldsymbol{h}^\star\boldsymbol{u}^{\mathsf{H}} + \boldsymbol{v}\boldsymbol{x}^{\star\mathsf{H}}, \boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{u}^{\mathsf{H}} + \boldsymbol{v}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\right\rangle}_{=:\gamma_3},
\end{aligned}$$

thereby motivating us to cope with these terms separately.

- Regarding γ_1, it is easily seen that

$$|\gamma_1| \leq \left\|\mathcal{P}_T\mathcal{A}^*\mathcal{A}\mathcal{P}_T - \mathcal{P}_T\right\| \cdot \left\|\boldsymbol{h}^\star\boldsymbol{u}^{\mathsf{H}} + \boldsymbol{v}\boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}}^2 \leq \frac{1}{100}\left(\|\boldsymbol{u}\|^2+\|\boldsymbol{v}\|^2\right),$$

  where the last inequality is obtained by invoking [LLSW19, Lemma 5.12].

- When it comes to γ_2, we observe that

$$\gamma_2 \geq -\left\|\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{u}^{\mathsf{H}} + \boldsymbol{v}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\right\|_{\mathrm{F}}^2 \geq -\frac{1}{100}\left(\|\boldsymbol{u}\|_2^2+\|\boldsymbol{v}\|_2^2\right)$$

  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

$$\begin{aligned}
&\left|\left\langle\mathcal{A}\left(\boldsymbol{v}\boldsymbol{x}^{\star\mathsf{H}}\right), \mathcal{A}\left(\boldsymbol{v}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\right)\right\rangle - \left\langle\boldsymbol{v}\boldsymbol{x}^{\star\mathsf{H}}, \boldsymbol{v}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\right\rangle\right| \\
&\quad\leq \left|\left\langle\mathcal{A}\left(\boldsymbol{v}\boldsymbol{x}^{\star\mathsf{H}}\right), \mathcal{A}\left(\boldsymbol{v}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\right)\right\rangle\right| + \left|\left\langle\boldsymbol{v}\boldsymbol{x}^{\star\mathsf{H}}, \boldsymbol{v}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\right\rangle\right| \\
&\quad\overset{\text{(i)}}{\leq} \left\|\mathcal{A}\left(\boldsymbol{v}\boldsymbol{x}^{\star\mathsf{H}}\right)\right\|_2 \left\|\mathcal{A}\left(\boldsymbol{v}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\right)\right\|_2 + \|\boldsymbol{x}^\star\|_2 \left\|\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\right\|_2 \|\boldsymbol{v}\|_2^2
\end{aligned}$$

$$
\begin{aligned}
&\overset{\text{(ii)}}{\leq} \sqrt{\sum_{j=1}^{m}\left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{v}\right|^2\left|\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\right|^2}\sqrt{\sum_{j=1}^{m}\left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{v}\right|^2\left|\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j\right|^2}+\frac{1}{200}\left\|\boldsymbol{v}\right\|_2^2 \\
&\overset{\text{(iii)}}{\leq} \sqrt{\left\|\boldsymbol{v}\right\|_2^2\max_{1\leq j\leq m}\left|\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\right|^2}\cdot\sqrt{\left\|\boldsymbol{v}\right\|_2^2\max_{1\leq j\leq m}\left|\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j\right|^2}+\frac{1}{200}\left\|\boldsymbol{v}\right\|_2^2 \\
&\overset{\text{(iv)}}{\leq} 20\sqrt{\log m}\cdot C\sqrt{\log m}\left(\lambda+\sigma\sqrt{K\log m}\right)\left\|\boldsymbol{v}\right\|_2^2+\frac{1}{200}\left\|\boldsymbol{v}\right\|_2^2 \\
&\leq \frac{1}{100}\left\|\boldsymbol{v}\right\|_2^2,
\end{aligned}
\tag{B.34}
$$

where the (i) and (ii) follow from the Cauchy-Schwarz inequality and (2.5b) that $\left\|\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\right\|_2 \lesssim \lambda + \sigma\sqrt{K\log m} \leq 1/200$; (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 18 and (2.5c); and the last inequality holds true as long as $\sigma\sqrt{K\log^3 m} \ll 1$.

2. Similarly, we can demonstrate that

$$
\begin{aligned}
&\left|\left\langle\mathcal{A}\left(\boldsymbol{h}^{\star}\boldsymbol{u}^{\mathsf{H}}\right),\mathcal{A}\left(\boldsymbol{v}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\right)\right\rangle-\left\langle\boldsymbol{h}^{\star}\boldsymbol{u}^{\mathsf{H}},\boldsymbol{v}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\right\rangle\right| \\
&\quad\overset{\text{(i)}}{\leq} \sqrt{\sum_{j=1}^{m}\left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^{\star}\right|^2\left|\boldsymbol{u}^{\mathsf{H}}\boldsymbol{a}_j\right|^2}\sqrt{\sum_{j=1}^{m}\left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{v}\right|^2\left|\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j\right|^2}+\frac{1}{200}\left\|\boldsymbol{u}\right\|_2\left\|\boldsymbol{v}\right\|_2 \\
&\quad\overset{\text{(ii)}}{\leq} \sqrt{\sum_{j=1}^{m}\left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^{\star}\right|^2\left|\boldsymbol{u}^{\mathsf{H}}\boldsymbol{a}_j\right|^2}\cdot C\sqrt{\log m}\left(\lambda+\sigma\sqrt{K\log m}\right)\left\|\boldsymbol{v}\right\|_2+\frac{1}{200}\left\|\boldsymbol{u}\right\|_2\left\|\boldsymbol{v}\right\|_2 \\
&\quad\leq \frac{1}{100}\left\|\boldsymbol{u}\right\|_2\left\|\boldsymbol{v}\right\|_2,
\end{aligned}
$$

where (i) holds for the same reason as Step (ii) in (B.34); (ii) arises due to the identity $\sum_{j=1}^m \left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{v}\right|^2 = \|\boldsymbol{v}\|_2^2$ and (2.5c); and the last inequality relies on the following claim.

Claim 4. With probability exceeding $1 - O\left(m^{-100}\right)$, the following inequality

$$
\left|\sum_{j=1}^{m}\left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^{\star}\right|^2\left|\boldsymbol{u}^{\mathsf{H}}\boldsymbol{a}_j\right|^2-\left\|\boldsymbol{u}\right\|_2^2\right| \lesssim \sqrt{\frac{\mu^2 K\log m}{m}}\left\|\boldsymbol{u}\right\|_2^2 \tag{B.35}
$$

holds uniformly for any $\boldsymbol{u}$.

Proof. See Appendix B.3.2. □

3. The next term we shall control is

$$
\left\langle\mathcal{A}\left(\boldsymbol{h}^{\star}\boldsymbol{u}^{\mathsf{H}}\right),\mathcal{A}\left(\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{u}^{\mathsf{H}}\right)\right\rangle-\left\langle\boldsymbol{h}^{\star}\boldsymbol{u}^{\mathsf{H}},\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{u}^{\mathsf{H}}\right\rangle=\sum_{j=1}^{m}\left(\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^{\star}\right)\left(\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\right)\left(\left|\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{u}\right|^2-\left\|\boldsymbol{u}\right\|_2^2\right).
$$

By virtue of the Bernstein inequality [Ver18, Theorem 2.8.2], we have

$$
\begin{aligned}
&\mathbb{P}\left(\left|\left\langle\mathcal{A}\left(\boldsymbol{h}^{\star}\boldsymbol{u}^{\mathsf{H}}\right),\mathcal{A}\left(\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{u}^{\mathsf{H}}\right)\right\rangle-\left\langle\boldsymbol{h}^{\star}\boldsymbol{u}^{\mathsf{H}},\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{u}^{\mathsf{H}}\right\rangle\right|\geq\tau\left\|\boldsymbol{u}\right\|_2^2\right) \\
&\quad\leq 2\max\left\{\exp\left(-\frac{\tau^2}{4\left\|\boldsymbol{B}\boldsymbol{\Delta}_{\boldsymbol{h}}\right\|_\infty^2}\right),\exp\left(-\frac{\tau}{4\left\|\boldsymbol{B}\boldsymbol{\Delta}_{\boldsymbol{h}}\right\|_\infty\left\|\boldsymbol{B}\boldsymbol{h}^{\star}\right\|_\infty}\right)\right\}
\end{aligned}
$$

for any $\tau \geq 0$. Let us choose τ to be

$$
\tau = 2\left\|\boldsymbol{B}\boldsymbol{\Delta}_{\boldsymbol{h}}\right\|_\infty\sqrt{2K\log m}+8\left\|\boldsymbol{B}\boldsymbol{\Delta}_{\boldsymbol{h}}\right\|_\infty\left\|\boldsymbol{B}\boldsymbol{h}^{\star}\right\|_\infty K\log m.
$$

In view of (2.5c) and (1.7), 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

$$\mathbb{P}\left(\left|\left\langle \mathcal{A}\left(\boldsymbol{h}^{\star}\boldsymbol{u}^{\mathsf{H}}\right), \mathcal{A}\left(\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{u}^{\mathsf{H}}\right)\right\rangle - \left\langle \boldsymbol{h}^{\star}\boldsymbol{u}^{\mathsf{H}}, \boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{u}^{\mathsf{H}}\right\rangle\right| \geq \frac{1}{100}\left\|\boldsymbol{u}\right\|_2^2\right) \leq 2\exp\left(-2K\log m\right). \tag{B.36}$$

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 [Ver18, 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

$$\left|\left\langle \mathcal{A}\left(\boldsymbol{h}^{\star}\boldsymbol{u}^{\mathsf{H}}\right), \mathcal{A}\left(\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{u}^{\mathsf{H}}\right)\right\rangle - \left\langle \boldsymbol{h}^{\star}\boldsymbol{u}^{\mathsf{H}}, \boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{u}^{\mathsf{H}}\right\rangle\right| \geq \frac{1}{100}\left\|\boldsymbol{u}\right\|_2^2$$

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

$$F\left(\boldsymbol{h}, \boldsymbol{u}\right) := \left\langle \mathcal{A}\left(\boldsymbol{h}^{\star}\boldsymbol{u}^{\mathsf{H}}\right), \mathcal{A}\left(\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{u}^{\mathsf{H}}\right)\right\rangle - \left\langle \boldsymbol{h}^{\star}\boldsymbol{u}^{\mathsf{H}}, \boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{u}^{\mathsf{H}}\right\rangle,$$

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| \\
&\quad \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| \\
&\qquad + \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| \\
&\qquad + \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| \\
&\quad \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 \\
&\qquad + \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 \\
&\quad \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) \\
&\quad \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 3. 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 $\left\langle \mathcal{A}\left(\boldsymbol{v}\boldsymbol{x}^{\star\mathsf{H}}\right), \mathcal{A}\left(\boldsymbol{\Delta_h}\boldsymbol{u}^{\mathsf{H}}\right)\right\rangle - \left\langle \boldsymbol{v}\boldsymbol{x}^{\star\mathsf{H}}, \boldsymbol{\Delta_h}\boldsymbol{u}^{\mathsf{H}}\right\rangle$ 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

$$|\gamma_3| \le \frac{1}{100}\|\boldsymbol{v}\|_2^2 + \frac{1}{100}\|\boldsymbol{u}\|_2\|\boldsymbol{v}\|_2 + \frac{1}{50}\|\boldsymbol{u}\|_2^2 + \frac{1}{100}\|\boldsymbol{u}\|_2\|\boldsymbol{v}\|_2 \le \frac{1}{25}\left(\|\boldsymbol{u}\|_2^2 + \|\boldsymbol{v}\|_2^2\right).$$

Combining the above results, we can continue the relation (B.33) to conclude that

$$\begin{aligned}
\|\mathcal{A}(\boldsymbol{Z})\|_2^2 &= \alpha_2 + \alpha_1 \\
&\ge \left\|\boldsymbol{h}^\star\boldsymbol{u}^{\mathsf{H}}\right\|_{\mathrm{F}}^2 + \left\|\boldsymbol{v}\boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}}^2 - \frac{1}{50}\left(\|\boldsymbol{u}\|_2^2 + \|\boldsymbol{v}\|_2^2\right) - |\gamma_1| - \frac{1}{100}\left(\|\boldsymbol{u}\|_2^2 + \|\boldsymbol{v}\|_2^2\right) - \frac{1}{25}\left(\|\boldsymbol{u}\|_2^2 + \|\boldsymbol{v}\|_2^2\right) \\
&\ge \frac{1}{2}\left(\|\boldsymbol{u}\|_2^2 + \|\boldsymbol{v}\|_2^2\right)
\end{aligned}$$

as claimed.

B.3.2 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 $\mathbb{E}\left[\left|\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{u}\right|^2 - \|\boldsymbol{u}\|_2^2\right] = 0$. In view of the Bernstein inequality (cf. [Ver18, Theorem 2.8.2]), we have

$$\begin{aligned}
&\mathbb{P}\left(\left|\sum_{j=1}^m \left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\right|^2 \left(\left|\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{u}\right|^2 - \|\boldsymbol{u}\|_2^2\right)\right| \ge \tau\|\boldsymbol{u}\|_2^2\right) \\
&\qquad \le 2\max\left\{\exp\left(-\frac{\tau^2}{4\|\boldsymbol{B}\boldsymbol{h}^\star\|_\infty^2\|\boldsymbol{u}\|_2^2}\right), \exp\left(-\frac{\tau}{4\|\boldsymbol{B}\boldsymbol{\Delta_h}\|_\infty^2\|\boldsymbol{u}\|_2}\right)\right\}
\end{aligned}$$

for any $\tau \ge 0$. 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

$$\mathbb{P}\left(\left|\sum_{j=1}^m \left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\right|^2 \left(\left|\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{u}\right|^2 - \|\boldsymbol{u}\|_2^2\right)\right| \ge \tau\|\boldsymbol{u}\|_2^2\right) \le 2\exp\left(-4K\log m\right). \tag{B.37}$$

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 [Ver18, Corollary 4.2.13]

$$|\mathcal{N}_0| \le \left(1 + \frac{2}{\epsilon_0}\right)^{2K}.$$

By taking the union bound over $\mathcal{N}_0$, we reach

$$\left|\sum_{j=1}^m \left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star\right|^2 \left(\left|\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{u}\right|^2 - \|\boldsymbol{u}\|_2^2\right)\right| \ge 4\|\boldsymbol{B}\boldsymbol{h}^\star\|_\infty\sqrt{2K\log m} + 16\|\boldsymbol{B}\boldsymbol{h}^\star\|_\infty^2 K\log m, \qquad \forall \boldsymbol{u} \in \mathcal{N}_0$$

with probability at least

$$1 - \left(1 + \frac{2}{\epsilon_0}\right)^{2K} e^{-4K\log m} \ge 1 - O\left(m^{-10}\right).$$

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) \\
&\quad \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 \\
&\quad \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 \\
&\quad \overset{\text{(iv)}}{\leq} \left(4K \log K + 20 \log m + 1\right) \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

$$\|\mathcal{A}\| \leq \sqrt{2K \log K + 10 \log m}, \tag{B.38}$$

which holds with probability at least $1 - O\left(m^{-10}\right)$ according to Lemma 3. Letting $\epsilon_0 = r / \left(m \log m\right)$ with $r = \lambda + \sigma \sqrt{K \log m}$, we note it satisfies

$$1 - \left(1 + \frac{2}{\epsilon_0}\right)^{2K} e^{-4K \log m} \geq 1 - O\left(m^{-10}\right).$$

Therefore, we conclude that: with probability at least $1 - O\left(m^{-10}\right)$, one has

$$\begin{aligned}
|\eta| &\leq 4 \|\boldsymbol{B} \boldsymbol{h}^\star\|_\infty \sqrt{2K \log m} + 16 \|\boldsymbol{B} \boldsymbol{h}^\star\|_\infty^2 K \log m + \left(4K \log K + 20 \log m + 1\right) \epsilon_0 \\
&\lesssim \sqrt{\frac{\mu^2 K \log m}{m}}
\end{aligned}$$

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 (1.7).

B.3.3 Proof of Lemma 16

Recall the definition of $\mathcal{T}^{\mathsf{debias}}$ in (A.8), obtained by subtracting the expectation from $\mathcal{T}$. For any fixed vectors $\boldsymbol{h}$ and $\boldsymbol{x}$, we make note of the following decomposition

$$\begin{aligned}
\boldsymbol{h} \boldsymbol{x}^{\mathsf{H}} - \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}} &= \left(\boldsymbol{\Delta}_{\boldsymbol{h}} + \boldsymbol{h}^\star\right) \left(\boldsymbol{\Delta}_{\boldsymbol{x}} + \boldsymbol{x}^\star\right)^{\mathsf{H}} - \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}} \\
&= \boldsymbol{h}^\star \boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}} + \boldsymbol{\Delta}_{\boldsymbol{h}} \boldsymbol{x}^{\star\mathsf{H}} + \boldsymbol{\Delta}_{\boldsymbol{h}} \boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}},
\end{aligned}$$

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

$$\mathcal{T}^{\text{debias}}\left(\boldsymbol{h}^\star \boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\right) = \sum_{j=1}^{m} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}} \left(\boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} - \boldsymbol{I}_K\right).$$

Consequently, for any fixed unit vectors $\boldsymbol{u}$, $\boldsymbol{v} \in \mathbb{C}^K$ one has

$$\boldsymbol{u}^{\mathsf{H}} \mathcal{T}^{\text{debias}}\left(\boldsymbol{h}^\star \boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\right) \boldsymbol{v} = \sum_{j=1}^{m} \left(\boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{v} - \boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}} \boldsymbol{v}\right),$$

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| \le 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| \le 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| \le 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

$$\begin{aligned}
|\omega_2| &= \left| \sum_{j=1}^{m} \mathbb{E}\left[\boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{v}\, \mathbb{1}_{\left\{|\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}} \boldsymbol{a}_j| > C_4 r \sqrt{\log m}\right\}} \right] \right| \\
&\overset{\text{(i)}}{\le} \sum_{j=1}^{m} \sqrt{ \mathbb{E}\left[\left| \boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{v} \right|^2 \right] \mathbb{P}\left[|\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}} \boldsymbol{a}_j| > C_4 r \sqrt{\log m} \right] } \\
&\overset{\text{(ii)}}{\le} \sum_{j=1}^{m} \left| \boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \right| \sqrt{ \left(2 \left| \boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}} \boldsymbol{v} \right|^2 + \|\boldsymbol{\Delta}_{\boldsymbol{x}}\|_2^2 \|\boldsymbol{v}\|_2^2 \right) 2 \exp\left(- \frac{C_4^2 r^2 \log m}{2 \|\boldsymbol{\Delta}_{\boldsymbol{x}}\|_2^2} \right) } \\
&\le \sum_{j=1}^{m} \left| \boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \right| \sqrt{ 6 \|\boldsymbol{\Delta}_{\boldsymbol{x}}\|_2^2 \exp\left(-50 \log m\right) } \\
&\overset{\text{(iii)}}{\le} \sum_{j=1}^{m} \left(\left| \boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \right|^2 + \left| \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \right|^2 \right) \frac{\sqrt{6} \|\boldsymbol{\Delta}_{\boldsymbol{x}}\|_2}{2 m^{25}} \\
&\overset{\text{(iv)}}{\le} \left(1 + \mu^2\right) \frac{\sqrt{6} \|\boldsymbol{\Delta}_{\boldsymbol{x}}\|_2}{2 m^{25}} \\
&\overset{\text{(v)}}{\le} \frac{\|\boldsymbol{\Delta}_{\boldsymbol{x}}\|_2}{m^{24}}.
\end{aligned} \tag{B.39}$$

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

$$\begin{aligned}
\mathbb{E}\left[\left| \boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{v} \right|^2 \right] &= \left| \boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \right|^2 \mathbb{E}\left[\left| \boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{v} \right|^2 \right] \\
&= \left| \boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{h}^\star \right| \left(2 \left| \boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}} \boldsymbol{v} \right|^2 + \|\boldsymbol{\Delta}_{\boldsymbol{x}}\|_2^2 \|\boldsymbol{v}\|_2^2 \right),
\end{aligned} \tag{B.40}$$

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 (1.7) 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

$$\left\|z_j - \mathbb{E}\left[z_j\right]\right\|_{\psi_2} \lesssim \left|C_4 r \sqrt{\log m}\left(\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\right)\left(\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^{\star}\right)\right| \leq C_4 \frac{\mu\sqrt{\log m}}{\sqrt{m}} r \left|\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\right|.$$

Therefore, by invoking Hoeffding's inequality (cf. [Ver18, Theorem 2.6.2]) we reach

$$\mathbb{P}\left(\left|\sum_{j=1}^{m} z_j - \mathbb{E}\left[z_j\right]\right| \geq t\right) \leq 2\exp\left(-\frac{ct^2}{\frac{C_4^2\mu^2 r^2\log m}{m}\sum_{j=1}^{m}\left|\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\right|^2}\right) = 2\exp\left(-\frac{ct^2}{\frac{C_4^2\mu^2 r^2\log m}{m}}\right)$$

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

$$\mathbb{P}\left(\left|\sum_{j=1}^{m} z_j - \mathbb{E}\left[z_j\right]\right| \geq \frac{C\mu r\sqrt{K}\log m}{\sqrt{m}}\right) \leq 2\exp\left(-10K\log m\right). \tag{B.41}$$

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 $\varepsilon_1 = r/\left(m\log m\right)$ and $\varepsilon_2 = 1/\left(m\log m\right)$. In view of [Ver18, 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

$$\left|\sum_{j=1}^{m} z_j - \mathbb{E}\left[z_j\right]\right| \geq \frac{C\mu r\sqrt{K}\log m}{\sqrt{m}},$$

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

$$\begin{aligned}
&\left|\sum_{j=1}^{m}\left(\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^{\star}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{|\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j| \leq C_4 r\sqrt{\log m}\right\}} - \boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^{\star}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{v}\right)\right| \\
&\qquad \leq \left|\sum_{j=1}^{m}\left(z_j - \mathbb{E}\left[z_j\right]\right)\right| + \left|\sum_{j=1}^{m}\mathbb{E}\left[\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^{\star}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{|\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j| \leq C_4 r\sqrt{\log m}\right\}}\right]\right| \\
&\qquad \leq \frac{C\mu r\sqrt{K}\log m}{\sqrt{m}} + \frac{\|\boldsymbol{\Delta}_{\boldsymbol{x}}\|_2}{m^{24}} \\
&\qquad \leq \frac{\lambda}{100}
\end{aligned} \tag{B.42}$$

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.39) and (B.41). For any $\boldsymbol{x}$ obeying the assumption $\max_j \left|\left(\boldsymbol{x} - \boldsymbol{x}^{\star}\right)^{\mathsf{H}}\boldsymbol{a}_j\right| \leq C_3 r\sqrt{\log m}$ 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 $\max\left\{\|\boldsymbol{u} - \boldsymbol{u}_0\|_2, \|\boldsymbol{v} - \boldsymbol{v}_0\|_2\right\} \leq \varepsilon_2$.

Given that $\max_j \|\boldsymbol{a}_j\|_2 \le 10\sqrt{K}$ with probability $1 - me^{-CK}$ for some constant $C > 0$, this yields that

$$\left|\boldsymbol{\Delta}_{\boldsymbol{x}_0}^{\mathsf{H}}\boldsymbol{a}_j\right| \le \left|\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j\right| + 10\varepsilon_1\sqrt{K} \le 2C_3\left(\lambda + \sigma\sqrt{K\log m}\right)\sqrt{\log m}.$$

Recalling $C_4 \ge 10C_3$, we have

$$\left|\boldsymbol{\Delta}_{\boldsymbol{x}_0}^{\mathsf{H}}\boldsymbol{a}_j\right| \le C_4\left(\lambda + \sigma\sqrt{K\log m}\right)\sqrt{\log m} = C_4 r\sqrt{\log m},$$

and hence $\mathbb{1}_{\left\{\left|\boldsymbol{\Delta}_{\boldsymbol{x}_0}^{\mathsf{H}}\boldsymbol{a}_j\right| \le C_4 r\sqrt{\log m}\right\}} = 1$, $\forall j$. Therefore, if we let

$$f(\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}) := \sum_{j=1}^{m}\left(\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star(\boldsymbol{x}-\boldsymbol{x}^\star)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{\left|(\boldsymbol{x}-\boldsymbol{x}^\star)^{\mathsf{H}}\boldsymbol{a}_j\right| \le C_4 r\sqrt{\log m}\right\}} - \boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star(\boldsymbol{x}-\boldsymbol{x}^\star)^{\mathsf{H}}\boldsymbol{v}\right),$$

then we can demonstrate that

$$\begin{aligned}
&|f(\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}) - f(\boldsymbol{x}_0,\boldsymbol{u}_0,\boldsymbol{v}_0)| \\
&\le \left|\sum_{j=1}^{m}\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star(\boldsymbol{x}-\boldsymbol{x}_0)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\right| + \left|\sum_{j=1}^{m}\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star(\boldsymbol{x}-\boldsymbol{x}_0)^{\mathsf{H}}\boldsymbol{v}\right| \\
&\quad + \left|\sum_{j=1}^{m}(\boldsymbol{u}-\boldsymbol{u}_0)^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star(\boldsymbol{x}_0-\boldsymbol{x}^\star)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\right| + \left|\sum_{j=1}^{m}(\boldsymbol{u}-\boldsymbol{u}_0)^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star(\boldsymbol{x}_0-\boldsymbol{x}^\star)^{\mathsf{H}}\boldsymbol{v}\right| \\
&\quad + \left|\sum_{j=1}^{m}\boldsymbol{u}_0^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star(\boldsymbol{x}_0-\boldsymbol{x}^\star)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}(\boldsymbol{v}-\boldsymbol{v}_0)\right| + \left|\sum_{j=1}^{m}\boldsymbol{u}_0^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star(\boldsymbol{x}_0-\boldsymbol{x}^\star)^{\mathsf{H}}(\boldsymbol{v}-\boldsymbol{v}_0)\right| \\
&\le \left(\|\mathcal{A}\|^2 + 1\right)\left(\|\boldsymbol{h}^\star\|_2\|\boldsymbol{x}-\boldsymbol{x}_0\|_2 + \|\boldsymbol{x}_0-\boldsymbol{x}^\star\|_2\|\boldsymbol{u}-\boldsymbol{u}_0\|_2 + \|\boldsymbol{x}_0-\boldsymbol{x}^\star\|_2\|\boldsymbol{v}-\boldsymbol{v}_0\|_2\right) \\
&\le (2K\log K + 10\log m + 1)(\varepsilon_1 + 2C_1 r\varepsilon_2),
\end{aligned}$$

where the last inequality arises from (B.38). Consequently,

$$\begin{aligned}
&\left|\boldsymbol{u}^{\mathsf{H}}\mathcal{T}^{\mathsf{debias}}\left(\boldsymbol{h}^\star(\boldsymbol{x}-\boldsymbol{x}^\star)^{\mathsf{H}}\right)\boldsymbol{v}\right| \\
&\quad = \left|\sum_{j=1}^{m}\left(\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star(\boldsymbol{x}-\boldsymbol{x}^\star)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{\left|(\boldsymbol{x}-\boldsymbol{x}^\star)^{\mathsf{H}}\boldsymbol{a}_j\right| \le C_4 r\sqrt{\log m}\right\}} - \boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^\star(\boldsymbol{x}-\boldsymbol{x}^\star)^{\mathsf{H}}\boldsymbol{v}\right)\right| \\
&\quad \le |f(\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}) - f(\boldsymbol{x}_0,\boldsymbol{u}_0,\boldsymbol{v}_0)| + |f(\boldsymbol{x}_0,\boldsymbol{u}_0,\boldsymbol{v}_0)| \\
&\quad \le (2K\log K + 10\log m + 1)(\varepsilon_1 + 2C_1 r\varepsilon_2) + \frac{\lambda}{100} \\
&\quad \le \frac{\lambda}{50},
\end{aligned}$$

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 (2.5), there holds

$$\left\|\mathcal{T}^{\mathsf{debias}}\left(\boldsymbol{h}^\star\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\right)\right\| = \sup_{\boldsymbol{u},\boldsymbol{v}\in\mathcal{S}^{K-1}}\boldsymbol{u}^{\mathsf{H}}\mathcal{T}^{\mathsf{debias}}\left(\boldsymbol{h}^\star(\boldsymbol{x}-\boldsymbol{x}^\star)^{\mathsf{H}}\right)\boldsymbol{v} \le \frac{1}{50}\lambda \tag{B.43}$$

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

$$\begin{aligned}
&\boldsymbol{u}^{\mathsf{H}}\mathcal{T}^{\mathsf{debias}}\left(\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{x}^{\star\mathsf{H}}\right)\boldsymbol{v} \\
&= \sum_{j=1}^{m}\left(\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v} - \boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{v}\right)
\end{aligned}$$

$$= \sum_{j=1}^{m} \Bigg(\underbrace{\boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{\Delta}_{\boldsymbol{h}} \boldsymbol{x}^{\star\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{v}\, \mathbb{1}_{\left\{ |\boldsymbol{x}^{\star\mathsf{H}} \boldsymbol{a}_j| \leq 20\sqrt{\log m} \right\}}}_{=: y_j} - \mathbb{E}\left[y_j \right] \Bigg)$$

$$- \underbrace{\sum_{j=1}^{m} \mathbb{E}\left[\boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{\Delta}_{\boldsymbol{h}} \boldsymbol{x}^{\star\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{v}\, \mathbb{1}_{\left\{ |\boldsymbol{x}^{\star\mathsf{H}} \boldsymbol{a}_j| > 20\sqrt{\log m} \right\}} \right]}_{=: \omega_4} + \sum_{j=1}^{m} \boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{\Delta}_{\boldsymbol{h}} \boldsymbol{x}^{\star\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{v}\, \mathbb{1}_{\left\{ |\boldsymbol{x}^{\star\mathsf{H}} \boldsymbol{a}_j| > 20\sqrt{\log m} \right\}}.$$

- For ω_4, similar to (B.39) we have

$$\begin{aligned}
|\omega_4| &= \left| \sum_{j=1}^{m} \mathbb{E}\left[\boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{\Delta}_{\boldsymbol{h}} \boldsymbol{x}^{\star\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{v}\, \mathbb{1}_{\left\{ |\boldsymbol{x}^{\star\mathsf{H}} \boldsymbol{a}_j| > 20\sqrt{\log m} \right\}} \right] \right| \\
&\overset{\text{(i)}}{\leq} \sum_{j=1}^{m} \sqrt{ \mathbb{E}\left[\left| \boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{\Delta}_{\boldsymbol{h}} \boldsymbol{x}^{\star\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{v} \right|^2 \right] \mathbb{P}\left(|\boldsymbol{x}^{\star\mathsf{H}} \boldsymbol{a}_j| > 20\sqrt{\log m} \right) } \\
&\overset{\text{(ii)}}{\leq} \sum_{j=1}^{m} \left| \boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{\Delta}_{\boldsymbol{h}} \right| \sqrt{ \left(2 \left| \boldsymbol{x}^{\star\mathsf{H}} \boldsymbol{v} \right|^2 + \left\| \boldsymbol{x}^{\star} \right\|_2^2 \left\| \boldsymbol{v} \right\|_2^2 \right) 2 \exp\left(-200 \log m \right) } \\
&\leq \sum_{j=1}^{m} \left| \boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{\Delta}_{\boldsymbol{h}} \right| \frac{4}{m^{100}} \\
&\overset{\text{(iii)}}{\leq} \sum_{j=1}^{m} \left\| \boldsymbol{b}_j \right\|_2 \times C_9 \sigma \times \frac{4}{m^{100}} && \text{(B.44)} \\
&\overset{\text{(iv)}}{\leq} \sqrt{\frac{K}{m}} \times m \times C_9 \sigma \times \frac{4}{m^{100}} \\
&\leq \frac{\lambda}{m^{99}}, && \text{(B.45)}
\end{aligned}$$

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.40), (iii) is due to the assumption (2.5c), and (iv) comes from the fact $\left\| \boldsymbol{b}_j \right\|_2 = \sqrt{K/m}$.

- Regarding the term $\omega_3 := \sum_{j=1}^{m} \left(y_j - \mathbb{E}\left[y_j \right] \right)$, we note that

$$\left\| \boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{\Delta}_{\boldsymbol{h}} \boldsymbol{x}^{\star\mathsf{H}} \boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}} \boldsymbol{v}\, \mathbb{1}_{\left\{ |\boldsymbol{x}^{\star\mathsf{H}} \boldsymbol{a}_j| \leq 20\sqrt{\log m} \right\}} \right\|_{\psi_2} \leq \frac{\mu \lambda}{\sqrt{m}} \log^2 m \times 20 \sqrt{\log m} \left| \boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \right|.$$

Hoeffding's inequality [Ver18, Theorem 2.6.3] tells us that

$$\mathbb{P}\left(\left| \sum_{j=1}^{m} \left(y_j - \mathbb{E}\left[y_j \right] \right) \right| \geq t \right) \leq 2 \exp\left(- \frac{c t^2}{400 \frac{\mu^2 \lambda^2}{m} \log^5 m \sum_{j=1}^{m} \left| \boldsymbol{u}^{\mathsf{H}} \boldsymbol{b}_j \right|^2} \right) = 2 \exp\left(- \frac{c t^2}{400 \frac{\mu^2 \lambda^2}{m} \log^5 m} \right)$$

for any $t \geq 0$. Setting $t = \frac{C \mu \lambda \sqrt{K}}{\sqrt{m}} \log^3 m$ for some sufficiently large constant $C > 0$ yields

$$\mathbb{P}\left(\left| \sum_{j=1}^{m} \left(y_j - \mathbb{E}\left[y_j \right] \right) \right| \geq C \frac{\mu \lambda \sqrt{K}}{\sqrt{m}} \log^3 m \right) \leq 2 \exp\left(-10 K \log m \right). \tag{B.46}$$

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 $\mathcal{B}_{\boldsymbol{h}}\left(\frac{C_5}{1-\rho}\eta r\right) := \left\{\boldsymbol{h} : \|\boldsymbol{h}-\boldsymbol{h}^\star\|_2 \le \frac{C_5}{1-\rho}\eta r\right\}$ 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

$$\begin{aligned}
&\left|\sum_{j=1}^m \left(\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{|\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j|\le 20\sqrt{\log m}\right\}} - \boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{v}\right)\right| \\
&\quad\le \left|\sum_{j=1}^m (y_j - \mathbb{E}[y_j])\right| + \left|\sum_{j=1}^m \mathbb{E}\left[\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{|\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j|> 20\sqrt{\log m}\right\}}\right]\right| \\
&\quad\le C\frac{\mu\lambda\sqrt{K}}{\sqrt{m}}\log^3 m + \frac{\lambda}{m^{99}} \\
&\quad\le \frac{\lambda}{100},
\end{aligned} \tag{B.47}$$

where the penultimate inequality comes from (B.45) and (B.46). Next, let us define

$$g(\boldsymbol{h},\boldsymbol{u},\boldsymbol{v}) := \sum_{j=1}^m \left(\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}(\boldsymbol{h}-\boldsymbol{h}^\star)\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{|\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j|\le 20\sqrt{\log m}\right\}} - \boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}(\boldsymbol{h}-\boldsymbol{h}^\star)\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{v}\right).$$

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 $\|\boldsymbol{h}-\boldsymbol{h}_0\|_2 \le \varepsilon_1$ and $\max\{\|\boldsymbol{u}-\boldsymbol{u}_0\|_2, \|\boldsymbol{v}-\boldsymbol{v}_0\|_2\} \le \varepsilon_2$, this guarantees that

$$\begin{aligned}
&|g(\boldsymbol{h},\boldsymbol{u},\boldsymbol{v}) - g(\boldsymbol{h}_0,\boldsymbol{u}_0,\boldsymbol{v}_0)| \\
&\quad\le \left|\sum_{j=1}^m \boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}(\boldsymbol{h}-\boldsymbol{h}_0)\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| + \left|\sum_{j=1}^m \boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}(\boldsymbol{h}-\boldsymbol{h}_0)\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{v}\right| \\
&\qquad+\left|\sum_{j=1}^m \left((\boldsymbol{u}-\boldsymbol{u}_0)^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}(\boldsymbol{h}-\boldsymbol{h}^\star)\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\}} - (\boldsymbol{u}-\boldsymbol{u}_0)^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}(\boldsymbol{h}-\boldsymbol{h}^\star)\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}}(\boldsymbol{h}-\boldsymbol{h}^\star)\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}(\boldsymbol{v}-\boldsymbol{v}_0)\,\mathbb{1}_{\left\{|\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j|\le 20\sqrt{\log m}\right\}} - \boldsymbol{u}_0^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}(\boldsymbol{h}-\boldsymbol{h}^\star)\boldsymbol{x}^{\star\mathsf{H}}(\boldsymbol{v}-\boldsymbol{v}_0)\right)\right| \\
&\quad\le \left(\|\mathcal{A}\|^2+1\right)\left(\|\boldsymbol{x}^\star\|_2\|\boldsymbol{h}-\boldsymbol{h}_0\|_2 + \|(\boldsymbol{h}-\boldsymbol{h}^\star)\|_2\|\boldsymbol{u}-\boldsymbol{u}_0\|_2 + \|\boldsymbol{h}-\boldsymbol{h}^\star\|_2\|\boldsymbol{v}-\boldsymbol{v}_0\|_2\right) \\
&\quad\le (2K\log K + 10\log m + 1)(\varepsilon_1 + 2C_1 r\varepsilon_2),
\end{aligned}$$

where the last inequality comes from (B.38). Since $\mathbb{P}\left(\left|\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\right| > 20\sqrt{\log m}\right) \le O\left(m^{-100}\right)$ (in view of (D.1)), we have, with probability exceeding $1 - O\left(m^{-10}\right)$, that

$$\begin{aligned}
\left\|\mathcal{T}^{\mathsf{debias}}\left(\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{x}^{\star\mathsf{H}}\right)\right\| &= \sup_{\boldsymbol{u},\boldsymbol{v}\in\mathcal{S}^{K-1}}\left|\sum_{j=1}^m \left(\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{|\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{a}_j|\le 20\sqrt{\log m}\right\}} - \boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{x}^{\star\mathsf{H}}\boldsymbol{v}\right)\right| \\
&\le \sup_{\boldsymbol{u},\boldsymbol{v}\in\mathcal{S}^{K-1}} |g(\boldsymbol{h},\boldsymbol{u},\boldsymbol{v}) - g(\boldsymbol{h}_0,\boldsymbol{u}_0,\boldsymbol{v}_0)| + |g(\boldsymbol{h}_0,\boldsymbol{u}_0,\boldsymbol{v}_0)| \\
&\le (2K\log K + 10\log m + 1)(\varepsilon_1 + 2C_1 r\varepsilon_2) + \frac{\lambda}{100} \\
&\le \frac{\lambda}{50}
\end{aligned} \tag{B.48}$$

holds uniformly over $\boldsymbol{h} \in \mathcal{B}_{\boldsymbol{h}}(C_1 r)$, where the last inequality is due to the choices $\varepsilon_1 = r/(m\log m)$, $\varepsilon_2 = 1/(m\log m)$ 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

$$\boldsymbol{u}^{\mathsf{H}}\mathcal{T}^{\mathsf{debias}}\left(\boldsymbol{\Delta_h}\boldsymbol{\Delta_x}^{\mathsf{H}}\right)\boldsymbol{v} = \sum_{j=1}^{m}\left(\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta_h}\boldsymbol{\Delta_x}^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v} - \boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta_h}\boldsymbol{\Delta_x}^{\mathsf{H}}\boldsymbol{v}\right),$$

which is 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 demonstrate that

$$\begin{aligned}
&\sum_{j=1}^{m}\Bigg(\underbrace{\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta_h}\boldsymbol{\Delta_x}^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{|\boldsymbol{\Delta_x}^{\mathsf{H}}\boldsymbol{a}_j|\le C_4 r\sqrt{\log m}\right\}}}_{=:s_j} - \boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta_h}\boldsymbol{\Delta_x}^{\mathsf{H}}\boldsymbol{v}\Bigg) \\
&\quad = \sum_{j=1}^{m}\left(s_j - \mathbb{E}\left[s_j\right]\right) + \sum_{j=1}^{m}\left(\mathbb{E}\left[\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta_h}\boldsymbol{\Delta_x}^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{|\boldsymbol{\Delta_x}^{\mathsf{H}}\boldsymbol{a}_j|\le C_4 r\sqrt{\log m}\right\}}\right] - \boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{h}^{\star}\boldsymbol{\Delta_x}^{\mathsf{H}}\boldsymbol{v}\right) \\
&\quad = \underbrace{\sum_{j=1}^{m}\left(s_j - \mathbb{E}\left[s_j\right]\right)}_{=:\omega_5} - \underbrace{\sum_{j=1}^{m}\mathbb{E}\left[\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta_h}\boldsymbol{\Delta_x}^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{|\boldsymbol{\Delta_x}^{\mathsf{H}}\boldsymbol{a}_j|> C_4 r\sqrt{\log m}\right\}}\right]}_{=:\omega_6}.
\end{aligned}$$

- With regards to ω_6, similar to (B.39) we have

$$\begin{aligned}
|\omega_6| &= \left|\sum_{j=1}^{m}\mathbb{E}\left[\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta_h}\boldsymbol{\Delta_x}^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{|\boldsymbol{\Delta_x}^{\mathsf{H}}\boldsymbol{a}_j|> C_4 r\sqrt{\log m}\right\}}\right]\right| \\
&\overset{\text{(i)}}{\le} \sum_{j=1}^{m}\sqrt{\mathbb{E}\left[\left|\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta_h}\boldsymbol{\Delta_x}^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\right|^2\right]\mathbb{P}\left[\mathbb{1}_{\left\{|\boldsymbol{\Delta_x}^{\mathsf{H}}\boldsymbol{a}_j|> C_4 r\sqrt{\log m}\right\}}\right]} \\
&\overset{\text{(ii)}}{\le} \sum_{j=1}^{m}\left|\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta_h}\right|\sqrt{\left(2\left|\boldsymbol{\Delta_x}^{\mathsf{H}}\boldsymbol{v}\right|^2 + \left\|\boldsymbol{\Delta_x}\right\|_2^2\left\|\boldsymbol{v}\right\|_2^2\right)2\exp\left(-\frac{C_4^2 r^2\log m}{2\left\|\boldsymbol{\Delta_x}\right\|_2^2}\right)} \\
&\le \sum_{j=1}^{m}\left|\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta_h}\right|\sqrt{6\left\|\boldsymbol{\Delta_x}\right\|_2^2\exp\left(-50\log m\right)} \\
&\le \sum_{j=1}^{m}\left\|\boldsymbol{b}_j\right\|_2\left|\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta_h}\right|\frac{\sqrt{6}\left\|\boldsymbol{\Delta_x}\right\|_2}{m^{25}} \\
&\overset{\text{(iii)}}{\le} \frac{\lambda\left\|\boldsymbol{\Delta_x}\right\|_2}{m^{24}},
\end{aligned}$$

  where (i) follows from Cauchy-Schwarz inequality, (ii) comes from the property of sub-Gaussian variable $\left|\boldsymbol{\Delta_x}^{\mathsf{H}}\boldsymbol{a}_j\right|$ and (B.40), and (iii) is due to the fact $\|\boldsymbol{b}_j\|_2 = \sqrt{K/m}$ and the assumption (2.5c).
- Regarding ω_5, we note that s_j is a sub-Gaussian random variable satisfying

$$\left\|s_j - \mathbb{E}\left[s_j\right]\right\|_{\psi_2} \lesssim C_4 r\sqrt{\log m}\left|\left(\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\right)\left(\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta_h}\right)\right| \le C_4\frac{\mu\sqrt{\log^5 m}}{\sqrt{m}}r\left|\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\right|.$$

  Therefore, invoking Hoeffding's inequality (cf. [Ver18, Theorem 2.6.3]) reveals that

$$\mathbb{P}\left(\left|\sum_{j=1}^{m}s_j - \mathbb{E}\left[s_j\right]\right| \ge t\right) \le 2\exp\left(-\frac{ct^2}{\frac{C_4^2\mu^2 r^2\log^5 m}{m}\sum_{j=1}^{m}\left|\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\right|^2}\right) = 2\exp\left(-\frac{ct^2}{\frac{C_4^2\mu^2 r^2\log^5 m}{m}}\right)$$

  for any $t \ge 0$. Setting $t = \frac{C\mu r\sqrt{K}\log^3 m}{\sqrt{m}}$ for some sufficiently large constant $C > 0$, we obtain

$$\mathbb{P}\left(\left|\sum_{j=1}^{m}s_j - \mathbb{E}\left[s_j\right]\right| \ge \frac{C\mu r\sqrt{K}\log^3 m}{\sqrt{m}}\right) \le 2\exp\left(-10K\log m\right). \tag{B.49}$$

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 $\mathcal{B}_{\boldsymbol{h}}\left(\frac{C_5}{1-\rho}\eta r\right) := \left\{\boldsymbol{h} : \|\boldsymbol{h}-\boldsymbol{h}^\star\|_2 \le \frac{C_5}{1-\rho}\eta r\right\}$, $\mathcal{N}_{\boldsymbol{x}}$ an ε_1-net of $\mathcal{B}_{\boldsymbol{x}}\left(\frac{C_5}{1-\rho}\eta r\right) := \left\{\boldsymbol{x} : \|\boldsymbol{x}-\boldsymbol{x}^\star\|_2 \le \frac{C_5}{1-\rho}\eta r\right\}$, 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 [Ver18, Corollary 4.2.13], these epsilon nets can be chosen to satisfy the following cardinality bounds

$$|\mathcal{N}_{\boldsymbol{h}}| \le \left(1+\frac{2C_5\eta r}{(1-\rho)\,\varepsilon_1}\right)^{2K}, \quad |\mathcal{N}_{\boldsymbol{x}}| \le \left(1+\frac{2C_5\eta r}{(1-\rho)\,\varepsilon_1}\right)^{2K} \quad \text{and} \quad |\mathcal{N}_0| \le \left(1+\frac{2}{\varepsilon_2}\right)^{2K}.$$

By taking the union bound, we show that with probability at least

$$1-\left(1+\frac{2C_5\eta r}{(1-\rho)\,\varepsilon_1}\right)^{4K}\left(1+\frac{2}{\varepsilon_2}\right)^{4K} e^{-10K\log m} \ge 1-O\left(m^{-100}\right),$$

the following bound

$$\left|\sum_{j=1}^m s_j - \mathbb{E}\left[s_j\right]\right| \ge \frac{C\mu r\sqrt{K}\log^3 m}{\sqrt{m}}$$

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

$$\left|\sum_{j=1}^m \left(\underbrace{\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{|\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j|\le C_4 r\sqrt{\log m}\right\}}}_{=:s_j} - \boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\boldsymbol{\Delta}_{\boldsymbol{h}}\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{v}\right)\right| \tag{B.50}$$
$$\le \frac{C\mu r\sqrt{K}\log^3 m}{\sqrt{m}} + \frac{\lambda\,\|\boldsymbol{\Delta}_{\boldsymbol{x}}\|_2}{m^{24}} \le \frac{\lambda}{100}$$

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 $\max_{1\le j\le m}\left|(\boldsymbol{x}-\boldsymbol{x}^\star)^{\mathsf{H}}\boldsymbol{a}_j\right| \le C_3 r\sqrt{\log m}$ 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 $\max\left\{\|\boldsymbol{h}-\boldsymbol{h}_0\|_2, \|\boldsymbol{x}-\boldsymbol{x}_0\|_2\right\} \le \varepsilon_1$ and $\max\left\{\|\boldsymbol{u}-\boldsymbol{u}_0\|_2, \|\boldsymbol{v}-\boldsymbol{v}_0\|_2\right\} \le \varepsilon_2$. Recognizing that $\|\boldsymbol{a}_j\|_2 \le 10\sqrt{K}$ with probability $1-O\left(me^{-CK}\right)$ for some constant $C>0$ (see (D.2)), we can guarantee that

$$\left|\boldsymbol{\Delta}_{\boldsymbol{x}_0}^{\mathsf{H}}\boldsymbol{a}_j\right| \le \left|\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j\right| + 10\varepsilon_1\sqrt{K} \le 2C_3\left(\lambda+\sigma\sqrt{K\log m}\right)\sqrt{\log m}.$$

Recalling that $C_4 \ge 10C_3$, we have

$$\left|\boldsymbol{\Delta}_{\boldsymbol{x}_0}^{\mathsf{H}}\boldsymbol{a}_j\right| \le C_4\left(\lambda+\sigma\sqrt{K\log m}\right)\sqrt{\log m} = C_4 r\sqrt{\log m},$$

and hence $\mathbb{1}_{\left\{|\boldsymbol{\Delta}_{\boldsymbol{x}_0}^{\mathsf{H}}\boldsymbol{a}_j|\le C_4 r\sqrt{\log m}\right\}} = 1$ for all $1\le j\le m$. Therefore, if we take

$$r\left(\boldsymbol{h},\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}\right) := \sum_{j=1}^m \left(\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{h}-\boldsymbol{h}^\star\right)\left(\boldsymbol{x}-\boldsymbol{x}^\star\right)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{|\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j|\le C_4 r\sqrt{\log m}\right\}} - \boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{h}-\boldsymbol{h}^\star\right)\left(\boldsymbol{x}-\boldsymbol{x}^\star\right)^{\mathsf{H}}\boldsymbol{v}\right),$$

then it follows that

$$\begin{aligned}
&\left|r\left(\boldsymbol{h},\boldsymbol{x},\boldsymbol{u},\boldsymbol{v}\right) - r\left(\boldsymbol{h}_0,\boldsymbol{x}_0,\boldsymbol{u}_0,\boldsymbol{v}_0\right)\right| \\
&\quad\le \left|\sum_{j=1}^m \boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{h}-\boldsymbol{h}_0\right)\left(\boldsymbol{x}-\boldsymbol{x}^\star\right)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\right| + \left|\sum_{j=1}^m \boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{h}-\boldsymbol{h}_0\right)\left(\boldsymbol{x}-\boldsymbol{x}_0\right)^{\mathsf{H}}\boldsymbol{v}\right| \\
&\qquad + \left|\sum_{j=1}^m \boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{h}_0-\boldsymbol{h}^\star\right)\left(\boldsymbol{x}-\boldsymbol{x}_0\right)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\right| + \left|\sum_{j=1}^m \boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}\left(\boldsymbol{h}_0-\boldsymbol{h}^\star\right)\left(\boldsymbol{x}-\boldsymbol{x}_0\right)^{\mathsf{H}}\boldsymbol{v}\right|
\end{aligned}$$

$$+\left|\sum_{j=1}^{m}\left((\boldsymbol{u}-\boldsymbol{u}_0)^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}(\boldsymbol{h}_0-\boldsymbol{h}^\star)(\boldsymbol{x}_0-\boldsymbol{x}^\star)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}-(\boldsymbol{u}-\boldsymbol{u}_0)^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}(\boldsymbol{h}_0-\boldsymbol{h}^\star)(\boldsymbol{x}_0-\boldsymbol{x}^\star)^{\mathsf{H}}\boldsymbol{v}\right)\right|$$

$$+\left|\sum_{j=1}^{m}\left(\boldsymbol{u}_0^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}(\boldsymbol{h}_0-\boldsymbol{h}^\star)(\boldsymbol{x}_0-\boldsymbol{x}^\star)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}(\boldsymbol{v}-\boldsymbol{v}_0)-\boldsymbol{u}_0^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}(\boldsymbol{h}_0-\boldsymbol{h}^\star)(\boldsymbol{x}_0-\boldsymbol{x}^\star)^{\mathsf{H}}(\boldsymbol{v}-\boldsymbol{v}_0)\right)\right|$$

$$\le\left(\|\mathcal{A}\|^2+1\right)\|\boldsymbol{h}-\boldsymbol{h}_0\|_2\|\boldsymbol{x}-\boldsymbol{x}_0\|_2+\left(\|\mathcal{A}\|^2+1\right)\|\boldsymbol{h}_0-\boldsymbol{h}^\star\|_2\|\boldsymbol{x}-\boldsymbol{x}_0\|_2$$

$$+\left(\|\mathcal{A}\|^2+1\right)\|\boldsymbol{x}_0-\boldsymbol{x}^\star\|_2\|\boldsymbol{u}-\boldsymbol{u}_0\|_2+\left(\|\mathcal{A}\|^2+1\right)\|\boldsymbol{x}_0-\boldsymbol{x}^\star\|_2\|\boldsymbol{v}-\boldsymbol{v}_0\|_2$$

$$\le(2K\log K+10\log m+1)\left(2(\varepsilon_1)^2+2C_1r\varepsilon_2\right), \tag{B.51}$$

where the last inequality arises from (B.38). This further leads to

$$\left|\boldsymbol{u}^{\mathsf{H}}\mathcal{T}^{\mathsf{debias}}\left((\boldsymbol{h}-\boldsymbol{h}^\star)(\boldsymbol{x}-\boldsymbol{x}^\star)^{\mathsf{H}}\right)\boldsymbol{v}\right|$$

$$=\left|\sum_{j=1}^{m}\left(\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}(\boldsymbol{h}-\boldsymbol{h}^\star)(\boldsymbol{x}-\boldsymbol{x}^\star)^{\mathsf{H}}\boldsymbol{a}_j\boldsymbol{a}_j^{\mathsf{H}}\boldsymbol{v}\,\mathbb{1}_{\left\{|\boldsymbol{\Delta}_{\boldsymbol{x}}^{\mathsf{H}}\boldsymbol{a}_j|\le C_4r\sqrt{\log m}\right\}}-\boldsymbol{u}^{\mathsf{H}}\boldsymbol{b}_j\boldsymbol{b}_j^{\mathsf{H}}(\boldsymbol{h}-\boldsymbol{h}^\star)(\boldsymbol{x}-\boldsymbol{x}^\star)^{\mathsf{H}}\boldsymbol{v}\right)\right|$$

$$=|r(\boldsymbol{h},\boldsymbol{x},\boldsymbol{u},\boldsymbol{v})-r(\boldsymbol{h}_0,\boldsymbol{x}_0,\boldsymbol{u}_0,\boldsymbol{v}_0)|+|r(\boldsymbol{h}_0,\boldsymbol{x}_0,\boldsymbol{u}_0,\boldsymbol{v}_0)|$$

$$\le(2K\log K+10\log m+1)\left(2(\varepsilon_1)^2+2C_1r\varepsilon_2\right)+\frac{\lambda}{100}$$

$$\le\frac{\lambda}{50},$$

where the last inequality follows from (B.50) and (B.51). As a consequence, for any point $(\boldsymbol{h},\boldsymbol{x})$ satisfying (2.5), we have, with probability exceeding $1-O\left(m^{-10}+me^{-CK}\right)$, that

$$\left\|\mathcal{T}^{\mathsf{debias}}\left((\boldsymbol{h}-\boldsymbol{h}^\star)(\boldsymbol{x}-\boldsymbol{x}^\star)^{\mathsf{H}}\right)\right\|=\sup_{\boldsymbol{u},\boldsymbol{v}\in\mathcal{S}^{K-1}}\boldsymbol{u}^{\mathsf{H}}\mathcal{T}^{\mathsf{debias}}\left((\boldsymbol{h}-\boldsymbol{h}^\star)(\boldsymbol{x}-\boldsymbol{x}^\star)^{\mathsf{H}}\right)\boldsymbol{v}\le\frac{1}{50}\lambda. \tag{B.52}$$

To finish up, combining the bounds obtained in (B.43), (B.48) and (B.52), we arrive at

$$\left\|\mathcal{T}^{\mathsf{debias}}\left(\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}-\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}\right)\right\|\le\frac{\lambda}{50}+\frac{\lambda}{50}+\frac{\lambda}{50}<\frac{\lambda}{8}.$$

B.3.4 Proof of Lemma 17

Recall the definition of $\mathcal{T}^{\mathsf{debias}}$ in (A.8). Letting

$$\boldsymbol{p}=\frac{1}{\|\boldsymbol{h}\|_2}\boldsymbol{h}\qquad\text{and}\qquad\boldsymbol{q}=\frac{1}{\|\boldsymbol{x}\|_2}\boldsymbol{x} \tag{B.53}$$

and rearranging terms, we can write

$$\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}+\mathcal{T}^{\mathsf{debias}}\left(\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}-\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}\right)+\mathcal{A}^*(\boldsymbol{\xi})=\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}+\lambda\boldsymbol{p}\boldsymbol{q}^{\mathsf{H}}+\boldsymbol{R} \tag{B.54}$$

for some matrix $\boldsymbol{R}$. In addition, in view of the small gradient assumption (2.5a), one has

$$\left[\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}+\mathcal{T}^{\mathsf{debias}}\left(\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}-\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}\right)+\mathcal{A}^*(\boldsymbol{\xi})\right]\boldsymbol{x}=\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}\boldsymbol{x}+\lambda\boldsymbol{h}-\boldsymbol{r}_1 \tag{B.55a}$$

$$\left[\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}+\mathcal{T}^{\mathsf{debias}}\left(\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}-\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}\right)+\mathcal{A}^*(\boldsymbol{\xi})\right]^{\mathsf{H}}\boldsymbol{h}=\boldsymbol{x}\boldsymbol{h}^{\mathsf{H}}\boldsymbol{h}+\lambda\boldsymbol{x}-\boldsymbol{r}_2 \tag{B.55b}$$

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

$$\|\boldsymbol{r}_1\|_2=\left\|\lambda\boldsymbol{h}-\left(\mathcal{T}\left(\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}-\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}\right)+\mathcal{A}^*(\boldsymbol{\xi})\right)\boldsymbol{x}\right\|_2\le\|\nabla f(\boldsymbol{h},\boldsymbol{x})\|_2\le C\frac{\lambda}{m^{10}}, \tag{B.56a}$$

$$\|\boldsymbol{r}_2\|_2=\left\|\lambda\boldsymbol{x}-\left(\mathcal{T}\left(\boldsymbol{h}^\star\boldsymbol{x}^{\star\mathsf{H}}-\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}\right)+\mathcal{A}^*(\boldsymbol{\xi})\right)^{\mathsf{H}}\boldsymbol{h}\right\|_2\le\|\nabla f(\boldsymbol{h},\boldsymbol{x})\|_2\le C\frac{\lambda}{m^{10}}. \tag{B.56b}$$

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

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

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

where $\boldsymbol{p}$ and $\boldsymbol{q}$ are unit vectors defined in (B.53). Recognizing that $\left\|\boldsymbol{h}\right\|_2=\left\|\boldsymbol{x}\right\|_2$ (cf. (2.5b)), we can use (B.54) and (B.55) to obtain

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

These together with (B.56) yield

$$\left\|\mathcal{P}_T\left(\boldsymbol{R}\right)\right\|_{\mathrm{F}}\le\left\|\boldsymbol{p}^{\mathsf{H}}\boldsymbol{R}\right\|_2+\left\|\boldsymbol{R}\boldsymbol{q}\right\|_2\le 2\left\|\nabla f\left(\boldsymbol{h},\boldsymbol{x}\right)\right\|_2\le 2C\frac{\lambda}{m^{10}}. \tag{B.57}$$

2. We them move on to control $\mathcal{P}_{T^{\perp}}\left(\boldsymbol{R}\right)$. Continue the relation (B.54) to derive

$$\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}+\mathcal{T}^{\mathsf{debias}}\left(\boldsymbol{h}^{\star}\boldsymbol{x}^{\star\mathsf{H}}-\boldsymbol{h}\boldsymbol{x}^{\mathsf{H}}\right)+\mathcal{A}^{*}\left(\boldsymbol{\xi}\right)-\mathcal{P}_T\left(\boldsymbol{R}\right)=\boldsymbol{p}\left(\left\|\boldsymbol{h}\right\|_2\left\|\boldsymbol{x}\right\|_2+\lambda\frac{\left\|\boldsymbol{h}\right\|_2}{\left\|\boldsymbol{x}\right\|_2}\right)\boldsymbol{q}^{\mathsf{H}}+\mathcal{P}_{T^{\perp}}\left(\boldsymbol{R}\right), \tag{B.58}$$

where we have used the assumption $\left\|\boldsymbol{h}\right\|_2/\left\|\boldsymbol{x}\right\|_2=1$ (cf. (2.5b)). Combine this with Lemma 16, Lemma 19 and (B.57) to derive

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

where the last inequality invokes the assumption (B.2g). Invoking (B.58) and Weyl's inequality give

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

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

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

This indicates that at least $K-1$ singular values of $\boldsymbol{p}\left(\left\|\boldsymbol{h}\right\|_2\left\|\boldsymbol{x}\right\|_2+\lambda\left\|\boldsymbol{h}\right\|_2/\left\|\boldsymbol{x}\right\|_2\right)\boldsymbol{q}^{\mathsf{H}}+\mathcal{P}_{T^{\perp}}\left(\boldsymbol{R}\right)$ 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

$$\left\|\mathcal{P}_{T^{\perp}}\left(\boldsymbol{R}\right)\right\|\le\lambda/2.$$

C Proof for minimax lower bounds (Theorem 3)

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 $\boldsymbol{h}^{\star},\boldsymbol{x}^{\star}\in\mathbb{R}^K$. 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. [CP11, Lemma 3.11]) gives

$$\begin{aligned}\inf_{\widehat{\boldsymbol{x}}} \sup_{\boldsymbol{x}^\star \in \mathbb{C}^K} \mathbb{E}\left[\|\widehat{\boldsymbol{x}} - \boldsymbol{x}^\star\|_2^2 \,\middle|\, \boldsymbol{A}\right] &= \frac{1}{2}\sigma^2 \left(\mathsf{tr}\left[\left(\widetilde{\boldsymbol{A}}_{\mathsf{re}}^\top \widetilde{\boldsymbol{A}}_{\mathsf{re}}\right)^{-1}\right] + \mathsf{tr}\left[\left(\widetilde{\boldsymbol{A}}_{\mathsf{im}}^\top \widetilde{\boldsymbol{A}}_{\mathsf{im}}\right)^{-1}\right]\right) \\ &\geq K\sigma^2 / \max\left\{\left\|\widetilde{\boldsymbol{A}}_{\mathsf{re}}\right\|^2, \left\|\widetilde{\boldsymbol{A}}_{\mathsf{im}}\right\|^2\right\},\end{aligned} \tag{C.1}$$

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 (C.1) gives

$$\inf_{\widehat{\boldsymbol{x}}} \sup_{\boldsymbol{x}^\star \in \mathbb{C}^K} \mathbb{E}\left[\|\widehat{\boldsymbol{x}} - \boldsymbol{x}^\star\|_2^2 \,\middle|\, \boldsymbol{A}\right] \gtrsim \sigma^2 / \log m.$$

In turn, this oracle lower bound implies that, with high probability,

$$\begin{aligned}\inf_{\widehat{\boldsymbol{Z}}} \sup_{\boldsymbol{Z}^\star \in \mathcal{M}^\star} \mathbb{E}\left[\left\|\widehat{\boldsymbol{Z}} - \boldsymbol{Z}^\star\right\|_{\mathrm{F}}^2 \mid \boldsymbol{A}\right] &\gtrsim \inf_{\widehat{\boldsymbol{x}}} \sup_{\boldsymbol{x}^\star \in \mathbb{C}^K} \mathbb{E}\left[\left\|\boldsymbol{h}^\star \widehat{\boldsymbol{x}}^{\mathsf{H}} - \boldsymbol{h}^\star \boldsymbol{x}^{\star\mathsf{H}}\right\|_{\mathrm{F}}^2 \mid \boldsymbol{A}\right] \asymp \inf_{\widehat{\boldsymbol{x}}} \sup_{\boldsymbol{x}^\star \in \mathbb{C}^K} \mathbb{E}\left[\|\widehat{\boldsymbol{x}} - \boldsymbol{x}^\star\|_2^2 \|\boldsymbol{h}^\star\|_2^2 \mid \boldsymbol{A}\right] \\ &\gtrsim \sigma^2 K / \log m.\end{aligned}$$

D Auxiliary lemmas

In this section, we collect several auxiliary lemmas that are useful for the proofs of our main theorems.

Lemma 18. *Consider any fixed vector $\boldsymbol{x}$ independent of $\{\boldsymbol{a}_l\}_{1\leq l\leq m}$. Then with probability at least $1 - O\left(m^{-100}\right)$, we have*

$$\max_{1\leq l\leq m} \left|\boldsymbol{a}_l^{\mathsf{H}} \boldsymbol{x}\right| \leq 20\sqrt{\log m}\, \|\boldsymbol{x}\|_2 . \tag{D.1}$$

Additionally, there exists some constant $C > 0$ such that with probability at least $1 - O\left(me^{-CK}\right)$, we have

$$\max_{1\leq l\leq m} \|\boldsymbol{a}_l\|_2 \leq 10\sqrt{K}. \tag{D.2}$$

Proof. The first result follows from standard Gaussian concentration inequalities as well as the union bound. The second claim results from [Ver18, Theorem 3.1.1]. □

Lemma 19. *Suppose that $m \gtrsim K \log^3 m$. With probability at least $1 - O\left(m^{-100}\right)$, one has*

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

Proof. See Appendix D.1. □

Lemma 20. *Fix an arbitrarily small constant $\epsilon > 0$. Suppose that $m \geq C\mu^2 K \log^2 m / \epsilon^2$ for some sufficiently large constant $C > 0$. Then one has*

$$\|\mathcal{P}_T \mathcal{A}^* \mathcal{A} \mathcal{P}_T - \mathcal{P}_T\| \leq \epsilon,$$

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

Proof. This has been established in [ARR13, Section 5.2]. □

D.1 Proof of Lemma 19

We intend to invoke [KLT+11, 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 [Ver18, Chapter 2], we have

$$B_{\boldsymbol{Z}} := \left\| \|\xi_j \boldsymbol{b}_j \boldsymbol{a}_j^{\mathsf{H}}\| \right\|_{\psi_1} = \left\| \xi_j \|\boldsymbol{b}_j\|_2 \|\boldsymbol{a}_j\|_2 \right\|_{\psi_1} \le \|\xi_j\|_{\psi_2} \left\| \|\boldsymbol{a}_j\|_2 \right\|_{\psi_2} \sqrt{\frac{K}{m}} \lesssim \sigma \frac{K}{\sqrt{m}}.$$

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. [Ver18, Theorem 3.1.1]). In addition, simple calculation yields

$$\left\| \sum_{j=1}^m \mathbb{E}\big[\boldsymbol{Z}_j \boldsymbol{Z}_j^{\mathsf{H}}\big] \right\| = \left\| \sum_{j=1}^m \mathbb{E}\Big[|\xi_j|^2 \, \boldsymbol{b}_i \boldsymbol{a}_i^{\mathsf{H}} \boldsymbol{a}_i \boldsymbol{b}_i^{\mathsf{H}} \Big] \right\| = \left\| \sum_{j=1}^m \mathbb{E}\big[|\xi_j|^2\big] \mathbb{E}\big[\|\boldsymbol{a}_j\|_2^2\big] \boldsymbol{b}_j \boldsymbol{b}_j^{\mathsf{H}} \right\| \asymp K\sigma^2,$$

$$\left\| \sum_{j=1}^m \mathbb{E}\big[\boldsymbol{Z}_j^{\mathsf{H}} \boldsymbol{Z}_j\big] \right\| = \left\| \sum_{j=1}^m \mathbb{E}\Big[|\xi_j|^2 \, \boldsymbol{a}_j \boldsymbol{b}_j^{\mathsf{H}} \boldsymbol{b}_j \boldsymbol{a}_j^{\mathsf{H}} \Big] \right\| = \left\| \sum_{j=1}^m \mathbb{E}\big[|\xi_j|^2\big] \, \|\boldsymbol{b}_j\|_2^2 \, \mathbb{E}\big[\boldsymbol{a}_j \boldsymbol{a}_j^{\mathsf{H}}\big] \right\| \asymp K\sigma^2,$$

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 [KLT+11, Proposition 2] to derive

$$\left\| \sum_{j=1}^m \xi_j \boldsymbol{b}_j \boldsymbol{a}_j^{\mathsf{H}} \right\| \lesssim \sigma_{\boldsymbol{Z}} \sqrt{\log m} + B_{\boldsymbol{Z}} \log\left(\frac{B_{\boldsymbol{Z}}}{\sigma_{\boldsymbol{Z}}}\right) \log m \lesssim \sigma \sqrt{K \log m} \tag{D.3}$$

with probability exceeding $1 - O(m^{-20})$, where the last inequality holds as long as $m \gtrsim K \log^3 m$.